Behaviorally Informed Machine Learning for Human Mobility

Ekin Uğurel

A dissertation proposal
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Cynthia Chen, Chair

Shuai Huang, Co-Chair

Qi "Ryan" Wang, Member

Program Authorized to Offer Degree:

Civil & Environmental Engineering

University of Washington

Abstract

Behaviorally Informed Machine Learning for Human Mobility

Ekin Uğurel

Co-Chairs of the Supervisory Committee:
Professor Cynthia Chen
Civil & Environmental Engineering

Professor Shuai Huang
Industrial & Systems Engineering

Large-scale digital trace datasets hold considerable promise for long-range transportation planning, offering the potential to observe mobility at metropolitan scales with far greater temporal and spatial resolution than traditional household travel surveys while capturing the regularities in daily travel behavior that underlie trip-making. Passively-collected mobile (PCM) data (e.g., location signals from smartphones and in-vehicle GPS) are central to this promise. Their usefulness, however, is limited by discontinuities in individual trajectories, privacy constraints that restrict data sharing and integration, and representativeness biases that distort inferred patterns of regional travel demand and travel behavior across population groups. This dissertation addresses these limitations through four contributions. First, it develops a multi-task Gaussian Process-based imputation method (grounded in recurring daily, weekly, and seasonal travel behavior patterns) capable of handling both short- and long-duration gaps in GPS traces, significantly improving the completeness and usability of mobility data. Second, it introduces an individualized, physics-regularized learning framework that produces high-fidelity mobility traces reflective of observed movement patterns. These generated trajectories can be scaled to build richer, more diverse mobility datasets for

developing and validating activity-based models. Third, it investigates the predictive signal linking mobility patterns as expressions of travel behavior to sociodemographic attributes that shape those behaviors, and develops imputation strategies for enriching PCM datasets with these inferred labels. This enrichment supports both more detailed planning analyses and a clearer diagnosis of representativeness biases in passively collected data. Finally, through a qualitative study of long-range transportation planners, this dissertation investigates barriers to the adoption of big data products and provides recommendations for their effective integration into planning processes. Together, these contributions bridge methodological advances in machine learning with insights from travel behavior research and the practical needs of public agencies, offering a more transparent and behaviorally coherent foundation for data-driven planning and travel behavior analysis.

TABLE OF CONTENTS

Page

List of Figures 1

List of Tables 5

Chapter 1: Introduction 8
1.1 Motivation 8
1.2 State of the Art 11
1.3 Research Objectives 15
1.4 Organization of the Dissertation 15

Chapter 2: Correcting Missingness in Passively-generated Mobile Data with Multi-Task Gaussian Processes 17
2.1 Introduction 17
2.2 Related Work 22
2.3 Methodological Framework 25
2.4 Implementation 33
2.5 Dataset 34
2.6 Experiments 36
2.7 Discussion 46

Chapter 3: Learning to Generate Synthetic Human Mobility Data: a Physics-regularized Gaussian Process approach based on Multiple Kernel Learning 51
3.1 Introduction 51
3.2 Related Work 55
3.3 Methodology 58
3.4 Numerical Experiments 71
3.5 Discussion 81

Chapter 4: On Predicting Sociodemographics from Mobility Signals 84
4.1 Introduction 84
4.2 Literature Review 88
4.3 Datasets 89
4.4 Methodology 91
4.5 Experiments 101
4.6 Conclusion 106

Chapter 5: MPO's uses of and needs for big data 111
5.1 Introduction 111
5.2 Literature Review 112
5.3 Methods 116
5.4 Findings 117

Chapter 6: Conclusion, Discussion, and Future Work 124
6.1 Discussion 125
6.2 Future Work 127

Appendices 153

Appendix A: Supplementary Material for Chapter 2 154
A.1 Initialization Strategy & Parameter Optimization 154
A.2 Algorithm to Determine Training Data for Experiments in 2.6.2 155
A.3 Detailed Imputation Accuracy Metrics for Robustness Experiments 156
A.4 Benchmark Methods and Model Parameters 156
A.5 Max Speed Threshold Sensitivity Analysis 156

Appendix B: Supplementary Material for Chapter 3 165
B.1 List of Notation Used 165
B.2 Dataset and Software Implementation 165
B.3 Descriptive Statistics 167
B.4 Data Preprocessing and Modeling Considerations 168
B.5 Predictor Variables 170
B.6 Algorithms 171

B.7 Handling Nonstationarity with GPs 174

Appendix C: Supplementary Material for Chapter 4 176
C.1 Experimental Details and Full Model Results 176

LIST OF FIGURES

Figure Number Page

1.1 The relationships among chapters in this dissertation 16

2.1 Overview of imputation workflow for passively-generated mobile data with varying levels of missingness (long and short gaps). Data Preprocessing reduces the computational complexity of the imputation task while also improving the accuracy of training points. Model Development specifies the kernel and mean functions and learns the kernel parameters through marginal log-likelihood maximization. Gap Imputation "corrects" the raw data by imputing the missing locations. 26

2.2 Descriptive analysis of Spectus data for all $2,000$ users. (Top Left) Cumulative distribution of location accuracy. (Top Right) Distribution of observations within each day of the week. (Bottom Left) Histogram of gaps categorized by size. (Bottom Right) Boxplots of temporal occupancy with three temporal resolutions. 37

2.3 Examples of trips fit with an underestimated lengthscale (left) and appropriate lengthscale (right) 39

2.4 (Left) Boxplot of total RMSE for trips in different mobility metric clusters (Right) Boxplot of optimized lengthscales for each trip cluster. 40

2.5 Scatterplots of DTW and temporal occupancy for various gap lengths. (a) $\tau = 1$ week, (b) $\tau = 1$ day, (c) $\tau = 6$ hours, (d) $\tau = 5$ minutes. 43

2.6 Predictions (red) and ground truth test data (blue) for a user tested at $\tau = 6$ hours. MTGP with the proposed kernel (left) outperforms exponential smoothing (right). 46

3.1 (top) Proposed algorithm for greedy multiple kernel learning; (bottom) example greedy learning tree with three steps (checkmark denotes lowest BIC option) 68

3.2 Physics-regularized GP inference for passively-generated mobile data. Circles indicate variables, while diamonds indicate estimation (kernel function) or inference (kernel matrix) processes. 70

3.3 Components of the single-task temporal kernel in a 3-D space 72

3.4 (a) Loss function progression and convergence behavior of different functional forms for a composite kernel; (b) Generating one week of trips using discovered kernel and the temporal GP; (c) Derived mobility metrics by kernel. Bold denotes the optimal kernel. r_g denotes radius of gyration, B_s denotes spatial burstiness, and ΔH denotes distance between user's imputed and actual home location. 73

3.5 Speed and Bearing Constraints in King County, Washington 75

3.6 Speed and Bearing Predictions and True Observations for a Spectus User . . 76

3.7 Evaluation metrics for all models being compared. Y-axis is shown in log-scale. The lower the better for all metrics. 81

4.1 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. 92

4.2 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 95

4.3 Illustration of selected metrics; (left) example travel day. In this case, $n_{\text{trips}} = 7$ and $f_{\text{comp}} = 3/7$; (right) each color denotes a tour to/from the anchors. In this case, $n_{\text{tour}} = 4$ and $f_{\text{mm}} = 2/4$. 96

4.4 Shared-trunk multitask architecture. Mobility descriptors are mapped to a shared representation $h = g(X;\theta)$ by a two-layer feed-forward network with ReLU activations and dropout. Four task-specific heads f_t produce class probabilities via softmax. Training minimizes the cross-entropy loss $\sum_t w_t \cdot \ell_t$ where we set the weights to be equal. 100

4.5 Largest magnitude Spearman rank correlations (all $p < 0.001$) 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. 102

4.6 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) Overall split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data. 108

4.7 Reliability diagrams for representative settings. (TOP) Overall 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. 109

4.8 Performance of the MT variant (in blue) compared to ST variants (in orange) at different fractions of training data. (top four rows) Overall split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data. . 110

5.1 Bar plot of answers mentioned in response to "In which area does your MPO operate?" 117

5.2 Word cloud of answers to "What word comes to mind when you think of "big data"? (98 responses) 119

5.3 Bar plot of factors mentioned in response to "What would help you gain confidence or have trust in using big data?" 120

A.1 Mobility error metrics for $\tau = 1$ week 156

A.2 Mobility error metrics for $\tau = 1$ day 157

A.3 Mobility error metrics for $\tau = 6$ hours 157

A.4 Mobility error metrics for $\tau = 1$ hour 158

A.5 Mobility error metrics for $\tau = 30$ minutes 158

A.6 Mobility error metrics for $\tau = 15$ minutes 159

A.7 Mobility error metrics for $\tau = 5$ minutes 159

A.8 Classical error metrics for $\tau = 1$ week 160

A.9 Classical error metrics for $\tau = 1$ day 160

A.10 Classical error metrics for $\tau = 6$ hours 160

A.11 Classical error metrics for $\tau = 1$ hour 161

A.12 Classical error metrics for $\tau = 30$ minutes 161

A.13 Classical error metrics for $\tau = 15$ minutes 161

A.14 Classical error metrics for $\tau = 5$ minutes 162

A.15 The number of observations eliminated by varying speed limits across the users' data analyzed in Section 2.6 162

B.1 Violin plots of training and testing set sizes across our passively-collected mobile data experiments 167

B.2 Descriptive analysis of Spectus data for all 2,000 users. (a) Cumulative distribution of location accuracy. (b) Distribution of observations within each day of the week. 168

B.3 Cumulative distribution of location accuracy for Spectus users in the Experiments sample (blue) and the whole dataset (grey) 169

B.4 Normalized observation counts by time of day for Spectus users in the Experiments sample (blue) and the whole dataset (orange) 169

B.5 Evolution of time interval between two consecutive points Week 1 to 27. Lines show the percentiles for the experiment sample, while the green and blue fillings shows the 70-80 and 15-35 percentile range for the entire dataset, respectively. 170

B.6 GP regression using cyclical features and a composite kernel accurately captures nonstationary behavior in CO2 concentration of the Mauna Loa volcano . . . 175

C.1 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. 177

LIST OF TABLES

Table Number Page

2.1 Temporal dimensions used in the robustness experiments. 34
2.2 Summary of trip clusters. 40
2.3 Median error with respect to the testing sets. 47
3.1 Optimal kernel parameters for long gap imputation example. Asterisk* denotes the highest lengthscale among categorical/binary variable. 74
3.2 Models and median test set metrics 78
4.1 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) 90
4.2 Selected linear regression models 103
5.1 Distribution of answers to exploratory questions about the audience 118
5.2 Types of questions MPO staffers have solved or plan on solving, categorized 122
A.1 Parameters of benchmark methods. 163
A.2 Parameters used during robustness experiments. 164
B.1 Variable dictionary 166
B.2 Temporal and physical dimensions used in our experiments. Asterisk* denotes a dimension that was not employed in the GeoLife experiments. 171
C.1 Performance across feature sets and models for Age (overall split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 178
C.2 Performance across feature sets and models for Gender (overall split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. . . 179
C.3 Performance across feature sets and models for HH Income (overall split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 180
C.4 Performance across feature sets and models for Number of Children (overall split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 181

C.5 Performance across feature sets and models for Age (2017–2019 train, 2023 test split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 182
C.6 Performance across feature sets and models for Gender (2017–2019 train, 2023 test split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 183
C.7 Performance across feature sets and models for HH Income (2017–2019 train, 2023 test split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 184
C.8 Performance across feature sets and models for Number of Children (2017–2019 train, 2023 test split). Values are mean±sd across folds; best per model in **bold**; best in metric in **red**. 185

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my primary advisor, Dr. Cynthia Chen, for her mentorship throughout my time at the University of Washington. Her academic rigor and continued support were the driving forces that allowed this work to reach completion. I also thank my co-advisor and ISE mentor, Dr. Shuai Huang, and my committee members, Dr. Qi "Ryan" Wang and Dr. Yen-Chi Chen, for their invaluable guidance and the thoughtful questions that pushed me to sharpen my thinking. Additionally, the methodological ideas in Chapter 4 benefited greatly from discussions with Dr. Filipe Rodrigues, whom I also thank for hosting me as a visiting Ph.D. student at the Technical University of Denmark.

This journey would have been far more difficult without the friends and colleagues who provided a vibrant community, making the Ph.D. experience richer than I could have imagined. I am grateful to my former housemates, Michele Martino and J. Felipe Montano-Campos, for their friendship and the memorable times we shared. Thanks also to the wonderful lab mates in Wilcox Hall for their camaraderie and support.

Last, but certainly not least, I am grateful for the endless support of my family. To my parents, Necdet and Aynur Ugurel, thank you for believing in me. To my wonderful partner, Rubina Singh, and our dog, Nia–thank you for your patience and love; I surely would not have made it across the finish line without you.

This research was generously supported by the National Science Foundation on the project entitled "A Whole-Community Effort to Understand Biases and Uncertainties in Using Emerging Big Data for Mobility Analysis" (Award no.: 2114260). The author was also supported by the Center for Teaching Old Models New Tricks (TOMNET), a Tier-1 University Transportation Center sponsored by the U.S. Department of Transportation under grant 69A3551747116, as well as the Center for Understanding Future Travel Behavior and Demand (TBD), a National University Transportation Center sponsored by the US Department of Transportation under grant 69A3552344815 and 69A3552348320.

Chapter 1

INTRODUCTION

1.1 Motivation

For decades, the transportation planning industry has tried to answer some of the most difficult questions that surround the future of urban areas. These include, for example, how to set investment priorities to accommodate population growth while maintaining level of service standards, how to balance competing demands between different transportation modes in limited urban space, how to ensure equitable access to transportation services across diverse socioeconomic groups, and how to develop resilient infrastructure that can adapt to changing climate conditions and emerging technologies like autonomous vehicles. As Rittel and Webber (1973) noted five decades prior, these are rather difficult questions that tend not to have a single correct answer; they represent inherently normative dilemmas where potential solutions are judged not as true or false, but as good or bad depending on stakeholder perspective.

Traditionally, long-range transportation planners (LRTPs) have relied on household travel surveys (HTS) to illuminate the behavioral mechanisms underlying daily travel. These surveys ask residents to record detailed information about their trips, including purposes, modes, timings, and destinations. Yet despite their richness, HTS datasets have well-known limitations: they are costly to administer, achieve modest response rates, and depend on respondents accurately recalling their travel (Li et al., 2025; Richardson et al., 1996; Stopher et al., 2008). Moreover, they capture only a small sample of the population and are collected infrequently (e.g., every eight years in the U.S.; Alberini et al., 2021), and thus cannot fully represent the spatial or temporal heterogeneity of traveler behavior.

The emergence of GPS-based devices has fundamentally transformed our ability to collect mobility data at scale. Unlike the actively-collected HTS, these devices passively generate location data as a by-product of their operation and billing purposes, which planners can then repurpose to infer mobility patterns (Chen et al., 2016a, 2026). This passively-collected data offers unprecedented insights into travel behavior at fine spatial and temporal resolutions, capturing longitudinal information on (1) when people travel, (2) where they go, (3) how long they stay at those destinations, and (4) what routes they take to get there. The past 15

years has seen a surge of studies leveraging this rich data source to analyze collective and individual mobility patterns (Alessandretti et al., 2020; Alexander et al., 2015b; Hao et al., 2020; Sulis et al., 2018).

However, the promise of GPS-based data comes with significant methodological and data-related challenges that must be carefully addressed before it can be used to solve transportation planning problems. First, despite its continuous nature, GPS data often suffers from both short- and long-gaps in observations that can distort the derived mobility metrics (McCool et al., 2022). These gaps in data collection can occur for multiple reasons: user-driven choices (such as restricting location access to when apps are actively in use; Kim et al., 2019), technical limitations (including aggressive background app management by mobile operating systems; Zhou et al. 2020), and environmental factors (like signal loss in urban canyons or tunnels; Ben-Moshe et al., 2011). The temporal distribution of observations is often uneven, with characteristic patterns showing peaks during commute hours on weekdays and fewer observations during nights and weekends. This missingness is not random—it systematically varies across users, time periods, and geographic contexts, creating complex patterns of data sparsity. Such gaps can significantly impact our understanding of mobility patterns. When observations are missing, derived metrics like stay durations, travel distances, and activity patterns may be systematically underestimated. While brief gaps might have minimal impact, longer periods of missingness can substantially bias our understanding of movement patterns, potentially leading to incorrect conclusions about travel behavior and infrastructure needs. This sparsity issue is the motivation behind the imputation method proposed in Chapter 2.

Second, the fine spatiotemporal nature of location traces raises serious privacy concerns. Data aggregators and providers increasingly face strict privacy regulations and user-centric demands for data protection, leading to more restrictive data sharing agreements and easier opt-out options for users. This shift is driven by heightened consumer awareness of data privacy implications and industry-led initiatives to enhance user control over personal data sharing (Apple, 2021). The resulting decrease in available data has profound implications for transportation planning applications, particularly in developing synthetic populations for large-scale mobility simulations. Traditional approaches, such as survey-estimated activity-based models (ABMs) or probabilistic models derived from aggregated mobile data, struggle to reproduce the fine-grained dynamics needed for realistic simulation (Bhat and Koppelman, 1999; Pendyala et al., 1997). Their reliance on coarse aggregates or sparse survey samples leaves

them unable to represent how personal characteristics, situational factors, and environmental context jointly shape movement. These limitations highlight the need for synthetic data methods that both preserve privacy and faithfully reconstruct the underlying data-generating process at the individual level. This is the motivation for Chapter 3, which prescribes a method for generating synthetic data that preserves statistical properties while protecting individual privacy.

Finally, even when privacy-protected GPS data can be collected and processed, the resulting mobility metrics often exhibit systematic biases that must be addressed (Li et al., 2024; Wang et al., 2025). These biases stem from multiple sources: the data may over-represent certain demographic groups while under-sampling others or disproportionately reflect particular travel patterns based on when and where data collection occurs. Unfortunately, the extent of such bias is typically unknown because demographic information is not directly available in passively collected datasets. This missing dimension fundamentally limits the kinds of analyses transportation planners seek to conduct. Metropolitan planning organizations (MPOs) increasingly rely on mobile data to evaluate network investments or assess the distributional impacts of new projects. Many of these applications require linking mobility traces to travelers' demographic attributes in order to gauge, for example, whether tolling policies differentially affect low-income households or whether expanded transit lines serve underserved communities. Without demographic context, agencies risk making decisions based on travel patterns that reflect only the behaviors of overrepresented subpopulations. Consequently, to quantify and ultimately correct for these biases, it becomes essential to infer travelers' underlying demographic attributes from their observed mobility behavior alone. To this end, in Chapter 4, we introduce a standardized set of higher-order mobility descriptors to strengthen demographic signal extraction, calibrate predictive models to characterize uncertainty, and apply multitask learning to improve model generalization across contexts.

Besides the domain context, the common denominator across the technical chapters 2-4 is a shared methodological backbone in behavioral integration. Each technical contribution is guided by empirical regularities and theoretical insights long established in the travel behavior literature. For example, daily and weekly mobility patterns motivate the structured kernels and periodic priors used for imputing missing data. Principles of activity scheduling and space-time constraints inform the individualized, physics-guided mechanisms that shape our synthetic data generation framework. Likewise, decades of work documenting how sociodemographic attributes manifest in activity patterns motivate the higher-order mobility

descriptors and uncertainty-aware predictive models developed for demographic inference. Across all chapters, behavioral theory serves primarily as a structuring device: it constrains model classes to reflect known regularities in human movement, guiding the choice of priors and feature representations. These behavioral constraints help ensure that the resulting models remain grounded in plausible mobility patterns rather than purely data-driven artifacts.

Building models around behavioral principles addresses only part of the challenge; equally important is understanding how such methods are received and evaluated within the institutions that rely on them. Despite the growing availability of passively generated mobile data, many planning agencies remain cautious about integrating these products into core decision-making processes. Their hesitancy is not solely due to the technical limitations discussed above (i.e., sparsity, privacy constraints, and demographic bias/uncertainty). It also reflects deeper questions about how these datasets fit within existing institutional, professional, and regulatory structures. Planners must judge whether new data sources are trustworthy, whether they align with established norms of practice, and how they should be used in the absence of clear federal or state guidance. These questions become even more acute for smaller Metropolitan Planning Organizations (MPOs), which often lack the staff capacity or analytical resources to independently assess complex data products. Chapter 5 examines these institutional and organizational dimensions through an in-person workshop at the Association of Metropolitan Planning Organizations (AMPO) Planning Tools & Training Symposium in Albuquerque, New Mexico, in May 2024, using this setting to understand what planners need in order to responsibly and confidently adopt big data products.

1.2 State of the Art

This dissertation develops behaviorally informed machine learning approaches to address issues in passively-collected GPS traces. The goal is to make these data usable and reliable for transportation planning applications. Specifically, with respect to these datasets, this dissertation aims to (1) fix sparsity (both short and long gaps); (2) overcome privacy concerns by generating synthetic data; (3) characterize representativeness bias by imputing sociodemographics from mobility signals; and (4) understand how MPOs use big data, what they need from it, and what barriers shape its adoption. Correspondingly, I will review several existing research areas that intersect with these challenges.

1.2.1 Correcting Missingness

Missing data imputation for mobile datasets generally involves three strategies: (1) time-series smoothing and interpolation, (2) leveraging external datasets, and (3) machine learning methods, including kernel- and deep learning-based approaches. Traditional time-series models (e.g., SARIMAX) offer interpretability but struggle with the non-linear, irregular nature of mobility data (Huo et al., 2010; Kohn and Ansley, 1986; McCool et al., 2022). External datasets, including transit smart card records or mobile data from other providers (Gong et al., 2020), offer a promising way to bolster sparse mobile traces. Yet they also pose substantial obstacles: access to these data can be limited, privacy rules may restrict sharing, and cross-source similarity matching tends to be non-trivial (Cao et al., 2016; Kondor et al., 2020). Machine learning approaches like RNNs or GNNs effectively capture spatiotemporal dependencies but often operate as black-boxes, with limited interpretability (Ren et al., 2021; Sun et al., 2021). In Chapter 2, we proposed a novel Gaussian Process (GP)-based framework to impute both short- and long-gaps, capturing correlations across user coordinates via multi-task learning and leveraging a distance-based compression algorithm to manage scalability. We find that our Bayesian framework outperforms a host of competing approaches based on how closely we recover the underlying mobility statistics of partially-observed trajectories.

1.2.2 Generating Synthetic Data

While imputation addresses sparsity, many planning applications still require privacy-preserving representations of mobility that can be shared, simulated, and analyzed without exposing individual traces. This has motivated a growing literature on generating synthetic trajectories that reproduce heterogeneous travel patterns. Early work relied on structured stochastic models that separate temporal routines from spatial movement, learning diary-like Markov chains over activity states and then assigning locations via preferential exploration and return rules (Pappalardo and Simini, 2018). More recent studies rely on deep generative models. Recurrent neural networks (RNNs) have been trained to simulate individual trajectories and, in extensions, full synthetic populations over multiple days (Berke et al., 2022; Kulkarni and Garbinato, 2017). Generative adversarial networks (GANs) augment these sequence models with discriminators and mobility-regularization terms to better match empirical distributions and support downstream tasks such as epidemic simulation (Feng et al., 2020). Diffusion models push fidelity further by learning denoisers that map noise to realistic GPS trajectories (Zhu et al., 2023). Finally, transformer architectures have also been

adapted to treat activities, locations, and modes as tokens (conditioned on sociodemographics) to generate rich mobility profiles (Zhang et al., 2024b).

Despite this progress, two gaps remain for applications of methods datasets in transportation contexts. First, most existing generators are global models that do not explicitly tailor their representation to each traveler's idiosyncratic temporal and spatial patterns, even though heterogeneity in routines is central to travel behavior (Kroesen, 2014; Lee et al., 2007; McGuckin and Murakami, 1999a). Second, these models seldom embed simple physical or built-environment constraints and rarely quantify uncertainty, which limits their use for risk-sensitive planning applications. Chapter 3 addresses these gaps by recasting synthetic trajectory generation as a Bayesian learning problem. We learn individualized Gaussian Process kernels that adapt to each person's observed mobility, and we incorporate physical and temporal attributes of the built environment when shaping these kernels. Sampling from the resulting posterior yields synthetic traces that reflect traveler-specific structure, respect basic physical constraints, and come with principled uncertainty quantification.

1.2.3 Inferring Demographics from Mobility Behavior

Complementing efforts to generate privacy-preserving synthetic mobility data, a growing body of research has explored inferring sociodemographic attributes from digital behavioral traces, motivated by the need to link mobility data with population characteristics for planning and policy analysis. Within this literature, mobility-focused approaches rely on *where and when* people travel, transforming raw coordinates into spatial, temporal, and semantic descriptors that proxy for demographic traits. Spatial metrics (e.g., radius of gyration, location diversity), temporal regularities (e.g., commuting rhythms), and semantic signals (e.g., POI visit frequencies, tour structures) have each been shown to carry predictive power for attributes such as age, gender, household structure, and work status (Auld et al., 2015; Ding et al., 2019; Doi et al., 2021; Wu et al., 2019). However, these associations are often context-specific and ad hoc, and few studies systematically evaluate the robustness or incremental value of each feature family across settings.

Methodologically, the field has evolved from classical statistical models using hand-crafted features (as in Auld et al., 2015) to representation learning techniques that encode daily trajectories as structured sequences or graphs (as in Solomon et al., 2018). Neural embedding methods and recurrent models have improved predictive performance by capturing temporal and contextual regularities (Ding et al., 2019; Xu et al., 2020), but most prior work trains

separate models for each demographic attribute, limiting data efficiency and transferability under distributional shifts. Moreover, uncertainty is often handled superficially: while most studies report aggregate accuracy, few quantify how confident the model is in its predictions for individual travelers. Recent advances in calibration and theoretical separability provide tools to address this gap, enabling more uncertainty-aware inference. Building on these developments, Chapter 4 evaluates the predictive contribution of different mobility feature families, leverages multitask learning to exploit shared structure across demographic targets, and incorporates calibration methods to produce more reliable, interpretable uncertainty estimates.

1.2.4 Uses and Adoption of Big Data by MPOs

MPOs increasingly view big data as a way to strengthen core planning tools, but most documented uses remain targeted and cautious. Agencies commonly deploy probe-based speed and volume data to calibrate and validate regional travel demand models (Bauer et al., 2014; FHWA, 2020). Some MPOs have used vendor products to estimate AADT or construct OD matrices for corridor studies, subarea analyses, or to cross-check regional commuting patterns, often treating these datasets as a "second source" rather than a primary input (KimleyHorn, 2021; StreetLight, 2025). Planners also see potential to use smartphone traces to study ride-hailing and to evaluate accessibility across communities (Boyd et al., 2024). In practice, however, many of these aspirations remain exploratory and are constrained by concerns about representativeness bias (Richardson, 2021; Wang et al., 2025).

Despite the proliferation of big-data products, adoption within MPOs is limited by intertwined technical, organizational, and institutional barriers. Planners frequently cite opaque vendor methods, uncertain data quality, and shifting proprietary algorithms as major obstacles to trusting and integrating these datasets into long-range planning workflows (KimleyHorn, 2021; Singh et al., 2022). High and often opaque cost structures make it difficult for smaller agencies to justify long-term subscriptions (Bauer et al., 2014; KimleyHorn, 2021). Many MPOs also lack in-house data engineering or advanced analytics capacity, facing difficulties in recruiting and retaining staff with big-data expertise (Pecheux et al., 2020; Richardson, 2021). These constraints interact with organizational culture and leadership; skepticism among executives about the value of big data, siloed data practices, and low psychological safety around experimentation all reduce agencies' absorptive capacity and confine the use of big data to pilot projects (Pecheux et al., 2020). Chapter 5 engages directly

with these issues by examining how planners articulate the uses they see for big data, how they evaluate vendor products, and which institutional, technical, and organizational factors most strongly shape adoption.

1.3 Research Objectives

The objectives of this dissertation are:

1. To develop and validate a multi-task Gaussian process-based imputation method for short- and long-gaps in passively-collected GPS traces;

2. To create an individualized, physics-regularized framework for generating high-fidelity synthetic traces that replicate observed patterns;

3. To develop and evaluate methods for inferring travelers' sociodemographic attributes solely from mobility behavior, using higher-order mobility descriptors, uncertainty calibration, and multitask learning to improve predictive accuracy and generalization across contexts, and;

4. To investigate and document the barriers to adoption of big data products among transportation planning organizations, providing insights into how these tools can be more effectively integrated into planning processes.

1.4 Organization of the Dissertation

This dissertation is presented in a multiple manuscript format. Chapters 2-5 are written as individual research papers, including an abstract, a main body, and references. The relationships among these chapters are depicted in Figure 1.1. Chapter 2, proposes a Bayesian ML method to correct (i.e., impute) missing data in passively-generated GPS traces, taking advantage of multiple periodicities that exist in humans' travel trends. Chapter 3 extends this work with individualized kernel learning and physics-regularization, facilitating the generation of synthetic traces that replicate users' observed travel patterns. This allows the research community to avoid compromising individual privacy while working with passively-generated GPS traces and share their synthetic counterparts. Chapter 4 investigates the extent to which key sociodemographic attributes (such as age, gender, income, and household structure) can be inferred solely from individuals' mobility patterns. The results highlight the potential

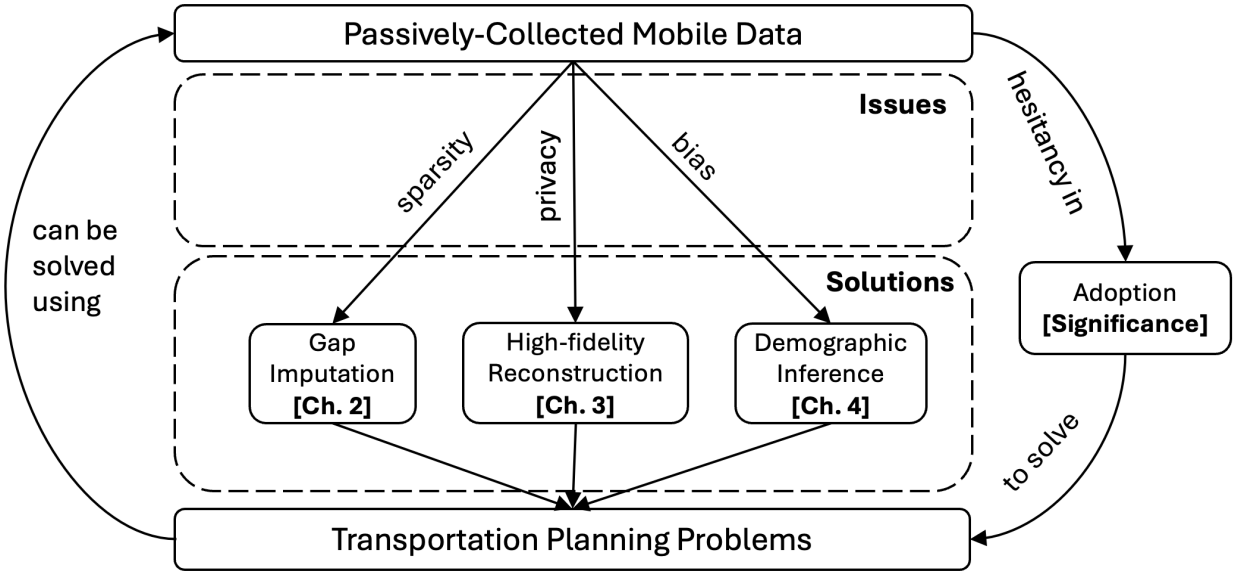


Figure 1.1: The relationships among chapters in this dissertation

to enrich passively collected mobility datasets with inferred sociodemographic information, enabling more comprehensive behavioral analyses without relying on direct survey linkage. Finally, Chapter 5 presents a qualitative study of long-range transportation planners' uses of and needs for big data products, tackling the limited adoption of these datasets among the planning community.

Chapter 2

CORRECTING MISSINGNESS IN PASSIVELY-GENERATED MOBILE DATA WITH MULTI-TASK GAUSSIAN PROCESSES

The prevalence of mobile devices and the ubiquity of network connectivity have generated a massive amount of temporally- and spatially-stamped data. A key characteristic of this mobile data is the prevalence of sparsity—for every recorded user, there is a significant amount of missing data. Sparsity leads to bias in the inferred mobility patterns and thus, correcting bias by imputing the missing data potentially creates opportunities for directly using the corrected mobile trajectory data for large-scale simulations in various applications. We propose a multi-task Gaussian process regression model to correct missingness in mobile data. Gaussian processes (GPs) allow for flexible modeling of diverse (i.e., non-linear, locally periodic) data patterns and quantify prediction uncertainty in an interpretable manner. We develop a methodological framework for applying GPs to mobile data. In doing so, we consider the correlations between users' coordinates (latitudes and longitudes) through multi-task learning and adjust for individual-level differences in data characteristics through parameter initialization and optimization. We introduce and demonstrate the use of smooth (i.e., rational quadratic) and periodic kernels in modeling human mobility data. Relatedly, we analyze our model's parameters and imputation accuracy with respect to different types of trips (e.g. walking trips vs. highway trips). We also demonstrate our model's performance in three experiments with real app-based data, in which it outperforms alternative imputation methods.

2.1 Introduction

The past two decades have brought a range of technological advancements that have made it easy to gather large sets of time- and location-stamped mobile data. Coupled with the ubiquity of network connectivity, this has led to a massive increase in individual trace sightings. Those datasets include tracks generated by Global Positioning Systems (GPS) or other position-signaling devices, geo-tagged posts from social media platforms, and call detail records (CDRs) derived from triangulation of cellular tower positions. Human mobility

patterns inferred from such datasets have been used for many applications, including for example, to quantify urban vitality (Sulis et al., 2018), understand commuting patterns (Frias-Martinez et al., 2012), and predict pandemic spreading (Hao et al., 2020).

Unlike household travel surveys, which design and implement a set of pre-defined questions to answer (thus called "actively-solicited data"), mobile data is passively-generated as a byproduct of processes that have little to do with answering research questions (Chen et al., 2016b). As a result, while mobile data offers sample sizes orders of magnitudes greater than traditional travel surveys, there exist many issues relating to its representativeness. Chief among these issues is sparsity—for a substantial portion of users, there is a significant amount of missing data. In other words, sparsity occurs when the observation frequency of a device is low.

The observation frequency of mobile data varies greatly across users, periods, and geographies for various reasons. Users commonly have the choice of only allowing location transmissions while apps are in use (rather than continuously in the background). Perceived benefits in network externalities (e.g., increased accuracy for Google Maps travel time estimation) have been linked to greater willingness to provide personal information, such as location data (Kim et al., 2019). Therefore, third-party apps that do not provide any perceived benefits to the data-solicited user may receive limited access to location data. Certain versions of the Android operating system (OS) have also been found to aggressively shut down apps running in the background, preventing continuous data collection (Zhou et al., 2020). In addition to stochastic events like battery drain, user- and OS-based decisions are likely to lead to long observation gaps for users, ranging anywhere from a few hours to multiple days or weeks.

Additionally, app-based datasets have large variations in their intra- and inter-day sparsity—observations tend to cluster temporally, resulting in frequency 'peaks' and 'valleys'. Ban et al. (2018) offer two insights regarding the temporal distribution of observations in app-based datasets: First, during weekdays, observations peak in the morning (7-9 AM) and in the evening (4-6 PM), while during weekends there is only one mid-day peak (12-3 PM). There are also significantly fewer observations overnight than during the day. Second, weekdays have more observations than weekends—Fridays tend to have the most observations, while Sundays tend to have the fewest, with the exception to the rule being holidays. This clustering tendency can be caused by users' mobile device usage patterns or the update frequency of the data provider. The problem is further shrouded by the opaqueness of those providers in disclosing neither the source of the data nor the reason for its generation.

Finally, missingness may also be caused by a range of geographical factors. Shorter gaps in continuity are often the result of temporary signal loss, which may occur in dense urban areas, while traveling through tunnels and other enclosed infrastructure, or due to the "cold-start problem"—when signals are dropped due to a lack of clear line of sight from GPS satellites. Additionally, the "urban canyon effect" describes the tendency of receivers in Global Navigation Satellite Systems (GNSS) to output erroneous location estimations while operating in areas with a high density of tall buildings, which block the satellite's direct line of sight to the receiving device (Ben-Moshe et al., 2011).

When a mobile dataset has missing data, derived mobility patterns are vulnerable to bias, meaning that they may misrepresent individuals' actual movements. Several outputs may be altered due to missing data, such as the duration of the inferred stays (Ban et al., 2018), the extent of one's travel (McCool et al., 2022), and users' interactions with the built environment (e.g., the types of locations visited; Merrill et al., 2020). While short and infrequent observation gaps have minimal influence on derived mobility metrics, longer and more frequent gaps will result in a downward bias (McCool et al., 2022). On the other hand, improvements in data quality, such as decreases in rates of missingness or increases in location data accuracy (i.e., the closeness of observation to the real location), have been linked to more accurate calculations of related mobility metrics, such as home census tracts and trip rates (Ban et al., 2018). Correcting missingness in mobile data is thus important to researchers and practitioners alike.

Before prescribing a method, we briefly outline some of the challenges that are present while modeling individual mobile data. In general, people tend to use various modes of transportation to get around. Different modes have different underlying physics (i.e., average velocity, acceleration behavior), meaning that they possess varying data-generating mechanisms. Additionally, people tend to change modes, either within the same trip or between different trips. These realities suggest that any method to correct missingness needs to be flexible enough to capture different data-generating mechanisms and the transition between these states.

A further complexity involves individuals' heterogeneous travel behavior. Previous works have explored the predictability of human mobility, proposing various models and outlining their properties (Barbosa et al., 2018; Gonzalez et al., 2008; Song et al., 2010a). In general, however, no single model has outperformed every other model in every context—in contrast, each model has had success in different domains. Thus, a key takeaway is that there may not

exist universal "laws" of human mobility—different people have different tendencies to explore and exploit their environments and therefore the distributions of their data are different. Oftentimes, the data distribution depends heavily on context and time (Hills et al., 2015). For example, an individual who has recently moved to a new city may be more interested in trying out a range of restaurants, since they have had no prior experience to guide their preferences. On the other hand, a 15-year resident of a neighborhood may have only one or two restaurants they regularly visit.

These individual-level complexities suggest that making rigid assumptions about the shape or form of the underlying distribution of an individual's mobile data may not be a successful approach. Rather, a non-parametric data-driven approach may be a promising one—such models are less restrictive, can accommodate complex patterns, and can better capture underlying relationships in the data, which may involve both temporal and spatial correlations. One example of a family of non-parametric methods includes Gaussian processes (GP). GPs can be understood as a generalization of multivariate Gaussian distributions to infinitely many variables. Two components define a GP—the mean function specifies the mean at any point in the input space, and the covariance function (or kernel) embeds a measure of similarity between any pair of points in a multi-dimensional space. In practice, specifying a kernel that can accurately capture relationships in the data is a rather difficult task. Simultaneously, a fitting kernel can seamlessly capture complex dependencies in mobile data.

In the context of mobile phone trajectory data, which often exhibits highly non-linear and context-dependent human mobility patterns, Gaussian Processes (GPs) offer an appropriate framework for capturing these complexities. For instance, various travel modes exhibit distinct underlying physics—cars move faster than bikes, and whereas a walker can move freely in any direction in a park, a car is limited to paved roads. GPs can manage these differences with kernel functions, facilitating context-dependent modeling of mobile data points. Moreover, GPs provide not only point estimates but also uncertainty estimates, which are invaluable for downstream analysis. The precision of the mobility patterns derived from trajectory data can be significantly influenced by these uncertainty measures. Additionally, there is an increasing necessity to integrate diverse data types, such as categorical, continuous, or even textual data, and GPs are well-equipped to handle this complexity effectively.

In this study, we propose a generalizable multi-task GP framework to correct missingness in mobile data at the individual level. Our framework imputes a user's latitudes and longitudes

simultaneously based on a set of predictor variables (primarily time), hence the word multi-task (the modeling of latitudes and longitudes as two tasks). We train the framework's parameters on longitudinal trajectory data. We use the word "generalizable" to refer to two of our framework's properties: (1) its ability to accurately correct varying levels of missingness (i.e. short and long gaps), (2) its propensity to capture changing data-generating mechanisms (i.e. due to mode changes). Both properties are achieved through the specification of an appropriate covariance function and the ensuing parameter optimization process. In this way, the model is kept general enough to apply to any user and unique enough to be accurate.

No model is perfect and GPs are no exception. One notable trade-off of using GPs is its demand on computational resources: though GPs have become more scalable, they can still be computationally demanding for extensive datasets. Our implementation addresses this challenge by leveraging GPyTorch (Gardner et al., 2018), an efficient GP implementation that reduces the asymptotic complexity of exact GP inference from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2)$. Another trade-off involves model complexity. Setting up GPs requires meticulous attention to kernel functions and hyperparameters, contributing to their perceived complexity. Our study demonstrates that kernel choices are directly related to the underlying behavioral patterns and the estimated hyperparameters exhibit systematic patterns that can be explained by human mobility behaviors (Sections 2.3.3 and 2.6.1). Additionally, compared to parametric regression methods, GPs are often seen as "black-box" models compared to parametric models, which might limit the interpretability of the results for some users. However, GPs are considered less of a black-box in comparison to deep neural networks, as choices of basic kernels are based on the nature of the phenomenon of interest (in this context, human mobility patterns) and their hyperparameters such as lengthscale are explainable (they directly tell us the periodicity of the underlying function, which characterizes the periodicity of individual mobility patterns). Additionally, in our context which is imputing missing data in the raw mobility trajectories, interpretability is not a primary concern. This is because the outputs we seek to obtain are repaired trajectories with the missingness filled; those repaired trajectories still need further processing to obtain meaningful mobility metrics such as the number of stays.

We summarize our contributions below:

- We develop a methodological framework to correct missingness in heterogeneous mobile data using a multi-task GP model. The resulting (repaired) data can be used for many transportation and mobility applications downstream. Our model considers the

correlations between users' coordinates and adjusts for individual-level differences in data characteristics;

- We analyze our model's key hyperparameters (e.g., lengthscale parameters) and imputation accuracy concerning different types of trips (e.g. walking trips vs. highway trips). Specifically, we show the linkage between trip types and lengthscale parameters;

- We introduce and demonstrate the combined use of smooth kernels (i.e., the rational quadratic kernel) and kernels with periodicities (i.e., the periodic kernel) in modeling human mobility data; and

- We benchmark our model against other missing data imputation methods, demonstrating its effectiveness under a range of missingness scenarios (short and long gaps).

The rest of this chapter is organized as follows. Section 2.2 summarizes related work on missing mobile data imputation. Section 2.3 outlines our methodology. Section 2.4 presents our implementation while Section 2.5 describes the dataset we used in our experiments. Section 2.6 shows our results. We discuss the implications of our work in Section 2.7.

2.2 Related Work

2.2.1 Methods for imputing missingness in the data

In general, missing data imputation methods for mobile data fall into one of three categories: (1) using time-series smoothing and interpolation methods, (2) leveraging external data sources to facilitate co-learning or (3) employing machine learning methods such as kernels or deep learning architectures. The first approach uses existing methods for smoothing and time-series interpolation to impute missing data points. These range from simpler methods like linear interpolation to more complex ones like seasonal auto-regressive integrated moving averages (SARIMAX). These models have a long history of being used to impute missing data in a range of contexts (i.e., in hydrology, finance, etc.) and therefore are more mature (Cipra et al., 1995; Huo et al., 2010; Kohn and Ansley, 1986; McCool et al., 2022). Though these methods are relatively interpretable, their interpretability depends on the validity of their assumptions, which tend to relate to the regularity of the data (i.e., seasonality or trend), linear dependence over time, and independent noises. Relatedly, the drawback of these

models is they are often not flexible enough to handle the non-linear nature of individual-level trajectory data.

The second approach leverages external data sources to make inferences about the original dataset. This strategy is especially common with CDRs, which provide limited spatial and temporal precision and thus struggle to capture fine-grained movement patterns. External datasets such as transit smart card records or mobile data from other providers can help enrich CDR or GPS trace datasets when the overlap between sources is strong. Gong et al. (2020) developed two indicators to measure the similarity of spatiotemporal trajectories across multiple data streams, and related work has explored the *matchability* of individuals and trajectories across sources using probabilistic techniques (Cao et al., 2016; Kondor et al., 2020). Despite this promise, multi-source fusion faces several practical obstacles: external datasets may not align spatially or temporally with the primary data, may be costly to obtain, or may be restricted by privacy regulations. Even when available, integrating and preprocessing multiple data sources can be resource-intensive, and robust cross-source matching remains technically challenging.

Finally, kernel- and deep architecture-based learning methods have also gained traction in the context of missing data. These methods often boil down to one of two approaches. The first involves using kernels or activation functions to predict the similarity between two spatial trajectories, then using the more complete trajectory to fill in gaps in the sparser trajectory. For example, Wang et al. (2020) utilize the head-direction information of trajectories together with the displacement attributed to an attention mechanism to learn from past trajectory points with different priorities. Similarly, Liu and Onnela (2021) bidirectionally sample discrete displacements for missing segments based on similar trajectories in linear time complexity. The drawback of this approach is that even if two trips have the same ground-truth trajectory, varying data sampling rates (due to the data provider) may distort the shape of the trajectory observed in the data. This could lead one to conclude that these two trips are not similar. Unless access to consistently high-sampling rate training data is available, using trajectory shapes as a measure of similarity is liable to produce errors.

The third approach involves using recurrent neural nets (RNN), attention mechanisms, or other related methods to capture long- and short-term temporal dependencies in users' locations. This is the approach we take in this paper, but rather than using time- and resource-intensive deep neural networks, we opt for multi-task GPs. We review two examples of the deep neural network approach. Sun et al. (2021) propose a model that captures

complex location transition patterns with graph neural networks and uses two attention mechanisms to capture the multi-level and shifting periodicity of human mobility. However, this model is less adept at fine-grained trajectory recovery, which involves accurately tracking and predicting very specific, detailed movements at a high temporal resolution. On the other hand, Ren et al. (2021) propose a framework that can recover and map match the fine-grained points in trajectories from coarse-grained GPS data. It uses a multi-task sequence-to-sequence learning model to capture the spatial and temporal dependencies of trajectories. However, their model requires knowledge of the underlying street network, which may not be available in a lot of scenarios (relating to the above paragraph on external data sources).

2.2.2 GPs in transportation and mobility applications

The use of Gaussian processes in transportation and mobility applications is not new. As early as 2009 (a few years after Rasmussen & Williams (2006) released their pivotal book), GPs were being proposed to predict travel times for arbitrary origin-destination pairs (Idé and Kato, 2009). Since then, GPs have been leveraged in a variety of domains, including in traffic operations and forecasting (Xie et al., 2010), transportation system estimation (Liu et al., 2022c), and pedestrian behavior modeling (Nasernejad et al., 2021).

Within the field of traffic operations and forecasting, Gaussian Processes (GPs) have emerged as a powerful tool for modeling and predicting traffic dynamics. Yuan et al. (2021) proposed a physics-regularized GP framework, pushing the boundaries of macroscopic traffic flow modeling. Beyond macroscopic flows, GPs have been used to predict fine-grained traffic speeds, as demonstrated by Le et al. (2016). Similar to the method of our paper, Rodrigues et al. (2018) leveraged multi-task GPs to impute missing traffic speeds while considering the spatial dependencies with nearby road segments. They found that using GPs to capture spatial correlations with nearby road segments led to substantial improvements in imputation performance over the benchmark methods.

Concurrently, GPs have demonstrated efficacy in handling uncertainties within transportation systems. Storm et al. (2022), for example, introduced an efficient method for evaluating stochastic traffic flow models using GP-based approximations. This not only enhanced computational efficiency but also refined the accuracy of stochastic traffic models. Expanding on uncertainty management, Steentoft et al. (2024) utilized GPs to provide uncertainty estimates for mobility flows derived from large-scale taxi data. Their approach enhanced the reliability of mobility flow predictions in addition to addressing the critical

issue of variable selection in complex transportation datasets.

GPs have also found their use in shared mobility and city-scale travel demand modeling. In addressing the challenges of micro-mobility planning, (Gammelli et al., 2020) focused on improving demand forecasting accuracy with censored data, which inherently contains a biased representation of the true demand due to supply constraints. They innovatively employed Gaussian Processes (GPs) to replace the traditional linear functional form in specifying the likelihood of data being censored. This approach enables the model to avoid bias due to censoring and improve the accuracy of demand predictions. On the other hand, Batista et al. (2022) employed GPs for estimating city-scale OD matrices, focusing on supply-related characteristics of urban networks. Their methodology provides an efficient alternative to Monte Carlo simulations, offering computational efficiency through iterative OD pair selection and shortest trip determination. Significantly, their method creates smaller synthetic sets, substantially reducing computational demands.

In our study, we also leverage Gaussian Processes, extending their application to a different facet of travel demand modeling–specifically, addressing missing data in mobile trajectory datasets. There is a thematic resonance between our work and that of Batista et al. (2022), particularly in the use of GPs to handle complex urban mobility data. Both studies underscore the flexibility of GPs in urban mobility contexts, whether it be in handling synthesizing trip sets or managing missing trajectory data. Uniquely, our method introduces multi-task learning to exploit correlations in user coordinates while also incorporating a distance-based compression algorithm to reduce the size of our training data, acknowledging the bounded nature of individual mobility patterns as suggested by González et al. (2008). Furthermore, our model utilizes a unique kernel designed to align with the characteristics of our dataset as well as findings from the human mobility literature, which suggest varying levels of regularity in individual travel behavior patterns (Kitamura and Van Der Hoorn, 1987; Song et al., 2010b; Teixeira et al., 2021). This tailored kernel selection echoes the importance of kernel choice in GP models, as also highlighted by Batista et al. (2022) in their work.

2.3 Methodological Framework

Our imputation workflow has three modules: Data Preprocessing, Model Development, and Gap Imputation, described in detail in Figure 2.1. The first module filters out erroneous and noisy data points, compresses a user's trajectory using a pairwise distance-based algorithm, and normalizes a user's coordinates. Model Development identifies appropriate kernel functions and

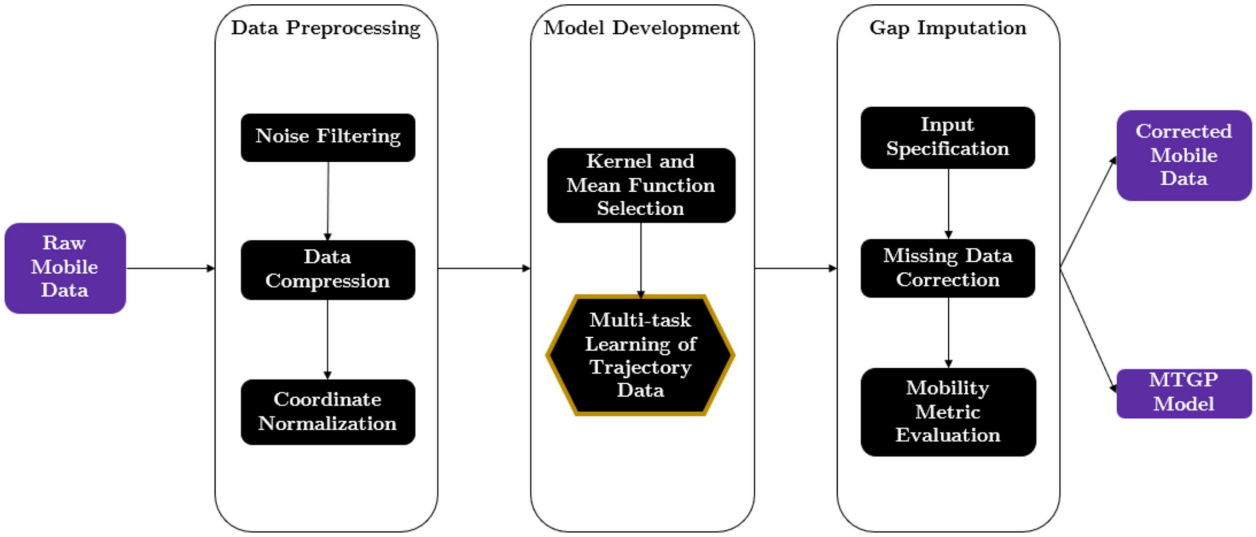


Figure 2.1: Overview of imputation workflow for passively-generated mobile data with varying levels of missingness (long and short gaps). Data Preprocessing reduces the computational complexity of the imputation task while also improving the accuracy of training points. Model Development specifies the kernel and mean functions and learns the kernel parameters through marginal log-likelihood maximization. Gap Imputation "corrects" the raw data by imputing the missing locations.

learns related parameters based on the marginal log-likelihood (the loss function) concerning the training data. Given a set of missing time inputs, Gap Imputation predicts the most likely location for those inputs.

2.3.1 Spatiotemporal Modeling of Human Mobility with Multi-Task GP Learning

We adopt a multi-task GP framework to capture the correlations between two highly correlated tasks in mobile data: predicting latitudes (ϕ) and longitudes (λ). Given a set $\mathbf{X}$ of n inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we define $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_\phi \\ \mathbf{y}_\lambda \end{bmatrix}^T = \begin{bmatrix} y_{1\phi}, & \ldots, y_{n\phi} \\ y_{1\lambda}, & \ldots, y_{n\lambda} \end{bmatrix}^T$, where y_{ij} is the response for the j^{th} task at i^{th} observation, and j refers to either latitudes ϕ or longitudes λ.

We assume the underlying data generation process for y_{ij} as follows:

$$y_{ij} = f_j(\mathbf{x}_i) + \epsilon_{ij} \tag{2.1}$$

where f_j is the latent function mapping inputs $\mathbf{x}_i$ to outputs y_{ij} and ϵ_{ij} is a white noise

process associated with the j^{th} task and $\epsilon_{ij} \sim \mathcal{N}(0,\ \delta_j^2)$ are independent random variables. y_{ij} is assumed to be normally distributed, or $y_{ij} \sim \mathcal{N}(f_j(\mathbf{x}_i),\ \delta_j^2)$.

In multi-task Gaussian process learning, a multivariate normal distribution is used as the prior for modeling multi-outputs (in our case, there are two: latitudes and longitudes). In this setting, we can consider both the temporal correlation within a task but also the correlation between tasks. Thus, we denote the covariance matrix for all n observations and all m tasks as $\mathbf{K}$, which can be expressed as follows:

$$\mathbf{K} = \mathbf{K}^f(\mathbf{y}_\phi, \mathbf{y}_\lambda) \bigotimes \mathbf{K}^x(\mathbf{X},\ \mathbf{X}) \tag{2.2}$$

where $\bigotimes$ is the Kronecker product, $\mathbf{K}^x$ is the covariance matrix of the training data for temporal correlations (n by n), $\mathbf{K}^f$ is the inter-task covariance matrix (Bonilla et al., 2007). In our context where $m = 2$, the resulting covariance matrix $\mathbf{K}$ is $2n$ by $2n$. The hyperparameters associated with $\mathbf{K}^x$ and $\mathbf{K}^f$ can be estimated by minimizing the negative marginal log-likelihood of the dataset (see Section 2.3.4).

In inference, for a new input $\mathbf{x}_*$, the predictive distribution of the output $\mathbf{y}_*$ is also Gaussian and can be computed as[1]

$$p(\mathbf{y}_* \mid \mathbf{x}_*,\ \mathbf{X},\ \mathbf{Y}) \sim \mathcal{N}(\boldsymbol{\mu}_*,\ \boldsymbol{\sigma}_*^2) \tag{2.3}$$

where

$$\boldsymbol{\mu}_* = (\mathbf{k}_j^f \bigotimes \mathbf{k}_*) \left(\mathbf{K}^f \bigotimes \mathbf{K}^x + \mathbf{D} \bigotimes \mathbf{I} \right)^{-1} vec(\mathbf{Y}), \tag{2.4}$$

$$\boldsymbol{\sigma}_*^2 = \left(\mathbf{k}_j^f \bigotimes \mathbf{k}_{**} \right) - (\mathbf{k}_j^f \bigotimes \mathbf{k}_*) \left(\mathbf{K}^f \bigotimes \mathbf{K}^x + \mathbf{D} \bigotimes \mathbf{I} \right)^{-1} (\mathbf{k}_j^f \bigotimes \mathbf{k}_*). \tag{2.5}$$

Here, $\mathbf{k}_j^f$ selects the j^{th} column of $\mathbf{K}^f$, $\mathbf{k}_* = \mathbf{k}(\mathbf{x}_*,\ \mathbf{X})$ is the covariance vector between the test point and the training inputs, $\mathbf{k}_{**} = \mathbf{k}(\mathbf{x}_*, \mathbf{x}_*)$ is the auto-covariance of the test point, $\mathbf{D}$ is a 2×2 diagonal matrix with the variances of the noise processes for latitude and longitude, and $\mathbf{I}$ is the identity matrix.

[1] For more background on the derivation of these equations, refer to Section 2.2 on Rasmussen and Williams (2006)

2.3.2 Using Kernels to Model Human Mobility

Human mobility observations at nearby time inputs tend to be spatially close to one another. This is an expanded application of Tobler's first law of geography—that near things are more related than distant things. To model this phenomenon, we use the rational quadratic kernel:

$$k_{RQ}(x,\ x') = \Omega^2 \left(1 + \frac{\left(x - x'\right)^2}{2\alpha l^2}\right)^{-\alpha} \tag{2.6}$$

where x and x' are two inputs, Ω is a scale parameter, determining the average distance of the function away from its mean, $\alpha > 0$ is the scale mixture parameter, determining the relative weighting of large-scale and small-scale variations in the data, and $l > 0$ is the lengthscale, specifying the smoothness of the function (i.e., the frequency of the gradient of the function to change sign). The lengthscale is commonly interpreted as the maximum distance the model can extrapolate beyond the training data without reverting to the mean. Note that as x and x' approach each other, the term with the exponent $-\alpha$ goes to 1, signifying the increasing likelihood of nearby points to covary.

This works fine for smooth trajectories with relatively low levels of sparsity. For prediction regions in which we have little-to-no observations, data from other days, weeks, and months for the same individual can be leveraged to improve model performance. This is possible because human mobility data exhibit patterns of periodicity.

Previous literature suggests that people's mobility patterns exhibit history dependency, meaning that where one is at time t depends upon their location at $t-1$ (Kitamura and Kermanshah, 1983). More specifically, activity locations center around frequently- and recently-visited locations such as home and work (Barbosa et al., 2018; Song et al., 2010a; Teixeira et al., 2021). There are rhythms of activity and travel patterns—people usually go to their workplaces and return home every day, inducing correlations between the days of the week. Other behavior like grocery shopping, going out with friends, or attending family gatherings happens less frequently, and hence has longer intervals between consecutive instances. These activities tend to be correlated between the weeks of the month and sometimes even the months of the year (i.e., Christmas, Thanksgiving, etc.).

To reflect these patterns, we incorporate categorical variables like days and weeks by representing them as sets of binary variables, using a one-of-k encoding. For example, as the days of the week x_d can take one of seven values, $x_d \in \{M,\ Tu,\ W,\ Th,\ F,\ Sa,\ Su\}$,

then a one-of-k encoding of x_d will correspond to seven binary inputs and one-of-k$(Tu) = [0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0]$. One approach to embed multiple inputs in a GP framework is to multiply kernels defined on each individual input. This family of kernels is called Automatic Relevance Determination (or ARD), so named due to the existence of a different lengthscale parameter for each input dimension d (Duvenaud, 2014). In this case, the kernel function $\mathbf{K}_{RQ}$ in Equation 2.6 becomes (for notational simplicity we drop the subscript i in $\mathbf{x}_i$):

$$k_{RQ-ARD}\left(\mathbf{x},\ \mathbf{x}^{'}\right) = \prod_{s=1}^{S} \Omega_s^2 \left(1 + \frac{\left(x_s - x_s^{'}\right)^2}{2\alpha l_s^2}\right)^{-\alpha} \tag{2.7}$$

where l_s is the lengthscale parameter for input dimension d. Thus, if the optimal lengthscale for a given categorical variable (i.e., day of the week) is small, the model has determined relatively low correlation between data in that category. On the other hand, categorical input dimensions with large lengthscales imply relatively little variation (in other words, high correlation) along those dimensions in the output variable (i.e., locations).

In addition to a one-of-k encoding strategy, we also use periodic kernels, which allow GPs to model functions that repeat themselves:

$$k_{PER}\left(x,\ x^{'}\right) = \Omega^2 \exp\left(-\frac{2\sin^2(\pi\left|x - x^{'}\right|/p)}{l^2}\right) \tag{2.8}$$

where the period length p determines the distance between repetitions of the function and the lengthscale l works in the same way as in the *RQ* kernel. The incorporation of the periodic kernel captures the various rhythmic patterns revealed by human mobility (people tend to conduct different activities and trips at varying frequencies). We note that the periodic kernel works best for continuous input dimensions that are unbounded in the input space.

2.3.3 Kernel and Mean Function Selection

Before prescribing a covariance function, we give some intuition on algebraic operations with kernels. The set of positive semi-definite kernels is closed under sum and product operations[2]. Multiplying two kernels can be interpreted as an AND operation while adding two kernels can be interpreted as an OR operation (Duvenaud, 2014). Let k_1 and k_2 be kernels which each

[2]For details on why this is the case, we refer the reader to Section 4.2.4 of Rasmussen and Williams (2006).

depend on a single input vector, $\mathbf{x}$ and $\mathbf{y}$, respectively. The product of k_1 and k_2 will result in a prior over functions that vary across both $\mathbf{x}$ and $\mathbf{y}$ and hence the function value $f(x_i, y_i)$ is only expected to be similar to some other function value $f(x_g, y_g)$ if x_i is close to x_g AND y_i is close to y_g. On the other hand, the sum $k_1 + k_2$ will result in a prior over functions which are a sum of one-dimensional functions, and hence the function $f(\mathbf{x}, \mathbf{y}) = f_x(\mathbf{x}) + f_y(\mathbf{y})$.

We design a composite kernel that is the sum of two nonlinear product kernels. We multiply the RQ-ARD kernel (introduced earlier) with a periodic kernel and then add a second identical component:

$$\mathbf{K}\left(\mathbf{x},\ \mathbf{x}^{'}\right) = \eta_1(\mathbf{K}_{RQ-ARD} \times \mathbf{K}_{PER}) + \eta_2(\mathbf{K}_{RQ-ARD} \times \mathbf{K}_{PER}) \tag{2.9}$$

where $\eta = \Omega^2$ denotes the weight of a kernel component.

As described by Duvenaud (2014), the product of a smooth kernel (like the RQ kernel) and the periodic kernel results in functions that are periodic but can slowly vary over the input space—that is, the shape of the repeating part of the function can change. This is an appropriate kernel to use as human mobility (as expressed in the way one's latitude and longitude changes over time) tends to be periodic (i.e., home-work-home routine), but with slight variations over time (Kitamura and Van Der Hoorn, 1987; Song et al., 2010b; Teixeira et al., 2021).

The two identical product terms in Equation 2.9 are to capture periodic patterns with different scales: the first periodic kernel is initialized with $p = 24$ hours (aiming to capture the home-to-work-to-home mobility pattern), while the second is initialized with $p = 7$ days (aiming to capture the weekly ebbs and flows of mobility). η_1 and η_2 are weights associated with the two product terms, respectively. When $\eta_1 = \eta_2$, the model places equal emphasis on the daily and weekly components. When $\eta_1 >> \eta_2$, the model identifies that the daily pattern is much more pronounced one than the weekly one and vice versa for $\eta_1 << \eta_2$. We constrain the weights such that their sum is always equal to one. We initialize the two components with different period lengths. In estimation, these two weights are treated as hyperparameters, like the lengthscale of a kernel component. Therefore, they are estimated in the same optimization procedure as described in Appendix A.1.

Furthermore, we use a constant mean function, for which the baseline value is set as the median value of the training set. The median latitude and longitude tend to correspond to (or are close to) a person's home location, as one's activity pattern tends to evolve around home. Thus, that becomes our model's baseline prediction in the absence of any information captured

in the covariance function. Compared to other mean functions (e.g., linear, multivariate orthogonal polynomial), the constant mean is a safe choice for data-driven models—in intra-trip inference, a majority of the prediction task is related to points temporally close to the training set and hence the covariance function explains most of the variation in the predictions. Our experiments with other mean functions showed that they tend to produce irrational predictions once the distance between the testing and training sets exceeds a certain threshold.

2.3.4 Model Training

In training our model, the goal is to optimally determine the set of hyperparameters Θ that minimize error with respect to the training data. The most straightforward method is to do so by minimizing the negative marginal log-likelihood (Rasmussen and Williams, 2006).

$$-\log p\left(\mathbf{Y} \mid \mathbf{X}, \Theta\right) = -\frac{1}{2} vec\left(\mathbf{Y}\right)^T \mathbf{\Sigma}^{-1} vec(\mathbf{Y}) - \frac{1}{2}\log|\mathbf{\Sigma}| - n\log\left(2\pi\right), \tag{2.10}$$

where Θ is the set of model parameters and $\mathbf{\Sigma} = \mathbf{K}^f \bigotimes \mathbf{K}^x + \mathbf{D} \bigotimes \mathbf{I}$ and $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix. Equation 2.10 provides a target function for kernel learning. The first component here estimates the model fit, while the second and third terms act as the regularization (Rasmussen and Ghahramani, 2000). Practically, the inverse term in the first component of this equation may pose a computational barrier when the size of the training data (and hence the kernel matrix) becomes too large. Traditional exact GP inference scales in $\mathcal{O}(n^3)$, while newer implementations take advantage of batched linear conjugate gradients to reduce this to $\mathcal{O}(n^2)$ (Gardner et al., 2018).

Various approximation methods have been proposed to overcome the computational limits of GPs on large datasets. Sparse Gaussian Processes (SGPs), for example, ease the computational burden by employing a subset of the training data ("inducing points") to reduce the dimensionality of the covariance matrix (Titsias, 2009). The key to the success of these methods is the careful selection of inducing points, which must be informative enough to explain most of the variation in the training data. Popular SGP methods use gradient-based optimization algorithms to choose optimal inducing points—these approaches run into problems when considering a large inducing point set size and/or high dimensional input spaces (Snelson and Ghahramani, 2005).

For passively-generated mobile data, most of the variation is generated from users' taking trips that differ from each other in a number of dimensions. Specifically, destination and

Algorithm 1: Trajectory Data Compression

Input: Raw trajectory data of one or multiple users; spatial radius parameter r
Output: Compressed trajectory data
LenTraj $\leftarrow$ Count the total number of observations in the raw trajectory;
Sorting: Sort by 'User ID' and 'Datetime';
Initialization: Create lists $[\phi]$ and $[\lambda]$, initializing them with ϕ_i and λ_i. Initialize $i = 0$ and $j = 1$;
for $i < LenTraj$ **do**
  for $j < LenTraj + 1$ **do**
    $d_{ij} \leftarrow$ Measure Haversine distance between (ϕ_i, λ_i) and (ϕ_j, λ_j);
    if $d_{ij} > r$ **then**
      break // End current for loop;
    end
    Add ϕ_j and λ_j to lists $[\phi]$ and $[\lambda]$;
    $j \leftarrow j + 1$;
  end
  $(\phi, \lambda) \leftarrow$ np.median($[\phi]$), np.median($[\lambda]$) // Replace each point in $[\phi]$ and $[\lambda]$ with the median point of each list;
  $i \leftarrow i + j$;
  $j \leftarrow i + 1$ // Update indices so that the compressed points are skipped in the next iteration;
end
return *The compressed trajectory data*;

mode choice, as well as departure time, are the three primary dimensions in which individuals' travel patterns vary. A wide range of literature has proposed parametric models to estimate these variables (Abkowitz, 1981; Daisy et al., 2018; Kitamura, 1988). In our model, we relate the variance of these variables to changes in a high-dimensional time matrix through the covariance function. However, we also do not need each available data point while training the model—many observations are close to each other in space, particularly if the sampling rate for a given period is high.

We reduce the size of trajectory data through a batched compression algorithm that aims to generate an approximated trajectory largely retaining the shape of the original trajectory. This algorithm is a variation of the well-known Douglas-Peucker algorithm (Douglas and Peucker, 1973), though instead of specifying an error requirement, we specify a Haversine distance—if multiple points are within a certain distance of one another, we replace them

using the median of those points. We chose this variation as it was faster to implement and used a more interpretable parameter based on distance. Algorithm 1 describes the full algorithm, which achieves the same outcome as a gradient-based optimization algorithm that carefully chooses training points which capture most of the variability. This reduces the computational burden of our model.

2.4 Implementation

We implement our algorithm and data analysis in Python. Specifically, we leverage GPyTorch, an efficient and modular implementation of GPs, as well as scikit-mobility, a data processing framework for GPS traces in the context of human mobility (Gardner et al., 2018; Pappalardo et al., 2019). The entire library of functions and classes we develop can be found at https://github.com/ekinugurel/GPSImpute. The requirements.txt file in the repository lists the required packages and their versions to run our program.

2.4.1 Data Preprocessing: Oscillation Correction, Noise Filtering, and Coordinate Normalization

We begin by preprocessing raw GPS traces to remove noisy data points—specifically, we filter by maximum velocity, using 200 km/h as the upper limit. Because segment velocities are calculated "as the crow flies" in our analysis (and hence expected to be smaller than the true value), a more conservative limit was appropriate (we provide more thoughts on this choice in Appendix A.5). This erases many of the oscillations or physically infeasible jumps that may be observed in mobile data due to the urban canyon effect. We then remove noisy points from the raw data by excluding observations more than 300 meters in precision (this is the location radius for which the data provider has 95% confidence). In practice, the potential of these noisy points to provide previously unobserved location information is shadowed by the problems they cause for model calibration, disrupting the continuity of smooth trajectories. Finally, prior to training the model, we normalize each user's training coordinates such that they have a mean of 0 and a variance of 1.

2.4.2 Gap Imputation: Input Specification

As described under Section 2.3.2, we represent time in multiple ways to encode the complexities of human mobility patterns (e.g., history dependency, periodicity) into the multi-task GP

Table 2.1: Temporal dimensions used in the robustness experiments.

Variable	Notation	Type	Model Inputs	Code Notation
Unix Time (normalized)	t_u	Continuous	$[0, 1, \ldots, n]$	`unix_min`
Second of the Day Sine	t_{ss}	Continuous	$[0, \ldots, 1]$	`sam_sin`
Second of the Day Cosine	t_{sc}	Continuous	$[0, \ldots, 1]$	`sam_cos`
Day of Week	t_d	Categorical	$[0, 1, 2, 3, 4, 5, 6]$	`dow`
Public Holiday	t_h	Binary	$[0, 1]$	`holiday`
Weekend	t_{we}	Binary	$[0, 1]$	`weekend`
AM Peak	t_{am}	Binary	$[0, 1]$	`am_peak`
PM Peak	t_{pm}	Binary	$[0, 1]$	`pm_peak`

model. We use a combination of continuous, categorical, and binary variables. The continuous variables include a monotonically-increasing Unix time, normalized to be 0 at users' first data point, as well as the number of seconds elapsed after midnight in a given day. Furthermore, we represent the days of the week using a one-of-k encoding as previously described, and they make up our categorical variables. Finally, we use binary variables to denote whether the given day is a public holiday, whether it is a weekend, and whether it is included in the AM or PM peaks (defined as 7-10am and 3-6pm, respectively). Table 2.1 summarizes the temporal dimensions used in our experiments.

With these in mind, we briefly remark on the previously-introduced kernel function (in Equation 2.9). The RQ-ARD kernels are specified to fit every input dimension (i.e. produce a unique lengthscale for each input vector), while the PER kernels fit only a monotonically-increasing Unix time variable $\mathbf{t}_u$, the only continuous and unbounded input in our input space. Also note that the PER kernels are multiplied by the RQ-ARD kernels—this can be loosely thought of as an AND operation, rather than an OR operation (which corresponds to addition). Therefore, the periodic kernels are only meant to supplement the hidden structures captured through the specification of additional temporal variables, not replace them.

2.5 Dataset

We employ privacy-protected, passively-generated mobile data from Spectus, a U.S.-based data solution provider specializing in geospatial analytics. The dataset contains time-stamped location traces of $2,000$ anonymous, opted-in individuals in the Greater Seattle Area between December 2019 and July 2020. More specifically, the data includes timestamps, unique device

identifiers, latitude and longitude coordinates, and a measure of data precision (i.e., a location radius for which we have 95% confidence). The dataset does not include any demographic or socioeconomic information. Locational data is sent to Spectus servers in encrypted form, via an HTTPS protocol through three options: an Application Protocol Interface (API), through a publisher that has licensed Spectus' software development kit, or via direct server-to-server integration (Spectus, 2022).

Spectus enhances users' privacy through two methods. First, they remove data from locations that do not meet privacy standards determined by their Sensitive Points of Interest (SPOI) policy. Though not an exhaustive list, the following is an example of data Spectus does not use for attribution: health-related, locations with vulnerable populations, sensitive lawful businesses, military-related, locations with first responders, correctional facilities, locations with firearms, churches/religious facilities, Native American reservations, sexual orientation-related, adult-oriented entertainment, and social demonstrations (Spectus, 2022). Second, Spectus anonymizes data in around home locations using patent-pending technology. What this means is that traces near identified home locations are replaced by points at the centroid of the address' census block group (CBG), thereby not revealing one's actual residence. CBGs are the second smallest geographical unit for which the U.S. Census Bureau publishes sample data, and they contain a nationwide average of 51 blocks.

2.5.1 Defining Missingness: Temporal Occupancy

Because Spectus' datasets are passively-generated, the gaps between any two adjacent observations are rarely equal in length. Thus, we need a convention to mathematically denote varying levels of missingness in mobile data. We discretize a user's total available data time $\mathcal{T}$ into P intervals of length τ, which we refer to as the "temporal resolution." The choice of τ is important—it decides the sparseness of a user's observed trajectory, in which each interval is assigned an indicator variable

$$I_p = \begin{cases} 1 & \text{if } p \text{ has at least one observation} \\ 0 & \text{otherwise} \end{cases}$$

We thus define temporal occupancy (or the inverse of sparsity) as

$$q_\tau = \frac{1}{P} \sum_{p=1}^{P} I_p \tag{2.11}$$

Note that $0 \leq q_\tau \leq 1$. A period with high q has fewer gaps in data and vice versa. For example, a period with $q_{30} = 0.9$ suggests that there exists at least one observation in 90% of the 30-minute intervals.

2.5.2 Descriptive Analysis: Quantifying Missingness

We provide some brief statistics on variables in the Spectus data, shown on Figure 2.2. Most data points are highly precise—95 percent of all traces are within a 65-meter radius of their true location (top left). This is a significant change from a 2018 study looking at a similar dataset, in which about 7 percent of all observations had a precision less than $1,000$ meters (Ban et al., 2018). In terms of the distribution of observations throughout the day (top right), our findings are consistent with that of Ban et al. (2018). We observe two peaks on weekdays, one in the morning (around 7am) and one in the evening (around 6pm). For weekends, there is one mid-day peak. For every day, there is a positive trend on the number of observations as the hours progress.

As expected, a mobile dataset's level of "missingness" depends on the temporal resolution one chooses. The bottom right figure in Figure 2.2 shows a boxplot of individual temporal occupancies of $2,000$ users from the Spectus dataset for three different temporal resolutions. As the temporal interval τ increases, so does the average temporal occupancy. The bottom left figure in Figure 2.2 shows that the gaps are not always short either—roughly 40 percent of all users recorded in Spectus data have at least one week that is continuously missing (i.e., there are no observations for a whole week). Meanwhile, 96 percent of all users have at least one day of missingness, and 99 percent have at least one six-hour period of missingness. Thus, we anticipate the need to prescribe a method to correct missingness in the general sense, whether the gaps are short or long.

2.6 Experiments

We do two sets of experiments to showcase our model. The first set explores the parameter space of multi-task GPs in the context of human mobility. Our goal here is to understand what model parameters work well for mobile data with certain characteristics (i.e., highway drives vs. urban walks), a benefit of which is developing a greater intuition on how to initialize GP model parameters in different contexts. The second set of experiments benchmarks our model against alternative missing data imputation methods, including simple exponential smoothing, exponential smoothing (Huo et al., 2010), Holt-Winters (Cipra et al., 1995),

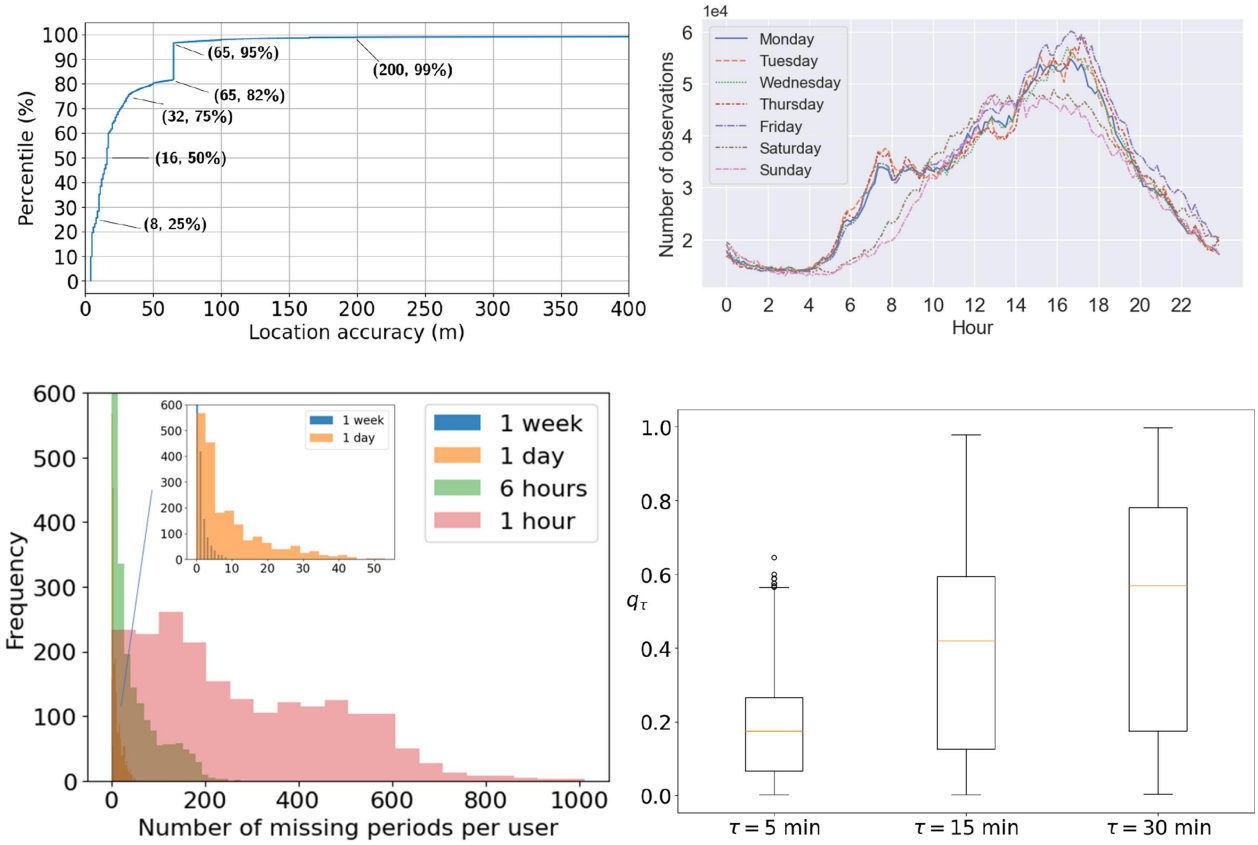


Figure 2.2: Descriptive analysis of Spectus data for all 2,000 users. (Top Left) Cumulative distribution of location accuracy. (Top Right) Distribution of observations within each day of the week. (Bottom Left) Histogram of gaps categorized by size. (Bottom Right) Boxplots of temporal occupancy with three temporal resolutions.

ARIMA and SARIMAX (Kohn and Ansley, 1986), as well as a multi-task GP using the standard RBF-ARD kernel on each input dimension (Duvenaud, 2014).

For the first set of experiments (Section 2.6.1), we first preprocess raw GPS traces to obtain two datasets for 10 randomly chosen users (1920 trips total): (1) a set showing all trip-related data points for a user, and (2) a set containing various trip information, including average velocity, total distance, total duration, among other metrics. We fit GPs to trip-related data points to predict missing locations. Then, we relate the optimized model parameter values with respect to the compressed trip information, which allows us to infer the type of trip (i.e., highway drive or a walk). For the second set (Section 2.6.2), we randomly select a subset of 50 users with at least $10,000$ observations and at least 3 months of data (from the first data point until the last). Within the selected 50, we only retain data from January and February 2020 to reduce the number of data points. We use this longitudinal data to train the model and make predictions on varying gap lengths.

2.6.1 Lengthscale Analysis

In this experiment, we determine well-fitting model parameters (i.e., ones that minimize the loss function) and relate them to a range of data characteristics, including average speed, total distance, and trip duration among other metrics. Identifying optimal kernel parameters is important for model accuracy. Figure 2.3 illustrates how a suboptimal lengthscale can deteriorate model fit. The left subplot shows a trip with noisy training data and a lengthscale value that is too low. The prediction's mean reverts to the constant mean (likely near zero) only a few seconds after a training observation.

Furthermore, the optimization process deduces high levels of noise, producing uncertainty even on training inputs. Both processes result in low prediction accuracy as the prediction mean remains distant from the testing set. The right subplot, on the other hand, shows a model with low levels of noise and an appropriate lengthscale for the same trip. The prediction mean shifts smoothly between the different trip segments and the confidence interval is narrow near training points.

Using preprocessed Spectus data, we leverage k-means clustering to group together similar trips using trip characteristics (shown in the columns of Table 2.2). We filter a subset of mobile data based on three criteria, which help avoid anomalies like airplane-based travel and out-of-state roadtrips: average velocity does not exceed 80 mph, total travel time does 6 hours, and the number of observations (in a trip) is greater than 4. Using the elbow method,

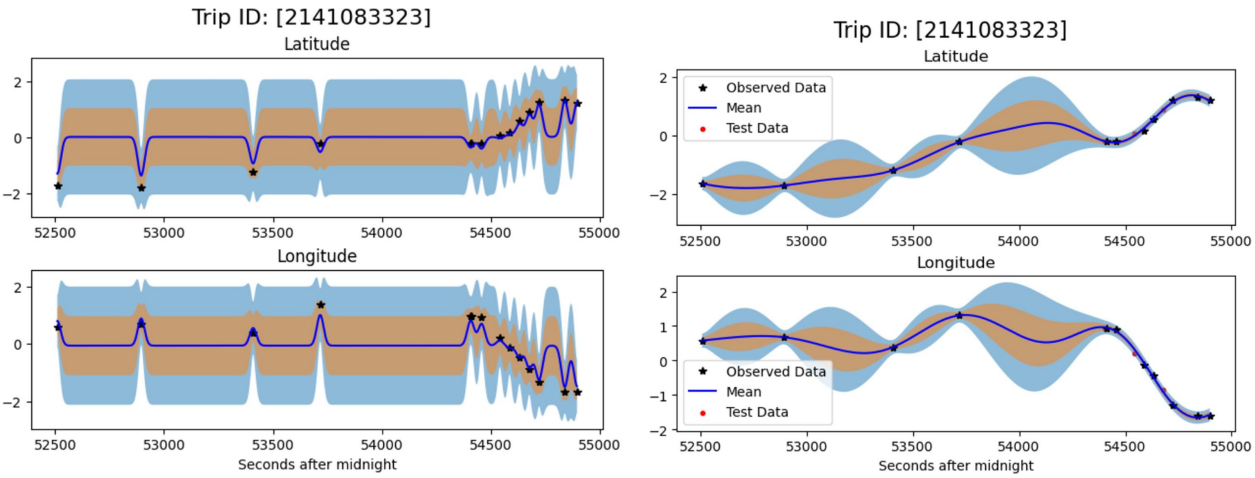


Figure 2.3: Examples of trips fit with an underestimated lengthscale (left) and appropriate lengthscale (right)

we identify three clusters from the filtered data as sufficient to explain most of the variation in trip-level data. Table 2.2 summarizes the average mobility metrics associated with the trips in the different clusters. The heading change rate shows the ratio of consecutive data points where a user changes direction with an angle exceeding a threshold (we use 0.33 rad). The velocity change rate shows the ratio of consecutive data points where the user exceeds a speed variation threshold (we use 26%). Finally, the stop rate represents the ratio of data points with an inferred velocity lower than a threshold (we use $0.89\frac{m}{s}$). These threshold values are chosen in accordance with findings from Zheng et al.[3] (2008) and are meant to discriminate between moving and being stationary.

We only used local data (i.e., observations from the same trip) in fitting GP models for these trips and thus our only input is $\mathbf{t}_u$ (normalized Unix time). Accordingly, we use the RBF kernel, a simpler form of the earlier-defined RQ kernel:

$$k_{RBF}(x,\ x^{'}) = \Omega^2 \exp\left(-\frac{\left(x-x^{'}\right)^2}{2l^2}\right) \tag{2.12}$$

Figure 2.4a reveals that walking trips exhibit the lowest average RMSE, indicating higher prediction accuracy, whereas vehicle trips present the highest RMSE. Correspondingly, Figure 2.4b's boxplot of optimized lengthscale parameters for each trip cluster indicates that faster

[3]See "Parameter Selection" on Zheng et al. (2008)

Table 2.2: Summary of trip clusters.

Cluster	**Avg. Vel. [m/s]**	**Distance [m]**	**Duration [s]**	**Heading Change Rate**	**Velocity Change Rate**	**Obs.**	**Stop Rate**
Slower, shorter trips	9.29	$8k$	1062	2e-3	2e-3	22.8	7e-4
Medium speed/ distance trips	13.9	$30k$	2362	7e-4	8e-4	49.9	2e-4
Fast, distant trips	17.9	$60k$	3449	5e-4	6e-4	142	1e-4

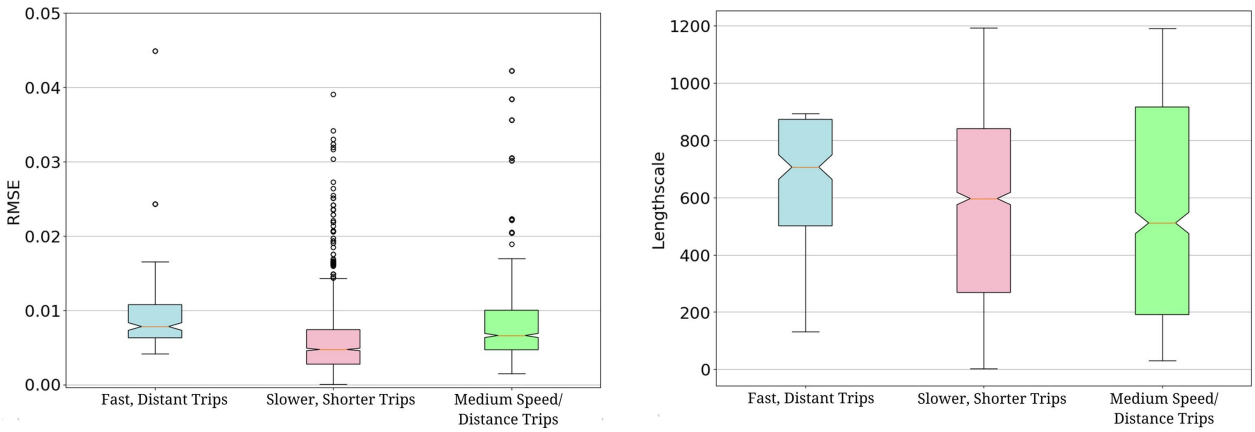


Figure 2.4: (Left) Boxplot of total RMSE for trips in different mobility metric clusters (Right) Boxplot of optimized lengthscales for each trip cluster.

trips, typically associated with vehicles, necessitate larger lengthscales, reflecting less frequent changes in mobility patterns. In contrast, walking trips and mixed trips generally require shorter lengthscales, suggestive of more variable and less predictable movements. Notably, mixed trips display the widest range of lengthscales, underscoring a significant variability in temporal correlation. This can be ascribed to the heterogeneity of mixed trips, which often involve a combination of transportation modes, each with distinct velocities and movement patterns. Such diversity introduces numerous non-smooth transitions, or 'kinks,' into the data, which the Gaussian Process (GP) must adeptly capture.

The variability in lengthscales observed for walking trips may be attributed to the wide

spectrum of walking behaviors. On one hand, some individuals walk with a clear destination and route in mind, resulting in relatively predictable paths. On the other, walking can be exploratory, with no pre-determined endpoint, leading to more stochastic patterns featuring frequent stops and directional changes. This dichotomy in walking behaviors is corroborated by the high heading change and stop rates for walking trips as detailed in Table 2.2. Despite the complexities introduced by such diverse activities, both walking and mixed trips achieve lower median RMSE values, suggesting that the GP model is effectively modeling these trip types. This indicates that the GP framework is robust enough to handle the stochasticity inherent in walking trips and the complexity of mixed trips, accurately reflecting the nuanced shifts in mobility without a significant loss in predictive performance.

2.6.2 Robustness Checks: Gap Imputation

The goal of this experiment is to assess the performance of our model against other imputation methods in a variety of missingness conditions. In doing so, we conduct a robustness analysis on the gap length. Our benchmark methods include simple exponential smoothing (SES), exponential smoothing (ES), the Holt-Winters (Holt) method, ARIMA, SARIMAX, and an RBF-ARD kernel. The first three (ES, SES, and Holt) work similarly—they weigh the average of past observations in producing forecasts, giving exponentially lower weights as the time gap increases. SES is most suitable for forecasting data with no clear trend or seasonal pattern, while ES allows for forecasting with a trend. The Holt-Winters method uses three types of exponential smoothing to model the level (the typical value), the trend (the slope), and the seasonality (repeating patterns) of data. ARIMA and SARIMAX are both linear regression models fit for forecasting univariate time series data. While ARIMA can handle data with a trend, its extension SARIMAX can also handle exogenous variables and seasonal components. Each of these methods is regarded as a reliable, well-studied alternative commonly used to impute missing values in time series, and therefore make a useful comparison. The inclusion of the RBF-ARD kernel is to demonstrate that an off-the-shelf kernel implementation does not perform as well as our composite formulation. Table A.1 in Appendix A.4 describes and quantifies the parameters of each method.

Our results demonstrate that our model outperforms existing methods in all scenarios with varying levels of missingness. Specifically, we simulate gaps in the selected 50 users by reserving a subset of their data for testing, and we choose this subset such that the temporal occupancy of their training data meets a lower target temporal occupancy (i.e., a decimal

Algorithm 2: Gap Simulation in Trajectory Data

Input: Trajectory data; Bin length τ
Output: Trajectory data with gaps; New q_τ
bins ← Create an array of integers spanning from $t_{u,1}$ to $t_{u,N}$ with step size τ;
bins_dict ← Use dictionary comprehension to map each data point to the relevant bin;
non_empty_bins_dict ← Make a similar dictionary with only the non-empty bins;
target_ocp ← `np.random.uniform`$(0, q_{curr})$ // Take decimal floor of randomly chosen value between 0 and the current temporal occupancy as the target occupancy;
while $q_\tau >$ *target_ocp* **do**
  `np.random.choice(non_empty_bins_dict.keys())` ← Randomly choose a bin to remove and remove all values in this bin from original data;
  bins_dict ← Update the original dictionary;
  non_empty_bins_dict ← Update the non-empty bins dictionary;
  new_ocp ← Calculate the new temporal occupancy with the gapped data;
end
return *The gapped trajectory data as well as the new* q_τ;

between 0 and the current temporal occupancy) according to the temporal resolution being tested. We test six different temporal resolutions (τ): one week, one day, six hours, one hour, thirty minutes, and fifteen minutes. Algorithm 2 describes the full process to remove points from users' data[4].

In the context of plotting and analyzing the results of various models with differing levels of data sparsity, we utilized Dynamic Time Warping (DTW) as an ancillary metric. DTW, a method established by (Müller, 2007), is suitable for measuring the similarity between two temporal sequences, accommodating variations in speed and timing. DTW values give us an indication of the dissimilarity between sequences; a lower DTW value indicates a closer match, whereas a higher value suggests a greater disparity between the compared sequences.

From the analysis depicted in Figure 2.5, we discern an inverse-linear relationship between temporal occupancy and DTW values, indicating that as temporal occupancy diminishes, DTW values—reflecting the disparity between testing set outcomes and model predictions—correspondingly escalate. This trend illustrates a progressive, rather than precipitous, decline in the model's predictive accuracy with sparser datasets. Notably, this inverse-linear relationship appears less distinct at smaller gap lengths, as illustrated when comparing Figure 2.5a for $\tau = 1$ week and Figure 2.5d for $\tau = 5$ minutes, suggesting a more stable model

[4]We also provide a step-by-step breakdown of this algorithm in Appendix A.2

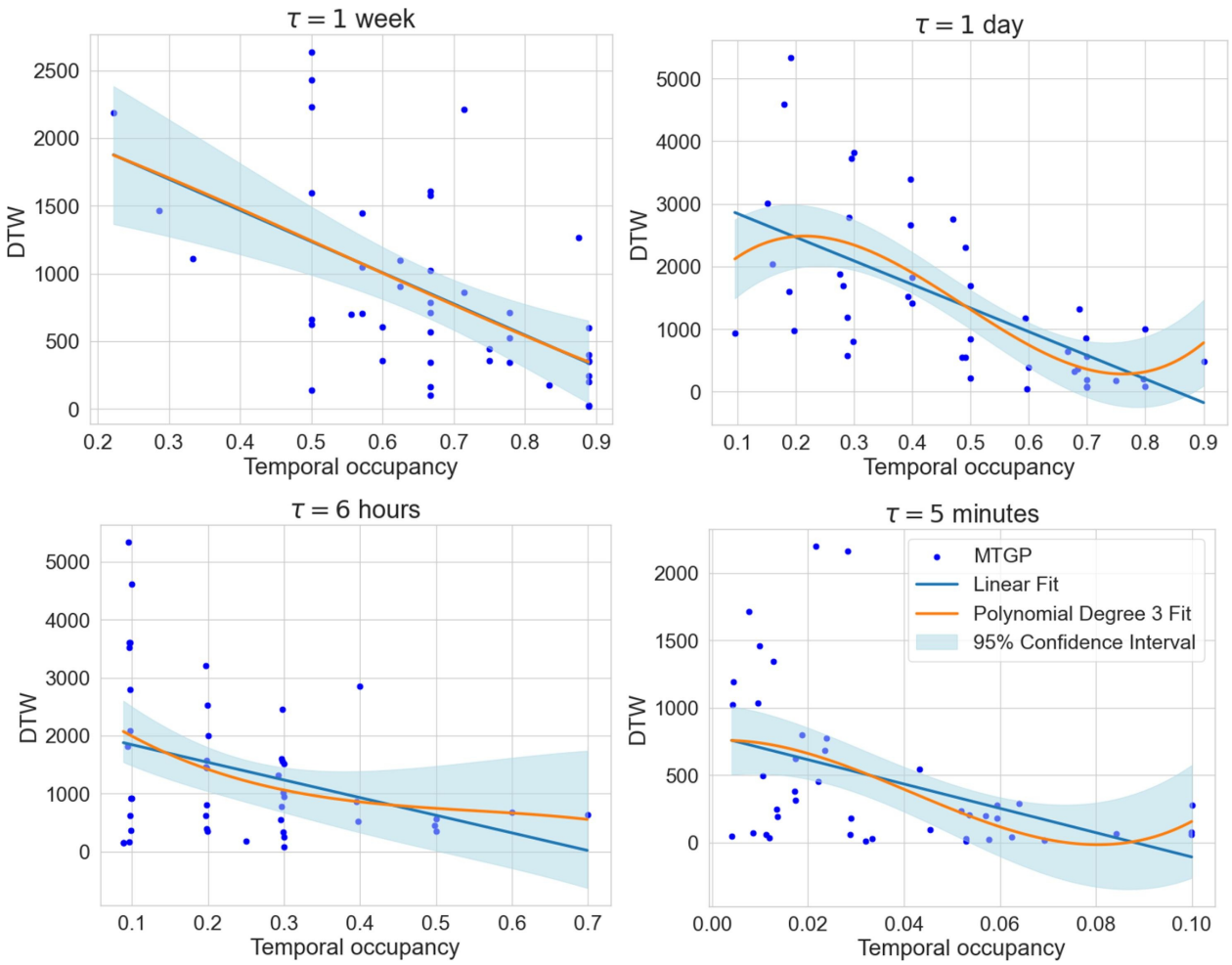


Figure 2.5: Scatterplots of DTW and temporal occupancy for various gap lengths. (a) $\tau = 1$ week, (b) $\tau = 1$ day, (c) $\tau = 6$ hours, (d) $\tau = 5$ minutes.

performance within these intervals. Despite the non-linear trends at the extremities of the temporal occupancies, potentially stemming from limited sample sizes, we do not observe a specific threshold beyond which the model's performance deteriorates drastically. Instead, the model's accuracy gradually tapers off with reduced temporal occupancy. This analysis emphasizes the model's consistent performance across different scenarios while indicating that ensuring a minimum level of temporal occupancy—adjusted according to the specific temporal resolution of the data—is crucial for achieving the best predictive accuracy from the model.

For the benchmark algorithms, we used only the monotonically increasing $\mathbf{t}_u$ as input since they do not have a straightforward way to deal with categorical or binary variables. For each individual, we trained an optimal benchmark model by maximum likelihood estimation (MLE) whenever applicable. In cases when we could not use MLE (i.e., determining the order parameters for ARIMA and SARIMAX), we used the grid search method to identify the optimal parameters (see Table A.2 in Appendix A.4).

We analyzed the accuracy of derived mobility metrics after the imputation of gapped training data compared to before simulating any gaps. Table 2.3 presents the median error results in each bin length. The boxplots of these error results can be found in Appendix A.4 (Figures A.1-A.7). The same 50 users' data was evaluated across the different gap lengths, allowing for an apples-to-apples comparison. In all tests, our model outperformed the other models: The average error was always the lowest, and the 5th and 95th percentile bounds tended to be closer. The metrics we compared across the different methods included the number of distinct locations, the radius of gyration, the straight-line distance traveled, real entropy, random entropy, and uncorrelated entropy. The number of distinct locations simply counted how many unique combinations of coordinate pairs existed in the imputed and original datasets. The straight-line travel distance traveled by an individual is computed as the sum of the distances traveled (Williams et al., 2015) and can be computed as

$$d_{SL} = \sum_{j=2}^{m} dist(r_{j-1}, r_j), \tag{2.13}$$

The radius of gyration indicates the characteristic travel distance of a user during a period

(Song et al., 2010b) and is given by the equation

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2}, \tag{2.14}$$

where r_i represents the $i = 1, \ldots, n$ locations recorded for the user and $r_{cm} = \frac{1}{n}\sum_{i=1}^{n} r_i$ is the center of mass of the period's trajectory. Finally, the concept of entropy is used to assess a user's predictability. The real entropy is given by the equation (Song et al., 2010b)

$$E_{real} = -\sum_{T} P(T') \log_2(P(T^i)) \tag{2.15}$$

where $P(T')$ is the probability of finding a particular time-ordered subsequence T' in the trajectory T. Therefore, the real entropy depends not only on the frequency of visitation, but also the order in which the locations were visited, and the time spent at each location, thus capturing the entire spatiotemporal order present in a user's mobility. The random entropy captures a user's degree of predictability if each of their distinct locations were visited with equal probability. The (temporal) uncorrelated entropy characterizes the heterogeneity of a user's visitation patterns (Eagle and Pentland, 2009). The two are given by the equations

$$E_{rand} = \log_2(L) \tag{2.16}$$

$$E_{unc} = -\sum_{j=1}^{L} P(j) \log_2(P(j)) \tag{2.17}$$

where L is the number of distinct locations a user visits and $P(j)$ is the historical probability that a location was visited by the user.

We see that, on average, our model outperforms the rest of the benchmarks by a significant margin in estimating the number of locations visited, the radius of gyration, the straight-line travel distance, real entropy, random entropy, and uncorrelated entropy. Moreover, on these metrics, the distance between our model and the second-best model is significant. The standard RBF kernel multiplied across all input dimensions does not capture the variability of trips in sparse regions, performing worse than simpler smoothing algorithms even in low sparsity regimes. This highlights the importance of using appropriate kernel specifications while dealing with high-dimensional, mixed-type datasets.

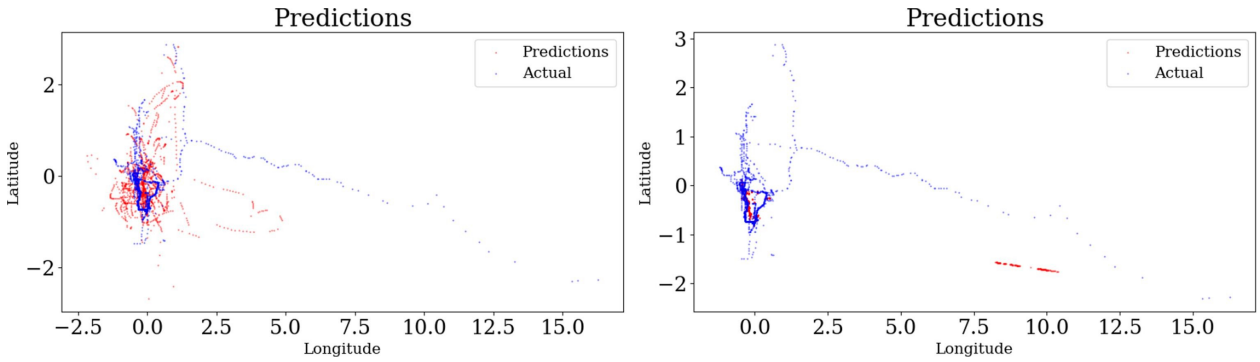


Figure 2.6: Predictions (red) and ground truth test data (blue) for a user tested at $\tau = 6$ hours. MTGP with the proposed kernel (left) outperforms exponential smoothing (right).

Figure 2.6 offers some perspective on the results presented here. In this example, a perfect predictor would have the red points align exactly with the blue points. The MTGP method (with the proposed kernel) better captures the variability of a user's location over time than Exponential Smoothing—hence why it is more accurate in metrics like straight-line travel distance (Equation 2.13) and radius of gyration (Equation 2.14). We also note that, in general, MTGP seems less accurate in capturing route choice of users, but more accurate in learning destination choice. This is what we attribute to the performance we see on metrics like real and uncorrelated entropy (Equations 2.15 and 2.17, respectively).

2.7 Discussion

The use of passively-generated mobile data for transportation applications and decision-making is an irreversible trend. And yet, such data exhibits significant and varying levels of missingness compared to actively solicited data such as household travel surveys with GPS loggers or smartphone data collection components. Left untreated, the resulting mobility metrics are biased (McCool et al., 2022). To correct missingness in the mobile data, we have developed a novel framework using multi-task Gaussian processes. More specifically, our framework leveraged the correlations between users' coordinates (thus multi-task), and allowed individual-level differences in data characteristics resulting from fundamentally different trip generation processes (e.g., different modes of transportation used). We introduced and demonstrated the effectiveness of RQ and PER kernels in the context of mobile data for human mobility modeling, and highlighted the ARD extension of these kernels for fitting

Table 2.3: Median error with respect to the testing sets.

Time Gap	Method	Number of Locations	Radius of Gyration	Straight Line Travel Distance	Random Entropy	Real Entropy	Uncorr. Entropy
	MTGP	26	-0.07	205.029	0.045	0.278	0.153
	RBF	-801	-0.835	-888.441	-9.647	-9.323	-9.527
	SES	-801	-0.835	-888.441	-9.574	-9.323	-9.527
1 week	Holt	-801	-0.835	-888.441	-9.574	-9.323	-9.527
	ES	-778	-0.621	-643.909	-4.989	-7.726	-4.955
	ARIMA	-801	-0.835	-888.441	-9.276	-9.323	-9.527
	SARIMAX	-801	-0.835	-888.441	-9.647	-9.323	-9.527
	MTGP	33	-0.245	236.805	0.036	0.227	0.117
	RBF	-1050	-0.909	-1303.06	-10.038	-9.612	-9.806
	SES	-1050	-0.871	-1303.06	-9.309	-9.612	-9.806
1 day	Holt	-1027	-0.718	-768.678	-4.875	-7.907	-5.225
	ES	-1050	-0.834	-1303.06	-8.506	-9.609	-9.803
	ARIMA	-1050	-0.834	-1303.06	-8.506	-9.609	-9.803
	SARIMAX	-1050	-0.909	-1303.06	-10.038	-9.612	-9.806
	MTGP	34	-0.187	-13.641	0.042	0.237	0.155
	RBF	-956.5	-0.645	-1223.47	-9.809	-9.493	-9.751
	SES	-954	-0.645	-1177.23	-9.139	-9.493	-9.608
6 h	Holt	-929	-0.645	-1177.19	-4.895	-7.869	-5.066
	ES	-929	-0.718	-768.56	-4.895	-7.869	-5.066
	ARIMA	-954	-0.645	-1178.65	-8.317	-9.493	-9.608
	SARIMAX	-956.5	-0.645	-1223.47	-9.901	-9.493	-9.751
	MTGP	38	-0.074	389.303	0.049	0.257	0.157
	RBF	-990	-0.994	-1319.47	-9.818	-9.548	-9.761
	SES	-901.5	-0.761	-1262.95	-7.989	-9.546	-9.642
1 h	Holt	-878	-0.627	-1161.74	-6.875	-9.149	-5.174
	ES	-898.5	-0.761	-1262.76	-6.801	-9.546	-9.613
	ARIMA	-898.5	-0.761	-1262.76	-6.801	-9.546	-9.613
	SARIMAX	-830.5	-0.644	-1161.14	-6.435	-9.455	-9.449
	MTGP	22	-0.299	-7.116	0.048	0.323	0.161
	RBF	-670	-2.15	-1112.56	-8.925	-8.871	-9.312
	SES	-660	-2.17	-1162.11	-7.734	-8.871	-8.994
15 min	Holt	-660	-2.099	-1162.07	-6.435	-8.871	-8.849
	ES	-637	-2.056	-513.035	-3.906	-7.711	-4.497
	ARIMA	-591	-1.931	-1162.07	-6.435	-8.871	-8.851
	SARIMAX	-670	-2.15	-1119.07	-9.39	-8.871	-9.416

high-dimensional data (whether continuous or binary). Our experiments highlighted the importance of specifying a fitting composite kernel for our domain, as using the generic RBF kernel resulted in sub-optimal predictions due to not being able to capture multi-level periodicities. In the results, we showed that trips made by different modes are associated with different, optimally estimated kernel parameters (i.e., the lengthscale) and provided guidelines on different parameter initializations for different kinds of trips. More importantly, we showed that our model outperformed six existing methods for correcting missingness in mobile data and assessed the temporal sensitivity of our model, confirming that our model achieves enhanced predictive performance when trained on a diverse temporal dataset.

While the current study is motivated by the need to directly address the significant sparsity issue exhibited in passively-generated mobile data (McCool et al., 2022), the implications of this work can be significant in several aspects. For one, it directly addresses the bias issue induced by data sparsity and the resulting, corrected trajectory data can be viewed as pseudo-ground truth from which mobility metrics can be derived. In another, it opens us the opportunity to create city-wide simulations of human mobility patterns using the generated mobile data directly. This is in contrast with the current process where parametric models are first estimated using the household travel survey data (which often represents less than 1% of a region's population) and then those models are extrapolated to a synthetic population. As noted earlier, the latter method is not scalable as travel surveys must be frequently collected and models must be updated. An added disadvantage is that those models do a poor job of capturing the nonlinearities embedded in the data.

Transportation behaviors, whether they refer to individual travel behaviors as studied in this paper or community- or region-level traffic phenomena, are highly complex and non-linear. Thus, expecting a universal model to capture all is likely unrealistic. Instead, context-dependent modeling is much needed for the field of transportation. Context-dependent modeling requires the development of flexible modeling frameworks that can capture not only heterogeneity but also adapt to changing contexts. We demonstrate that GP-based methods are one suitable framework for this purpose in the context of modeling individual mobility patterns.

Our work underscores the critical role of kernel specification in capturing the nuances exhibited by individual mobility patterns. In our study, we used a domain-based approach, i.e., using our domain knowledge in the travel behavior literature to guide the selection and the composition of the kernels. Through our experiments, we show that standard, off-the-

shelf kernels, such as the Radial Basis Function (RBF) kernel, frequently underperform in comparison to more complex kernels across a range of imputation scenarios. The process of specifying kernels is akin to a traditional, parametric modeling process where one would have to start with a hypothesized model structure. One may argue that even in data-driven modeling, one would still have to start with a modeling framework or architecture that define how input data are taken in and evolves over different layers, and whether there are feedback loops or not. Just as these models often require rigorous cross-validation to determine the most suitable hyperparameters, Gaussian Processes similarly benefit from a structured approach to kernel selection. Addressing this issue, notable advancements have been made in the development of automated methods for kernel determination. Duvenaud et al. (2013) have proposed a method for automating the construction of kernel expressions, thereby facilitating the identification of suitable kernel structures directly from the data. In a similar vein, Wilson and Adams (2013) introduced a framework that learns expressive covariance functions for GPs through spectral mixture kernels capable of automatically adapting to the structure of the data. These innovations are significant steps forward in simplifying the kernel selection process, thereby enhancing both the performance and the broader applicability of the model across various domains.

Nevertheless, our study also recognizes certain limitations. One limitation is the challenge in selecting kernels that accurately emulate the data's inherent features, which is compounded by the issue of model interpretability. Although our model offers improved interpretability over deep learning (DL) models, which are often criticized for their 'black box' nature, it lacks the transparency of simpler, more traditional statistical methods. The inherent complexity of Gaussian processes, particularly when employing sophisticated kernels such as Rational Quadratic (RQ) and Periodic (PER), introduces significant challenges in terms of interpretability. While model interpretability is not of paramount significance to our study (since the study objective is simply fixing missingness in raw mobile trajectories), potential applications of GPs in other transportation applications (such as those reviewed earlier in Section 2.2 for forecasting travel demand or traffic flows) would require higher levels of model interpretability. Thus, balancing model complexity with interpretability is an essential area for future research that involve GPs or other big data methods.

Another future direction of this work is to tackle computational complexity—the time it takes to train a model scales cubically with the size of the covariance matrix, which is a function of the size of the data. We outline a few ideas to explore this front. One is

to accelerate the process of GP parameter optimization through CUDA (Compute Unified Device Architecture), a parallel computing platform and API created by NVIDIA that allows the software to use certain types of graphical processing units (GPUs) for general-purpose processing. Doing so would allow future models to add additional layers in a GP framework (i.e., deep GPs), predicting covariates like velocity and bearing before producing a location output. We can also pre-process the input data to drastically reduce the number of times we have to re-evaluate Equation 2.10 above. An example is provided by Lee et al. (2017).

There is also the need to incorporate the underlying built environment including multi-modal road networks into the model, such that bodies of water, buildings, or locations of different types can be recognized. Methods like map-matching may be employed to post-process model results such that the predictions can only be made within allowable regions. However, this approach would have a trade-off with computational complexity—though scalable map-matching algorithms have been proposed in recent years (Fiedler et al., 2019; Zeidan et al., 2020), convenient implementations are not yet widely available. Alternatively, one may also explore additional selection and identification of composite kernels (like Equation 2.9) to potentially account for those contextual effects.

Application-wise, there are numerous ways that the corrected human trajectory data can be used. There are numerous applications of passively-generated mobile data in the existing literature, which are exclusively based on uncorrected trajectory data. It would thus be interesting to see whether the existing findings still stand if the corrected trajectory data is used. In a previous paper by one of the authors, we showed that not removing the oscillation phenomenon in the data leads to the non-negligible overestimation of the regularity of individuals' mobility (Wang and Chen, 2018). We similarly expect improvements in accuracy for downstream studies that use sparsity-corrected passively-generated mobile data. Future efforts could also be made to unify existing literature on applications of GPs to transportation problems, such as the works of Batista et al. (2022) and Gammelli et al. (2020), highlighting shared model properties, cautioning on common modeling mistakes, and establishing standards of benchmarking.

Chapter 3

LEARNING TO GENERATE SYNTHETIC HUMAN MOBILITY DATA: A PHYSICS-REGULARIZED GAUSSIAN PROCESS APPROACH BASED ON MULTIPLE KERNEL LEARNING

Passively-generated mobile data has grown increasingly popular in the travel behavior (or human mobility) literature. A relatively untapped potential for passively-generated mobile data is synthetic population generation, which is the basis for any large-scale simulations for purposes ranging from state monitoring, policy evaluation, and digital twins. And yet, this significant potential may be hindered by the growing sparsity or rate of missingness in the data, which stems from heightened privacy concerns among both data vendors and consumers (users of service platforms generating individual mobile data). To both fulfill the great potential and to address sparsity in the data, there is a need to develop a flexible and scalable model that can capture individual heterogeneity and adapt to changes in mobility patterns. We propose a conditional-generative Gaussian process framework that learns kernel structures characterizing individual mobile data and can provably replicate observed patterns. Our approach integrates physical knowledge to regularize the framework such that the generated data obeys constraints imposed by the built and natural environments (such as those on velocity and bearing). To capture travel behavior heterogeneity at the individual level, we propose a data-driven multiple kernel learning approach to determine the optimal composite kernel for every user. Our experiments demonstrate that: (1) the impact of kernel choice on mobility metrics derived from synthetic data is non-negligible; (2) physics-regularization not only reduces model bias but also improves uncertainty estimates associated with the predicted locations; and (3) the proposed method is robust and generalizes well to varying individuals and modes of travel.

3.1 Introduction

The ubiquity of sensor-equipped mobile devices has enabled the generation of passively-generated large-scale time- and location-stamped data (hereafter called "mobile data"), including GPS traces, geo-tagged posts from social media platforms, and call detail records

(Chen et al., 2016b). These datasets are often massive in size (covering millions of residents in and travelers passing through a region) and longitudinal that can last from days to months and even years. These two attributes are fundamentally different from household travel surveys which are snapshots of a region's travel patterns with a small sample (typically less than 1% of a region's population) and have been traditionally used for travel behavior and demand forecasting studies. These big mobile datasets offer great promises in moving travel behavior or human mobility studies forward in various applications. Existing studies of the past 15 years using these passively-generated mobile data have mostly focused on identifying models that can capture both regularities and variations of human mobility patterns (Gonzalez et al., 2008; Yuan and Raubal, 2012). The emergence of the COVID pandemic as well as AI in the last several years have stimulated additional studies that seek to identify cause and effect relationships (e.g., whether a change in the built environment leads to changes in mobility patterns; Wang et al., 2021a) and to understand changes in mobility patterns to support timely policy making (e.g., social distancing policy-making during pandemic requires rapid assessment of changes in people's mobility patterns; Morris et al., 2021).

An emerging application of using passively-generated mobile data is synthetic population generation for large-scale simulation of individual mobility patterns (Deng et al., 2021), whose applications are wide-ranging from state monitoring, to whole-region or targeted policy making, and long-term investment decisions. Using passively-generated mobile data, which are not only massive but also longitudinal, for this purpose has the potential to transform the current state of the art in the field that are primarily of two kinds. The first kind is the long-standing Activity-based Models (ABM) which are essentially a set of linked econometric models whose parameters are estimated using household travel survey data (Bhat and Koppelman, 1999; Pendyala et al., 1997; Timmermans and Arentze, 2011). Estimated econometric models reflect average trends and thus direct application of those models at the individual level result in large amounts of errors. Furthermore, ABM models are often estimated on small cross-sectional datasets and thus are static and cannot capture how behaviors evolve over time. The second kind are data-driven models that leverage population-level probabilistic distributions estimated from the big mobile data. These models capture individual heterogeneity through the expression of a distribution and fundamentally they cannot account for or adapt to nuances that are unique to an individual. These nuances may include factors unique to the individual (e.g., socio-demographics), the trip itself (e.g., the specific modes of transportation being used), and the underlying physical environment

at the moment (e.g., road networks regulate average velocity and bearing). Fundamentally, neither of the two kinds of the existing studies touches upon uncovering the underlying data generation process at the individual level, which is a critical task for passively-generated mobile data to be used for synthetic population in large-scale simulations. This constitutes the central aim of our study.

The great potential of the passively-generated mobile data as stated above can be hampered by the fact that the consumer data privacy protection landscape has become increasingly user-centric, bestowing individuals with much greater flexibility in how their data is handled (i.e., easily allowing users to opt out of location-based data sharing agreements). This is due to a combination of factors both from the supply and demand sides: From the latter, consumers are now much more aware of the implications of unlimited data collection (i.e., hyper-targeted advertisements) and, as a result, wary of opting into agreements from which they do not receive any benefits (Kim et al., 2019). Simultaneously, companies like Apple have spearheaded efforts to promote user-centric privacy, implementing a range of new features in 2021's iOS 15 to help users better control and manage access to their data (Apple, 2021). Though the jury is still out on whether the goal of Apple's new initiatives were to maximize stakeholder returns or avoid anti-trust legislation (or both), the end result is the same: there has been a significant decrease in the amount of data generated, or a significant increase in sparsity. This emphasizes, from another perspective, the need to develop models that can uncover the underlying data generation process of passively-generated mobile data, as such models can generate synthetic data where consumer privacy will be less of an issue.

In this paper, we propose a framework that can be used to potentially generate synthetic individual mobile data that replicates real travel behavior. This approach can achieve two ends: It can augment existing datasets with sparsity, and alternatively, it can be substituted for real raw data while conducting analyses without compromising user privacy. In modeling travel behavior or mobility patterns, individual heterogeneity has been well-documented in numerous studies, as affected by many systematic factors such as demographic, socioeconomic, temporal, cultural, and built environment-related factors (Bayarma et al., 2007; Kitamura and Van Der Hoorn, 1987; McGuckin and Murakami, 1999a; Nishii et al., 1988; Wallace et al., 2000), as well as spur of the moment factors (Lee and McNally, 2006, 2003). Thus, the challenge is to develop a flexible model framework that can adapt to individual uniqueness and easily scale up as more and new data come in. This model cannot take a parametric form that requires the pre-specification of a functional form. To this end, we propose a multi-task,

multiple kernel Gaussian process framework that discovers optimal kernel specification for each user. Gaussian Processes (GPs) are a generalization of multivariate Gaussian distributions to infinitely many variables. A GP usually consists of two main components—a mean function that specifies the mean at any point in the output space, and a covariance (kernel) function that embeds a measure of similarity between any pair of observations in a multi-dimensional space. The most crucial component of a GP model is its kernel, as it determines the class of functions the GP relies upon (Rasmussen and Williams, 2006). To learn the form of the covariance function, we combine candidate kernels in a domain-inspired and data-driven manner to optimize model performance and use the Kronecker product to reduce the time it takes to invert a complex covariance matrix.

While powerful, unconstrained GPs can generate predictions that do not align with real-world conditions. Discrete observations of mobile data tend to satisfy certain physical bounds, which stem from the underlying street network and the broader built and natural environment. Specifically, we deal with two dimensions of physics: (1) Land-based human mobility has inherent limitations on the speed of travel, mostly related to the underlying street network as well as the time of day[1]. For example, a driver cannot travel 100 kilometers per hour in a residential street. Similarly, trips at night may have higher average velocities due to reduced traffic. (2) The direction of a user between consecutive data points is limited by the mode of travel and the built environment. If there is no paved road going east at a juncture, a person traveling by car would not be able to traverse that direction, just as a person cannot walk into a lake. Thus, our approach integrates physical knowledge, such as segment velocities and bearings, into the GP inference process. Our conditional-generative framework, inspired by (but also differs from) the likes of Raissi et al. (2019), Lasserre et al. (2006), Wang et al. (2022b), works by first sampling a set of locations as a function of time (conditional model) and then using a generative process to simulate physical variable observations at a certain time and location. However, our framework differs from existing methods in that it does not generate new training data before posterior estimation. Instead, it samples training data during estimation and generates synthetic physical inputs for model inference.

Our Contributions

- We propose a conditional-generative GP framework to generate synthetic individual

[1]This applies to water-based and air-based human mobility as well. However, because these trips are not captured in the dataset we use, we do not consider them in our methodology.

mobile data that provably replicates observed travel patterns (Theorem 1);

- We formulate a flexible physics-regularization framework that permits the learning of composite kernels that can accommodate nonsmooth and nonstationary patterns in human mobility modeling;

- We develop a data-driven multiple kernel learning (MKL) approach to determine the best spatiotemporal kernel for each individual;

- Our experiments lead us to three findings: First, the impact of kernel choice on mobility metrics derived from synthetic data is non-negligible. Second, incorporating physics-based regularization not only diminishes model bias but also enhances the precision of uncertainty estimates in predicting locations. Third, the proposed method is robust and generalizes well to varying individuals and modes of travel.

3.2 Related Work

Machine learning techniques like artificial neural networks (ANN) and kernel-based approaches have been particularly promising in modeling human mobility, with notable recent developments in map-matching (Jiang et al., 2023) and trajectory completion (Ugurel et al., 2024a; Wang et al., 2020). The basic idea with many ANN-based models is to replace unknown functions with a neural network and then minimize a loss function with respect to the parameters of the network. Although ANN-based models have demonstrated empirical success in tackling complex problems, their performance can be inadequate for simpler problems if the ANN architecture is not tailored to the underlying domain (Wang et al., 2021b, 2022a). This goes hand-in-hand with the fact that the theory of ANNs has lagged behind its empirical success, with many ANN-based PDE solvers lacking rigorous theoretical guarantees, and convergence analyses often being limited to linear instances with constraints (Shin et al., 2020; Wang et al., 2022a). As Chen et al. (2021) points out, when compared to kernel methods, ANN methods suffer from limited theoretical foundations and present less favorable trade-offs between computational complexity and accuracy estimates. These limitations are especially salient in generative settings, where complex ANN-based architectures are now widely used to synthesize human mobility traces but often remain black boxes with limited interpretability or uncertainty quantification. The generative nature of kernel methods can

provide a more elucidated mathematical form of its model and can provide uncertainty quantification, probabilistic inference, and simulation.

Beyond tasks like map-matching and trajectory completion, a growing line of work uses data-driven models to generate synthetic human mobility data as a privacy-preserving proxy for raw trajectories. Early approaches build structured stochastic models that decouple temporal routines from spatial movement, for example diary-based frameworks that learn Markovian patterns of routine versus non-routine activity and then assign locations via preferential exploration and return mechanisms (Pappalardo and Simini, 2018). More recent studies rely on deep generative models. Recurrent neural networks have been used to simulate individual-level traces by learning next-location sequences (Kulkarni and Garbinato, 2017), and later extended to generate synthetic populations over multiple days while explicitly tuning randomness to balance statistical fidelity with person-level dissimilarity for privacy (Berke et al., 2022). Generative adversarial networks with mobility-aware regularization further integrate spatial relations and daily periodicity to reproduce a wide range of empirical mobility statistics and support downstream simulations such as epidemic spread (Feng et al., 2020). Diffusion-based models push this idea further by training powerful denoisers that produce high-fidelity synthetic GPS datasets for urban analysis (Zhu et al., 2023). Most recently, large language model style transformers have been adapted to mobility generation, treating sequences of activities, locations, and modes as tokens and conditioning on socio-demographic attributes to generate rich traveler profiles (Zhang et al., 2024b). A recent systematic review highlights that, despite their empirical realism, these ANN-based generators generally prioritize utility, offer limited formal privacy guarantees, and provide little in the way of calibrated uncertainty or interpretable structure (Kapp et al., 2023).

In contrast to these predominantly ANN-based generative models, Gaussian Process priors offer a Bayesian alternative for modeling and simulating complex spatio-temporal processes. Rasmussen & Williams (2006) wrote extensively on modeling and predicting complex interactions in space and time while accounting for spatial non-stationarity. The flexibility of Gaussian Processes in capturing nonlinearity and incorporating prior knowledge through kernel selection has demonstrated their efficacy in uncovering hidden patterns and predicting in dynamic geospatial contexts. The major limitation of using GPs is that the expert must choose a kernel to describe the underlying functional relationships in the data. An ill-fitting kernel can degrade GP model performance, leading to poor generalization and inadequate feature representation (Gönen and Alpaydın, 2011). In the human mobility

domain, only a few previous works have explored preliminary methods on this kernel selection problem, while many are limited to choosing a kernel from an established off-the-shelf list of common kernels. Liu and Onnela (2021) proposed an additive kernel with a daily and a weekly temporal component as well as a distance-based spatial component to impute missing values in longitudinal mobile data, while Batista et al. (2022) focused primarily on the RBF and Matern kernels. The literature lacks a systematic framework established to account for heterogeneity in travel behavior at the individual level.

As a data-driven method, the performance of GP models can suffer when the training data does not sufficiently reflect the complexity of the system (Wang et al., 2022b). Left unconstrained, a GP's behavior may not satisfy expected system outcomes (Ugurel, 2023). Furthermore, GPs may produce predictions that are inconsistent with real-world conditions, defying the laws of physics or the constraints imposed by the built and natural environments. Physical models, built using physics knowledge expressed through differential equations, can characterize the underlying principles governing a system without needing large amounts of training data (Lapidus and Pinder, 1999). Thus, marrying data-driven methods with physical models not only provides insights into the system's mechanics but also enhances prediction accuracy by overcoming data limitations, which addresses the issues arising from unconstrained GPs.

For synthetic mobility generation, the lack of explicit physical constraints in existing ANN-based models means that generated traces can violate simple laws of motion or built-environment constraints, even when they match aggregate statistics. Physics-regularization has been extremely popular in the context of ANNs (Cuomo et al., 2022; Raissi et al., 2019). Though several studies have proposed analogous approaches using GPs and other kernel methods (Chen et al., 2021; Nevin et al., 2021; Wang et al., 2022b; Yang et al., 2019), the methodology has not been developed as extensively. Specifically, the studies focusing on physics-regularized GPs have proposed methods for deriving physical knowledge for the estimation phase (i.e., to obtain the posterior). In this study, we use the human mobility domain to our advantage in overcoming this estimation problem, specifically by reducing the noise associated with direct estimation through proper data labeling (see Section 3.3.3). Conversely, the challenge of inference using this posterior, particularly when it involves physical variables that are dependent on both the inputs and the outputs, has not been extensively explored.

Within the transportation domain, Yuan et al. (2021) proposed a physics-regularized GP

analogous to that of Wang et al. (2022b) to model macroscopic traffic flow, while Zhu et al. (2022) proposed a physics-regularized multi-output GP to model the numbers of pickups, returns, and idle bikes of a large-scale bike-sharing system. Recently, Wu et al. (2024a) used anisotropic Gaussian Processes to model congestion propagation in traffic flow data. These studies highlight the growing interest in applying physics-regularization in diverse contexts, yet there remains a gap in its application to individual-level mobility data where each individual has their unique spatial coverage and the data can extend for days, weeks or months. To our knowledge, this study is one of the first to propose a method for generating GPS-based human mobility traces that (i) models individual-level heterogeneity through tailored kernels and (ii) explicitly regularizes the latent functions with physical attributes associated with the built environment, thereby complementing existing ANN-based synthetic mobility generators.

3.3 Methodology

3.3.1 Problem Definition

Mathematically, the problem of uncovering the underlying data generation process for synthetic individual mobile data generation translates into two related research questions:

1. Given time, how do we infer (predict) spatial locations?

2. How do we infuse physics (i.e., velocity and bearing) into the inference problem from time to location, as stated above?

The input time can be represented by a set of variables, in addition to being simply a time index (e.g., the Unix time). Those additional time-related variables include periodic elements like the day of the week and weeks of the month. They capture varying rhythms that have long been identified in travel behavior literature (Bayarma et al., 2007; Gonzalez et al., 2008; Hanson and Huff, 1988): e.g., people usually go to their workplaces and return home every day and other trips like grocery shopping, going out with friends, or attending family gatherings happen less frequently. Additionally, there are behavioral patterns correlated with certain weeks of a month or months of a year (i.e., Christmas, Thanksgiving, etc.). We denote with $\mathbf{T}$ the matrix of temporal variables, where $t_{u,i}$ is the u^{th} temporal dimension at the i^{th} observation. The physics-related variables such as average velocity and bearings are to capture the constraints that the built environment has on one's movement pattern. We

denote with $\mathbf{P}$ the matrix of physics-related variables, where v_i and β_i is the average velocity and bearing at i^{th} observation respectively. The output variable location $\mathbf{Y}$ is represented with latitude λ and longitudeϕ [2]:

$$\mathbf{T} = \begin{bmatrix} t_{1,1} & \dots & t_{d,1} \\ \vdots & \ddots & \vdots \\ t_{1,n} & \dots & t_{d,n} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_n \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} v_1 & \beta_1 \\ \vdots & \vdots \\ v_n & \beta_n \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_{\lambda,1} & y_{\phi,1} \\ \vdots & \vdots \\ y_{\lambda,n} & y_{\phi,n} \end{bmatrix}. \tag{3.1}$$

Human movements, as reflected in individual-level mobile phone trajectory data, are highly non-linear, context-dependent, and heterogeneous. Underlying the observed mobile data is the wide range of transportation modes used, each with different underlying physics (i.e., average velocity, acceleration behavior). When more than one mode of transportation is used, there is also a change of modes that need to be captured. Individual movement behaviors are also heterogeneous. While previous works have explored "universal" models on human movement patterns (Gonzalez et al., 2008; Song et al., 2010b), later studies have revealed many other models that can capture human movement patterns and no single model will outperform others across contexts. These individual-level complexities suggest that making rigid assumptions about the shape or form of the underlying distribution of an individual's mobile data is unlikely a fruitful approach. A non-parametric, data-driven method holds greater promise, being more flexible and less constrained by predefined assumptions. Such an approach is capable of accommodating complex patterns and more adeptly capturing the inherent relationships present within the data. Thus, to answer the first research question stated above, we propose a multi-task Gaussian Process that learns individualized kernels to infer a set of locations $\mathbf{Y}$ from a set of temporal variables $\mathbf{T}$. As noted earlier, identifying kernels, or more specifically the correct structure for a combination of kernels, is critical as a fitting kernel can seamlessly capture complex dependencies in mobile data while a non-fitting one will be inadequate. In this paper, we address this multiple kernel learning problem in the context of mobile trajectory data (see Section 3.3.4).

To answer the second research problem stated above, we need information on the physics-related variables, or $\mathbf{P}$. And yet, $\mathbf{P}$ is not directly observed. More specifically, this issue of non-observability is different in the estimation problem where the kernels are learned from

[2] For simplicity, we project λ and ϕ to a two-dimensional grid such that a user's location at time t is $\mathbf{y}_t = [y_\lambda, y_\phi]$. As travel distances between consecutive data points in mobile datasets tend to be short, this does not make a significant difference compared with extensions to spherical geometry.

$\mathbf{T}$ to $\mathbf{Y}$ versus the inference problem where $\mathbf{Y}$ is inferred based on $\mathbf{T}$. In the estimation problem, since $\mathbf{T}$ and $\mathbf{Y}$ are both observed in the training data, $\mathbf{P}$ can be calculated. This calculation, of course, is unavoidably tampered with noise, especially when the time gap between two consecutive observations is large. We address this issue by a sample labeling procedure that reduces bias in the estimation of physical variables (see Section 3.3.3). In the inference problem, $\mathbf{Y}$ is not available, since it is to be inferred. Thus, we must first estimate $\mathbf{P}$ by considering its dependency on the yet-to-be-predicted output variable $\mathbf{Y}$, as well as on $\mathbf{T}$. We achieve this by formulating a model to predict the unobserved physical constraints $\mathbf{P}$ as a function of $\mathbf{T}$ and $\mathbf{Y}$. We then incorporate these generated constraints as observed inputs to aid in the inference of locations. We detail this strategy in Section 3.3.5.

The rest of Section 3.3 is organized as such: Section 3.3.2 translates the defined problem into one that can be addressed with Gaussian processes; Section 3.3.3 describes how to incorporate $\mathbf{P} \rightarrow \mathbf{Y}$ in the modeling framework; Section 3.3.4 discusses our greedy kernel learning approach to learn individual kernel structures; Section 3.3.5 ties together Sections 3.3.1 through 3.3.4 and describes our model inference strategy.

3.3.2 Multi-task Gaussian Process for Mobile Data

We first focus on modeling the relationship $\mathbf{T} \rightarrow \mathbf{Y}$. Consider the task of learning a function $f_j : \mathbb{R}^d \rightarrow \mathbb{R}$ from a training set $\mathcal{D} = (\mathbf{T}, \mathbf{Y})$ where j refers to either latitudes ϕ or longitudes λ. The basic form of our learning problem is

$$y_{ji} = f_j(\mathbf{t}_i) + \epsilon_{ji}, \tag{3.2}$$

where f_j is a systematic function mapping inputs $\mathbf{t}_i$ to output y_{ji}, and $\epsilon_{ji} \sim \mathcal{N}(0, \delta_j^2)$ are independent random variables for noise associated with the j^{th} task. We place a GP prior on f_j such that $f_j \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$, where $m(\cdot) = \mathbb{E}[f_j(\cdot)]$ is the mean function (hereafter set to $\mathbf{0}$ assuming proper normalization), and $k(\cdot, \cdot)$ is the covariance (or kernel) function.

Assumption 1. *The set* $\mathbf{f}_j = [f_j(\mathbf{t}_1), \ldots, f_j(\mathbf{t}_n)]^\top$ *follows a multivariate normal distribution such that* $p(\mathbf{f}_j|\mathbf{T}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_t)$*, where* $\mathbf{K}_t$ *is the covariance matrix such that* $[\mathbf{K}_t]_{i,g} = k_t(\mathbf{t}_i, \mathbf{t}_g)$*, where* $subscript_t$ *is used to indicate temporal correlations and* $subscript_g$ *is an index like* $subscript_i$*.*

In our approach, we relate multiple tasks (each multivariate normal distributed) by leveraging the correlations between them. Thus, we specify the covariance matrix for all n

observations and two tasks as

$$\mathbf{K}_t = \mathbf{K}^t(\mathbf{T}, \mathbf{T}) \otimes \mathbf{K}^f(\mathbf{y}_\lambda, \mathbf{y}_\phi), \tag{3.3}$$

where $\mathbf{K}^t$ is the $n \times n$ covariance matrix of the training times using any valid PSD kernel, $\otimes$ is the Kronecker product, and $\mathbf{K}^f$ is a 2×2 PSD matrix containing the inter-task covariances, learned by jointly minimizing the negative marginal log-likelihood (Bonilla et al., 2007). The dimension of $\mathbf{K}_t$ for two tasks is then $2n \times 2n$. A key property of this model is that the joint distribution over the entire output space is not block-diagonal with respect to tasks. That is, observations of one task can affect the predictions on another task.

Assumption 2. *The inferred location $f_j(\mathbf{t}_*)$ of a new input $\mathbf{t}_*$ conditioned on the training data is assumed to be distributed with the following form:*

$$p(f_j(\mathbf{t}_*)|\mathbf{t}_*, \mathbf{T}, \mathbf{Y}, \delta_j^2) \sim N(\boldsymbol{\mu}_*, \boldsymbol{\sigma}_*^2), \tag{3.4}$$

where

$$\begin{aligned} \boldsymbol{\mu}_* &= (\mathbf{k}_j^f \otimes \mathbf{k}_*)(\mathbf{K}^f \otimes \mathbf{K}^t + \mathbf{D} \otimes \mathbf{I})^{-1} vec(\mathbf{Y}) \\ \boldsymbol{\sigma}_*^2 &= (\mathbf{k}_j^f \otimes \mathbf{k}_{**}) - (\mathbf{k}_j^f \otimes \mathbf{k}_*)(\mathbf{K}^f \otimes \mathbf{K}^t + \mathbf{D} \otimes \mathbf{I})^{-1}(\mathbf{k}_j^f \otimes \mathbf{k}_*). \end{aligned} \tag{3.5}$$

Here, $\mathbf{k}_j^f$ selects the j^{th} column of $\mathbf{K}^f$, $\mathbf{k}_* = k(\mathbf{t}_*, \mathbf{T})$ is the vector of covariances between the test point and the training set, $\mathbf{D}$ is a 2×2 diagonal matrix with the variances of the noise processes for latitude and longitude δ_j^2, and $\mathbf{k}_{**} = k(\mathbf{t}_*, \mathbf{t}_*)$. Finally, we minimize the negative marginal log-likelihood (MLL) of the output vectors with respect to the training data in determining the optimal hyperparameters.

$$-\log(p(\mathbf{Y}|\mathbf{T}, \Theta)) = \frac{1}{2}[\text{vec}(\mathbf{Y})^\top \boldsymbol{\Sigma}^{-1} \text{vec}(\mathbf{Y}) + \log(\det(\mathbf{K}_t)) + |\Theta| \log(2\pi)], \tag{3.6}$$

where Θ is the set of model parameters, $|\Theta|$ denotes the cardinality of the set of model parameters, which include both kernel-specific parameters as well as the weights associated with each kernel component (discussed further in Section 3.3.4), $\boldsymbol{\Sigma} = \mathbf{K}^f \otimes \mathbf{K}^t + \mathbf{D} \otimes \mathbf{I}$, and $\det(\mathbf{K}_t)$ is the determinant of the $\mathbf{K}_t$ matrix. We emphasize that the elements of $\mathbf{K}^f$ are treated as free parameters which are learned through Equation 3.6. The negative marginal log-likelihood provides a target function for kernel learning. The first component, $\text{vec}(\mathbf{Y})^T \Sigma^{-1} \text{vec}(\mathbf{Y})$, estimates the model fit, while the second and the third terms, $\log(\det(\mathbf{K}_t)) + |\Theta| \log(2\pi)$, act as the regularization (Rasmussen and Ghahramani, 2000).

3.3.3 Physics-regularized GP

Only using the relationship $\mathbf{T} \rightarrow \mathbf{Y}$ to generate synthetic data can be problematic, as unconstrained GPs may produce predictions that are inconsistent with real-world conditions. We can indeed regularize the framework we have thus described through measurements of velocity and bearing, which are realizations of physics-based constraints acting on every user. Incorporating physical knowledge helps reduce the risk of overfitting and improves the likelihood of model generalization since the physical knowledge provides data augmentation. We derive observations of these variables with

$$v_\lambda = \frac{dy_\lambda}{dt}, \quad v_\phi = \frac{dy_\phi}{dt}, \quad v = \sqrt{v_\lambda^2 + v_\phi^2} \tag{3.7}$$

$$\beta = \arctan\left(\frac{v_\phi}{v_\lambda}\right) = \arctan\left(\frac{dy_\phi}{dy_\lambda}\right) \tag{3.8}$$

where v_λ and v_ϕ denote the velocity in the horizontal and vertical direction, respectively, v denotes the overall velocity, and β denotes the bearing.

One issue with using physical variables is the noise that may be introduced by irregular sampling of training data. That is, if time gaps between data points are large, the estimated segment velocities and bearings may not be representative of the distribution of physics within that period. To get around this, Wang et al. (2022b) proposed to sample the physical observations from the posterior of the GP modeling $\mathbf{T} \rightarrow \mathbf{Y}$, thereby retaining control over where $\mathbf{P}$ is observed by utilizing a continuous function. A drawback of this approach is that the resulting likelihood function is difficult to solve, necessitating the calculation of the evidence lower bound (ELBO). Additionally, it significantly increases the computational demand, raising the complexity of the conditional GP from $O(n^3)$ to $O(n^3 + m^3)$, where m represents the number of sampled locations for $\mathbf{P}$.

Alternatively, Alvarez and colleagues (2013; 2009) propose a different strategy that sidesteps the problem by embedding physical knowledge directly into the kernel's structure, rather than incorporating it as part of the training data. While this method simplifies the model, it also imposes limitations on the selection of kernels. Specifically, it restricts the choice to simpler and smoother kernel functions, which in turn constrains the ability to incorporate complex physical knowledge into the model through more sophisticated and adaptable kernels.

To mitigate the introduction of noise from data points with extensive time gaps, we simply

assign each data point the same label until the time gap between the next observation exceeds a threshold (in our case, we used 1 hour). This essentially labels continuous trajectories with the same unique identifier. For the first point of each unique trajectory, we assign a velocity and bearing of zero as we cannot observe the physical variables from the most recent trajectory (which would be > 1 hour ago). This avoids having to drop training data with noisy physical variables without complicating the estimation procedure.

After ensuring the training data is efficiently sampled, adding physical variables to the posterior of the temporal GP translates to sampling from a larger covariance matrix that also considers physical correlations between the physical variables and their interactions with the temporal variables.

Assumption 3. *Let* $\mathbf{X} = \begin{bmatrix} \mathbf{T} & \mathbf{P} \end{bmatrix}$ *such that a data vector* $\mathbf{x}_i$ *has both temporal and physical variables. Given the set of physics observations* $\mathbf{P}$*, the set* $\mathbf{f}_j = [f_j(\mathbf{x}_1), \ldots, f_j(\mathbf{x}_n)]$ *follows a normal distribution such that* $p(\mathbf{f}_j|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{comp})$*, where* $\mathbf{K}_{comp}$ *is the covariance matrix with entries* $[\mathbf{K}_{comp}]_{i,g} = k(\mathbf{x}_i, \mathbf{x}_g)$*.*

The inclusion of the physical variables is compatible with the multi-task kernel structure, as the kernel in Equation 3.3 becomes $\mathbf{K}_{comp} = \mathbf{K}^x(\mathbf{X}, \mathbf{X}) \otimes \mathbf{K}^f(\mathbf{y}_\lambda, \mathbf{y}_\phi)$. This then changes the probability of observing a location $f_j(\mathbf{x}_*)$ at unobserved input $\mathbf{x}_*$ to be conditional on not only time but physics as well.

Assumption 4. *The inferred location* $f_j(\mathbf{x}_*)$ *of a new input* $\mathbf{x}_*$ *conditioned on the training data is assumed to be distributed with the following form:*

$$p(f_j(\mathbf{x}_*)|\mathbf{x}_*, \mathbf{T}, \mathbf{P}, \mathbf{Y}, \delta_j^2) \sim \mathcal{N}(\boldsymbol{\mu}_{**}, \boldsymbol{\sigma}_{**}^2), \tag{3.9}$$

where $\boldsymbol{\mu}_{**}$ *and* $\boldsymbol{\sigma}_{**}^2$ *are defined as before but using the larger covariance matrix* $\mathbf{K}_{comp}$*.*

This process transforms the MLL shown in Equation 3.6 to

$$-\log(p(\mathbf{Y}|\mathbf{T}, \mathbf{P}, \Theta)) = \frac{1}{2}[\text{vec}(\mathbf{Y})^\top \boldsymbol{\Sigma}^{-1} \text{vec}(\mathbf{Y}) + \log(\det(\mathbf{K}_{comp})) + |\Theta| \log(2\pi)]. \tag{3.10}$$

The impact of including physical variables is non-negligible, both for estimation and inference (the latter of which we discuss in Section 3.3.5). Only looking at the temporal learning problem $\mathbf{T} \rightarrow \mathbf{Y}$ risks letting the GP remain too unconstrained, leading to illogical travel behavior predictions or unfeasible routes. For example, a well-fit GP with no physics-regularization tends to predict a user hovering around an activity location (rather than staying

stationary while not on a trip), potentially leading to a false "walking trip" interpretation. On the other hand, a well-fit GP with physics-regularization largely solves this issue, as the model is statistically informed on whether the user is moving or remaining stationary.

3.3.4 Greedy Algorithm for Multiple Kernel Learning

Multiple kernel learning comprises two steps: the first step is to select a set of base kernels that are suitable to the context of the study (in this case, it is human mobility patterns) and the second step is to identify the optimal combination of the base kernels.

3.3.4.1 Selecting base kernels

The selection of base kernels is based on two key features of human mobility patterns. The first one relates to that observations of human mobility at nearby time intervals tend to be geographically close to each other. This concept extends Tobler's first law of geography (Tobler, 1970), which states that things in close proximity are more closely related than those far apart. Such dependencies can be captured rather straightforwardly using smooth functions derived from a variety of kernels—in this work, we consider the squared exponential (SE), rational quadratic (RQ), and Matern (MAT) 5/2 kernels:

$$k_{SE} = \eta \exp\left(-\frac{|x_i - x_g|^2}{2l^2}\right), \tag{3.11}$$

$$k_{RQ} = \eta \left(1 + \frac{|x_i - x_g|^2}{2\alpha l^2}\right)^{-\alpha}, \tag{3.12}$$

$$k_{MAT_{5/2}} = \eta \left(1 + \frac{\sqrt{5}|x_i - x_g|}{l} + \frac{5|x_i - x_g|^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}|x_i - x_g|}{l}\right), \tag{3.13}$$

where $|x_i - x_g|$ represents the Euclidian distance between any pair of inputs x_i and x_g; η is a scale parameter[3]; the lengthscale l determines the smoothness of the function; and the scale mixture α determines the relative weight of large- and small-scale variations in the data.

The SE kernel is characterized by infinite differentiability and is defined by a single lengthscale parameter. It is suitable for capturing smooth and gradual variations in the data. The RQ kernel introduces an additional scale mixture parameter that controls the smoothness

[3]The scaling parameter can also be thought of as a weight in the composite kernel, determining the influence of each base kernel.

of the kernel function and therefore can model both short- and long-range dependencies in the data. It can capture various patterns, from smooth variations to abrupt changes, making it suitable for datasets with diverse characteristics. Finally, the Matern 5/2 kernel exhibits differentiability up to the second order and is often favored when there is prior knowledge that the underlying process has a roughness level between that of the SE and the Matern 3/2 kernels. The Matern 5/2 kernel can capture moderate levels of non-smoothness in the data while still maintaining a certain degree of smoothness.

The second feature as noted in Section 3.3.1 is that individuals' travel behavior tends to center around a few important locations, such as home and work and there tend to be rhythms or cyclical patterns of different lengths with respect to how individuals conduct activities. To account for these, we further introduce categorical variables to capture cyclical patterns of different lengths. We represent calendar-based structures like days, weeks, and months as binary variables using a one-of-k encoding. For example, as the days of the week can take one of seven values $\{Mo, Tu, We, Th, Fr, Sa, Su\}$, a one-of-k encoding of We will correspond to $\{0, 0, 1, 0, 0, 0, 0\}$. We embed multiple categorical inputs in a GP framework by multiplying the same kernel across the one-hot encodings:

$$k_{RQ_{cat}} = \prod_{u=1}^{d} k_{RQ_u}. \tag{3.14}$$

Equation 3.14 is a product kernel containing a unique lengthscale parameter for each input dimension being multiplied across the function. A small optimal lengthscale for a categorical variable implies a relatively low correlation among data in that category, while a large optimal lengthscale implies a high correlation. In addition to the one-hot encoding strategy, we also use the periodic kernel (Equation 3.15), which allows GPs to model functions that repeat themselves:

$$k_{PER} = \exp\left(-\frac{2\sin^2(\pi|x_i - x_g|/p)}{l^2}\right), \tag{3.15}$$

where the period length p determines the distance between repetitions of the function. Due to its sinusoidal nature, the periodic kernel is a better fit for continuous input dimensions that are unbounded in the input space.

3.3.4.2 Learning multiple kernel structures

The basic rationale of MKL builds on the possibility to create composite kernels comprising base kernels. So, the task at hand is kernel function discovery, i.e., finding the optimal structure for the composite kernel function that comprises a set of base ones. Kernel function discoveries have been done in other domains that aim to find concrete analytic forms of the kernels (Barla et al., 2003; Gibbs, 1998; Ong et al., 2002; Wilson, 2014), but there have been no specialized kernel functions developed for individual mobile data. In this paper, we develop a multiple kernel learning algorithm that is data-driven.

A GP's kernel is only valid if it has a PSD covariance matrix (Rasmussen and Williams, 2006). The direct sum or the tensor product of two valid kernels is also a valid kernel (for proof, we refer the reader to Rasmussen & Williams (2006) section 4.2.4). To put it another way, the set of PSD kernels is closed under sum and product operations. Multiplying two kernels can be interpreted as an **AND** operation while adding two kernels can be interpreted as an **OR** operation (Duvenaud, 2014). Let k_1 and k_2 be kernels which each depend on a single input vector, $\mathbf{x}$ and $\mathbf{y}$, respectively. The product of k_1 and k_2 will result in a prior over functions that vary across both $\mathbf{x}$ and $\mathbf{y}$ and hence the function value $f(x_i, y_i)$ is only expected to be similar to some other function value $f(x_g, y_g)$ if x_i is close to x_g **AND** y_i is close to y_g. On the other hand, the sum $k_1 + k_2$ will result in a prior over functions which are a sum of one-dimensional functions, and hence the function $f(\mathbf{x}, \mathbf{y}) = f_x(\mathbf{x}) + f_y(\mathbf{y})$.

Our MKL learning is essentially learning about six components: target function, the base learner, the learning method, the functional form, the training method, and the computational complexity (Gönen and Alpaydın, 2011). We use the negative marginal log-likelihood (Equation 3.6) as our target function to assess the model's goodness of fit. However, simply comparing MLLs after optimizing kernel parameters would result in a bias in favor of more complex models. We avoid this overfitting issue by approximating the integral of the marginal likelihood over all free parameters using the Bayesian information criterion (BIC) (Schwarz, 1978):

$$\mathrm{BIC}(n, |\Theta|) = |\Theta| \log(n) - \log(p(\mathbf{Y}|\mathbf{X}, \Theta)). \tag{3.16}$$

Our greedy multi-kernel learning is essentially iteratively adding or multiplying new kernels until the BIC is minimized. The functional form of the derived kernel expressions can be both linear (i.e., additive) or non-linear (i.e., product)—each component has a weight parameter that determines its relevance and proportion of importance. We use the Adaptive Moment

Estimation algorithm (Kingma and Ba, 2014) as the training method, which optimizes both the kernel weight parameters $\boldsymbol{\eta}$ and the base kernel parameters simultaneously. The computational complexity is a function of the algebraic operations involved, which determines the number of times the composite covariance matrix has to be inverted. We initialize the algorithm with a set of base kernels B, the maximum number of steps M, and a set of algebraic operations A[4]. This optimization procedure can be formally stated as:

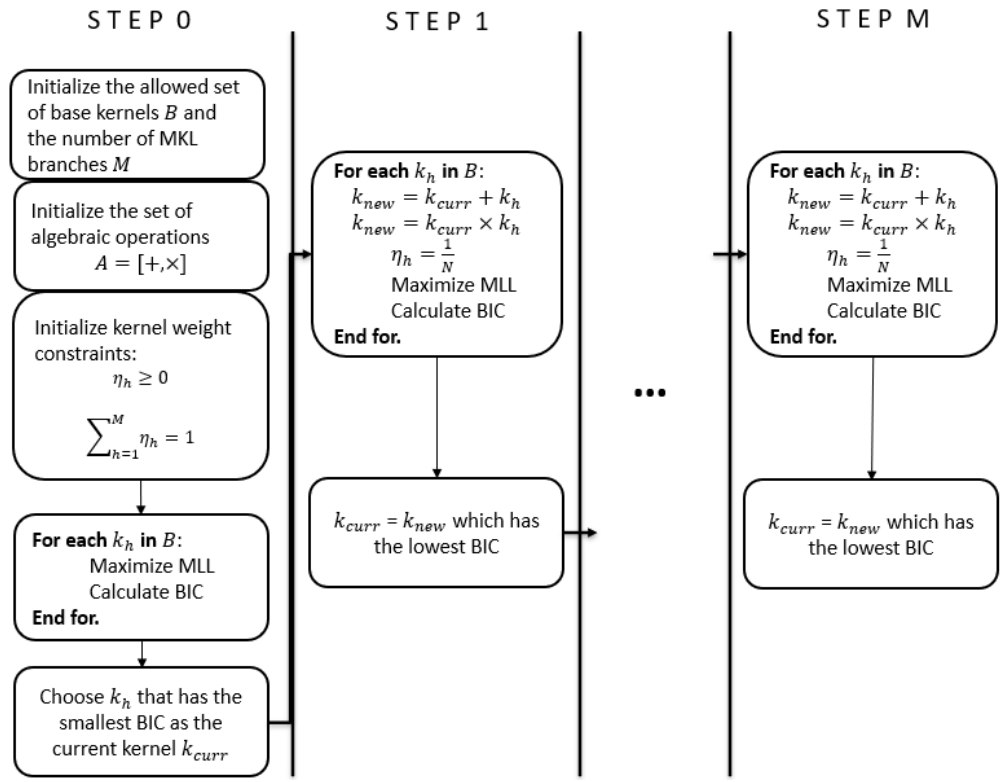
$$\begin{aligned} &\underset{\boldsymbol{\eta},\mathbf{K},\Theta}{\operatorname{argmin}} \quad -\log(p(\mathbf{Y}|\mathbf{X},\Theta,\boldsymbol{\eta},k_{curr}))+|\Theta|\log(n), \\ &\quad s.t. \quad k_{curr}=\sum_{h=1}^{M}\eta_h k_h \quad \text{where} \quad k_h=\begin{cases} k_{prev}+k_{base}, & \text{if addition at step } h \\ k_{prev}\times k_{base}, & \text{if multiplication at step } h \end{cases} \\ &\qquad \sum_{h=1}^{M}\eta_h=1, \quad \eta_h\geq 0 \quad \forall h\in 1,\ldots,M. \end{aligned} \tag{3.17}$$

As illustrated in Figure 3.1, we first estimate the BIC of each kernel in B and choose the lowest one. We then begin the greedy procedure: we apply each operation in the set of algebraic operations A to the chosen kernel and every available kernel in B (that is, they are added or multiplied) and choose the combination with the lowest BIC. This procedure is repeated until no new kernels result in a lower BIC than that of the existing kernel or until we reach M. The convergence properties of this algorithm depend largely on the complexity of the training data—if the data includes a large timeframe and the user displays multiple periodicities, trends, or random variations, the algorithm may not converge until we reach M. Otherwise, only one or two steps can be sufficient in describing the user's behavior.

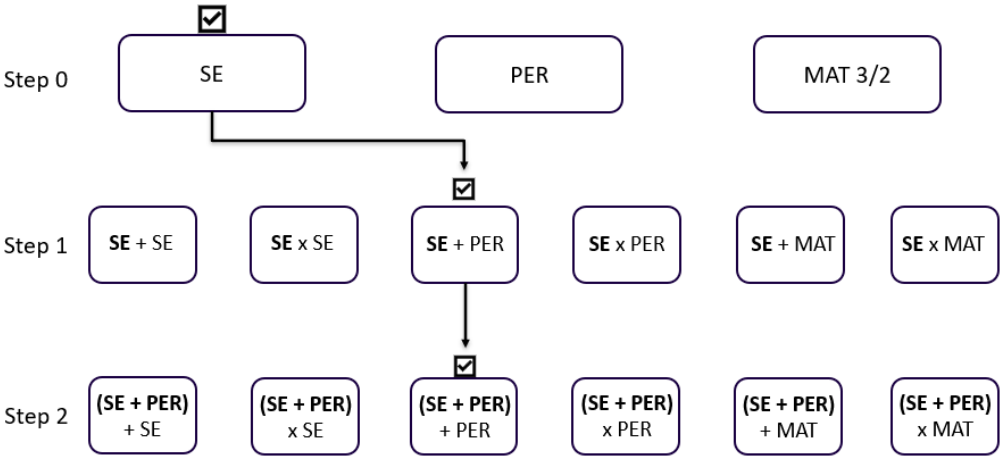
3.3.5 Model Inference

So far, we have dealt with the estimation of the posterior for the GP model. However, inference using physical variables presents a significant obstacle as it is a random variable determined by both temporal and spatial factors (locations), the latter of which are unobserved. Thus, prior to generating a set of synthetic locations $\mathbf{Y}_{gen}$ using the conditional model, we must specify a model to predict the unobserved physical constraints $\mathbf{P}_{gen}$ as a function of time and space. For the rest of the paper, $\mathbf{T}$, $\mathbf{P}$, and $\mathbf{Y}$ will be represented with $\mathbf{T}_{cond}$, $\mathbf{P}_{cond}$, and

[4]For now, we only allow addition and multiplication as other operations do not always result in PSD kernels.



(a)



(b)

Figure 3.1: (top) Proposed algorithm for greedy multiple kernel learning; (bottom) example greedy learning tree with three steps (checkmark denotes lowest BIC option)

$\mathbf{Y}_{cond}$ to distinguish the fact that they belong to the conditional model, while sets relating to the generative model will have the subscript "*gen*".

Inspired by Lasserre et al. (2006), we propose a conditional-generative inference model. The generative component is for realizations of physical variables beyond the training data. Unlike LFM and the physics-regularized GP proposed by Wang et al. (2022b), the physical variables in our method are determined partially by the output of the conditional model (coordinates). Therefore, we first generate a set of noisy locations $\bar{\mathbf{Y}}_{gen} = [\bar{\mathbf{y}}_{gen,1}, \ldots, \bar{\mathbf{y}}_{gen,m}]^\top$ induced by a set of times $\mathbf{T}_{gen}$ using the conditional multi-task GP $f^{\bar{y}} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_t)$ via Equations 3.4 and 3.5, where $[\mathbf{K}_t]_{i,g} = k_t(\mathbf{t}_i, \mathbf{t}_g)$.

To estimate $\mathbf{P}_{gen}$, we require a second posterior that takes in spatial and temporal observations $\mathbf{Z} = \begin{bmatrix} \mathbf{Y}_{cond} & \mathbf{T}_{cond} \end{bmatrix}$ and approximates a function $f^p : \mathbf{Z} \rightarrow \mathbf{P}_{cond}$. This is achieved by defining another multi-task GP ("hidden GP" in Figure 3.2) for the physical variables $f^p \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_p)$ where $[\mathbf{K}_p]_{i,g} = k_{comp}(\mathbf{z}_i, \mathbf{z}_g)$. Sampling from the posterior distribution

$$\mathbf{P}_{gen} \sim p(\mathbf{T}_{gen}) p(f^p | \bar{\mathbf{Y}}_{gen}, \mathbf{Y}_{cond}, \mathbf{T}_{cond}, \mathbf{P}_{cond})) \tag{3.18}$$

where we use the marginal independence of $\mathbf{T}_{gen}$ to simplify the joint conditional distribution. The set of generated physical variables are then incorporated as inputs to the physics-regularized GP model described in Section 3.3.3. Finally, we sample from the physics-regularized GP $f^y \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{comp})$ where $[\mathbf{K}_{comp}]_{i,g} = k_{comp}(\mathbf{x}_i, \mathbf{x}_g)$.

Figure 3.2 shows the workflow for the proposed model, which works in two passes. On the first pass, the physical knowledge is embedded as a function of time and space using a GP ("Physics Embedding"). On the second pass, synthetic data points are generated given $\mathbf{T}_{gen}$ and the estimated physical variables $\mathbf{P}_{gen}$.

Theorem 1. *The empirical distribution functions used for sampling the generated synthetic data* $\mathbf{Y}_{gen}$ *and* $\mathbf{P}_{gen}$ *are constructed only using the basis vectors generated from the conditional and hidden GPs, respectively.*

Proof. Let $F_{\mathbf{Y}_{gen}}$ and $F_{\mathbf{P}_{gen}}$ denote the empirical distribution functions for $\mathbf{Y}_{gen}$ and $\mathbf{P}_{gen}$, respectively. Due to Mercer's theorem[5]), each sample of $\mathbf{P}_{gen}$ and $\mathbf{Y}_{gen}$ is a linear combination of the basis functions determined by the covariance matrices $\mathbf{K}_p$ and $\mathbf{K}_{comp}$. Because $F_{\mathbf{Y}_{gen}}$ and $F_{\mathbf{P}_{gen}}$ are constructed from these samples, they will also be linear combinations of these basis vectors generated from the conditional and hidden GPs. □

[5]For details, see Section 4.3 of Rasmussen & Williams (2006

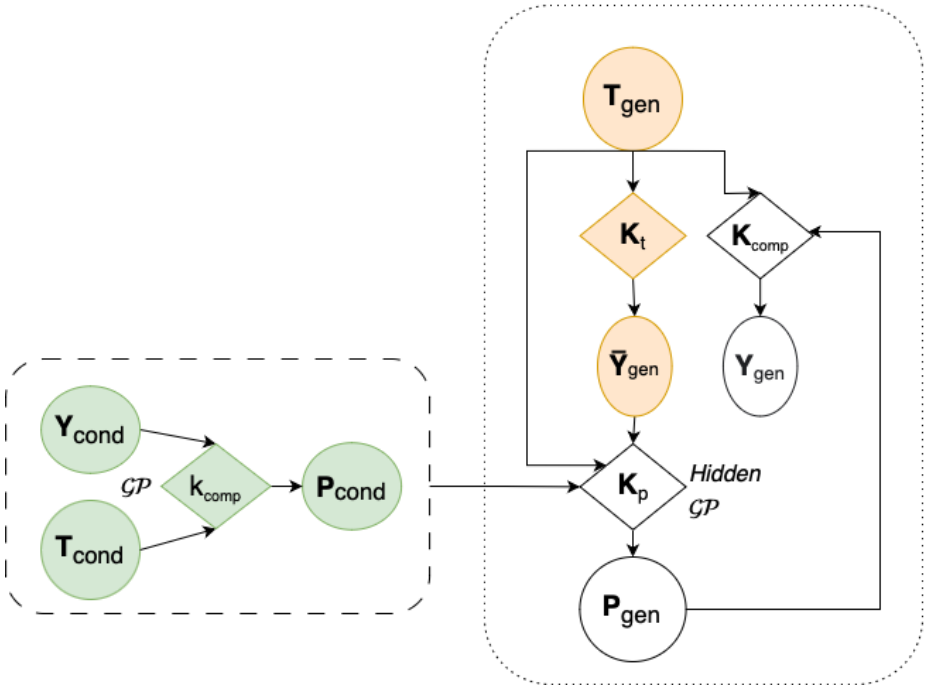


Figure 3.2: Physics-regularized GP inference for passively-generated mobile data. Circles indicate variables, while diamonds indicate estimation (kernel function) or inference (kernel matrix) processes.

The implication of Theorem 1 is three-fold. First, the generated $\mathbf{P}_{gen}$ and $\mathbf{Y}_{gen}$ reflect the patterns, trends, and noise characteristics present in the observed data $\mathbf{P}_{cond}$ and $\mathbf{Y}_{cond}$. Second, the model is not just statistically informed but also physically informed. This means that any generated data point respects the underlying physical laws or constraints that have been encoded through the model framework. These constraints might represent known relationships between variables that must be maintained, such as the laws of motion in a physical system. Third, by adhering to the learned relationships and physical constraints, the synthetic data generation process is not just mimicking the observed data but generalizing from it. This enables the model to create data points that are plausible under the model's understanding of the domain, even if they were not explicitly present in the training data.

3.4 Numerical Experiments

In this section, we document the results of four experiments that evaluate the formulation we have thus presented. In Sections 3.4.1 and 3.4.2, we discover the best-fitting covariance structure for temporal and physical features to create a composite kernel, respectively. In Section 3.4.3, we compare the performance of this composite kernel against alternative formulations over many users in a passively-collected mobile dataset. Finally, to analyze our framework's generalizability, we conduct another set of experiments in Section 3.4.4 on Microsoft Asia's GeoLife dataset using trips of varying modes. All class and function objects relevant to this experiment can be found on GitHub: https://github.com/ekinugurel/physics-regularized-MTGP. Additionally, Appendix B.6 outlines a set of algorithms we use throughout the experiments.

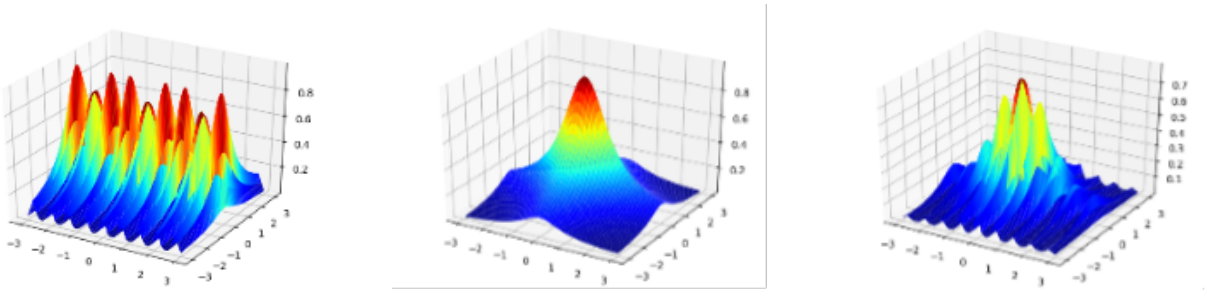
Before all experiments, we perform data standardization (z-score normalization) such that each feature is centered at 0 with a standard deviation of 1. For example, $v_i \in \mathbf{v}$ (the i-th observation of velocity in $\mathbf{v}$) can be standardized with

$$\tilde{v}_i = \frac{v_i - \bar{v}}{s_{\mathbf{v}}} \tag{3.19}$$

where $\bar{v}$ is the sample's average velocity and $s_{\mathbf{v}}$ is the standard deviation of the sample's velocity. As different variables have different units, doing so simplifies the process of optimization via gradient descent–allowing the algorithm to converge more efficiently by ensuring that the scales of all features are comparable. This standardization also aids in preventing certain variables with larger magnitudes from disproportionately influencing the optimization process, promoting a more balanced and effective model training.

3.4.1 Temporal Structure Discovery using Multiple Kernel Learning

We show the process of discovering the temporal structure associated with an individual's mobile data. Specifically, we conduct a test on data over three weeks, using the first two weeks for training the model, and measure its accuracy using the last week. We run the MKL algorithm on the training set with the following base kernels: SE (Equation 3.11), PER (Equation 3.15), and RQ (Equation 3.12). We initiate two instances of the PER kernel: one initialized with $p = 1,440$ (representing a day) and the other initialized with $p = 10,080$ (representing a week). After a sensitivity analysis, we settle on $M = 3$ (as more steps do not result in lower BICs) and initialize the kernel weight constraints. The resulting kernel is a



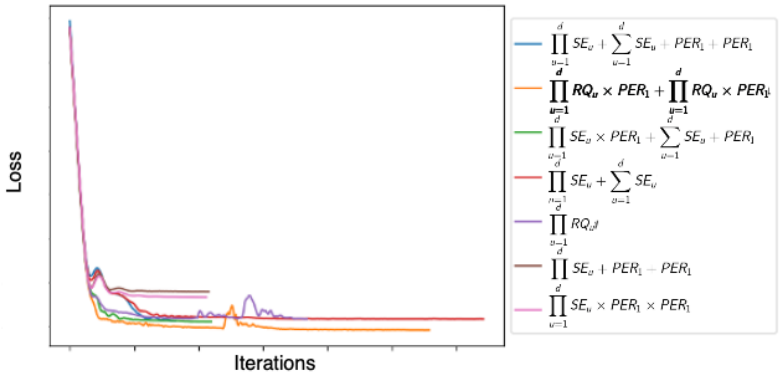
(a) $k_{PER,0} \times k_{RQ,1}$ (b) $k_{RQ,1} \times k_{RQ,2}$ (c) $k_{PER,0} \times k_{RQ,1} \times k_{RQ,2}$

Figure 3.3: Components of the single-task temporal kernel in a 3-D space

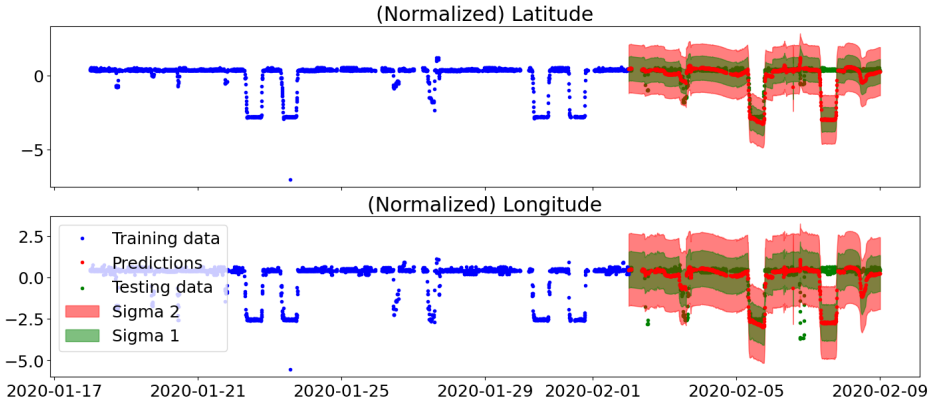
set of rational quadratic kernels multiplied on all input dimensions and one periodic kernel on the Unix time dimension. Figure 3.3 visualizes components of the discovered kernel in a single-task space. To evaluate the success of our algorithm, we compare this kernel with similar kernels that can be constructed using the above base kernel set.

We measure model accuracy with three mobility metrics: spatial burstiness B_s, representing the distribution pattern of travel distances between consecutive stay points (Kim and MacEachren, 2014); radius of gyration r_g, indicating the characteristic travel distance of a user during a period (Song et al., 2010b); and home location inaccuracy ΔH, measuring the haversine distance between user's imputed and actual home location. Our results are detailed in Figure 3.4, while the optimal kernel parameters are shown in Table 3.1. From Figure 3.4a, the convergence behavior of each kernel is discernible, with most of them stabilizing as the number of iterations increases, indicating that the models are reaching an optimal point in terms of the parameters being estimated. A few kernels exhibit some volatility in their loss values, even in later iterations, which could suggest a less stable model fit. Figure 3.4b shows the generated traces using the discovered kernel and the temporal GP based on the training data. Sigma 1 and Sigma 2 represent 1 and 2 standard deviations in the normalized distribution. The coefficients of variation using de-normalized generated traces are 0.127 for the latitudes and 0.026 for the longitudes, respectively, suggesting that GP does a good job in estimating the variance of the predictions.

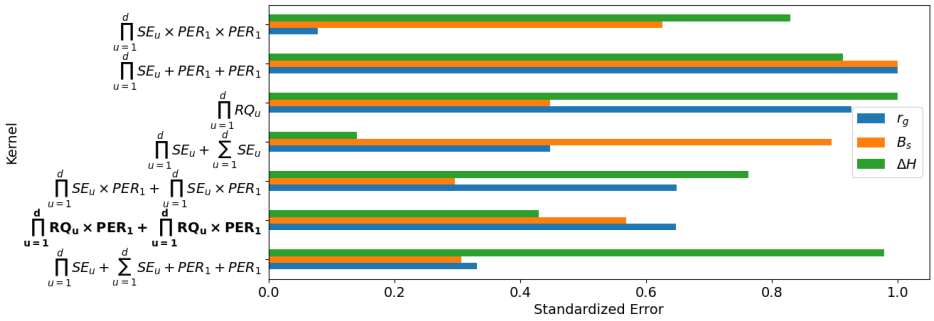
Figure 3.4c suggests that there is a trade-off in the performance of a composite kernel depending on the mobility metric being prioritized and that the impact of the chosen kernel is non-negligible in this context. Notably, while certain kernels may not be optimal for all



(a)



(b)



(c)

Figure 3.4: (a) Loss function progression and convergence behavior of different functional forms for a composite kernel; (b) Generating one week of trips using discovered kernel and the temporal GP; (c) Derived mobility metrics by kernel. Bold denotes the optimal kernel. r_g denotes radius of gyration, B_s denotes spatial burstiness, and ΔH denotes distance between user's imputed and actual home location.

Table 3.1: Optimal kernel parameters for long gap imputation example. Asterisk* denotes the highest lengthscale among categorical/binary variable.

Parameter	Value	Explanation
Initial $p_{a,1}$	1,440 mins = 1 day	Initial value for period length in PER_a
Initial $p_{b,1}$	10,080 mins = 1 week	Initial value for period length in PER_b
η_a	0.27	Optimal weight of kernel component a
η_b	0.73	Optimal weight of kernel component b
$p_{a,1}$	629 mins = 0.44 days	Optimal period length of PER_a
$p_{b,1}$	5,290 mins = 3.67 days	Optimal period length of PER_b
$l_{a,RQ,t_d=3}$	39.4*	Lengthscale of RQ_a for $t_d = 3$
$l_{a,RQ,t_{am}=1}$	18.4*	Lengthscale of RQ_a for $t_{am} = 1$
$l_{b,RQ,t_d=6}$	55.6*	Lengthscale of RQ_b for $t_d = 6$
$l_{b,RQ,t_{am}=1}$	18.7*	Lengthscale of RQ_b for $t_{am} = 1$

metrics, they still provide some degree of insight into the mobility patterns. A kernel that may have a higher loss or error for one metric might still be useful for understanding certain aspects of mobility through other metrics. This may be because the mobility metrics analyzed here are not necessarily aligned with the objective function of the MKL process. We discuss ideas for future studies on this topic in Section 3.5.

In practical applications, composing a kernel would then depend on the specific aspect of mobility one aims to capture. If the goal is to have a balanced model across various metrics, then kernels that show moderate performance across the board might be preferred. However, if the objective is to excel in one particular metric, such as predicting home location with high accuracy, then the kernel with the lowest error for ΔH would be chosen, despite potential compromises in other areas.

Analyzing optimal kernel parameters highlights the capacity of our model to uncover individual mobile data generation processes. The last four rows in Table 3.1 show the highest characteristic lengthscale along each categorical dimension—a high smoothing parameter for an input dimension suggests low variation along that dimension in the function being modeled, and vice versa. We notice that the example user has high correlations in their behavior on Thursdays ($t_d = 3$), Sundays ($t_d = 6$) and in the AM peaks ($t_{am} = 1$) compared to the other days of the week and the PM peak, respectively. The optimal period lengths also suggest that the user's periodicities occur at shorter intervals than initially assumed—the first periodic kernel is optimized at 0.44 days and the second at 3.67 days. Relatedly, the

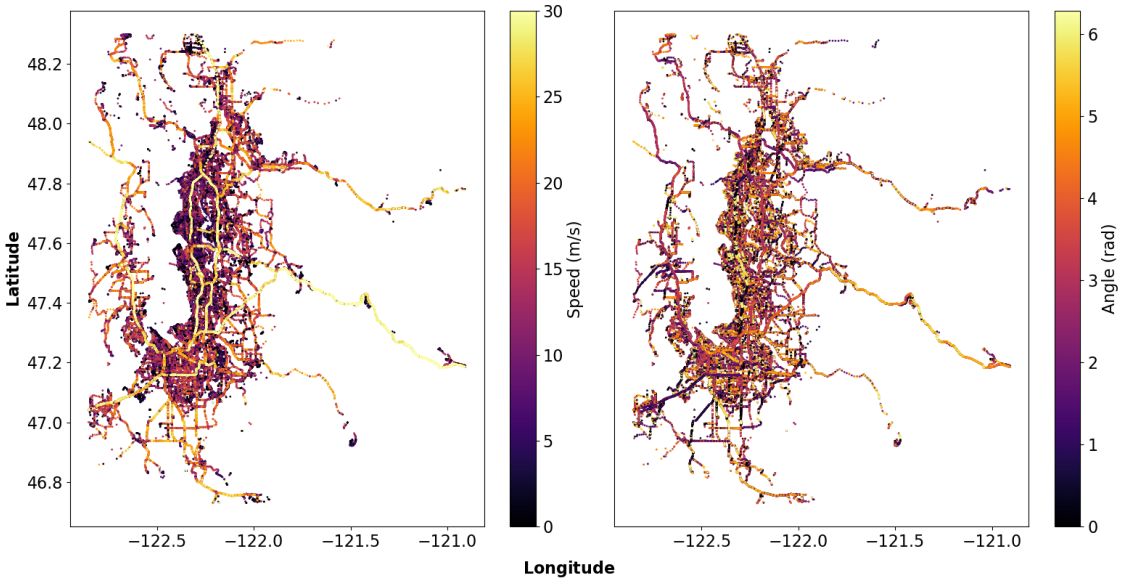


Figure 3.5: Speed and Bearing Constraints in King County, Washington

daily kernel component influences the model less than the weekly component judging by the weights η_a and η_b.

3.4.2 Learning the Physical Kernel

In this section, we discuss the process of learning the optimal physical kernel k_p. Figure 3.5 shows discrete observations of average segment speeds and angular directions observed from 10 users' trips over six months. We note that one can deduce roadway types by the average speed observed on them (e.g., the clear vertical lines on the left are SR-99 and I-5 respectively).

To learn the form of the covariance function for physics-based variables, we apply the greedy MKL algorithm on the training set with the following base kernels: SE (Equation 3.11), and RQ (Equation 3.12), MAT (Equation 3.13). This is one realization of the $\mathbf{P}_{cond} \rightarrow \mathbf{Y}_{cond}$ GP shown on the left side of Figure 3.2. The resulting kernel (Equation 3.20) has an additive form and uses the RQ (Equation 3.12) and MAT 5/2 (Equation 3.13) kernels. Figure 3.6 shows predictions on one week of velocities and bearings after the covariance matrix associated with Equation 3.20 was estimated for two weeks of training data.

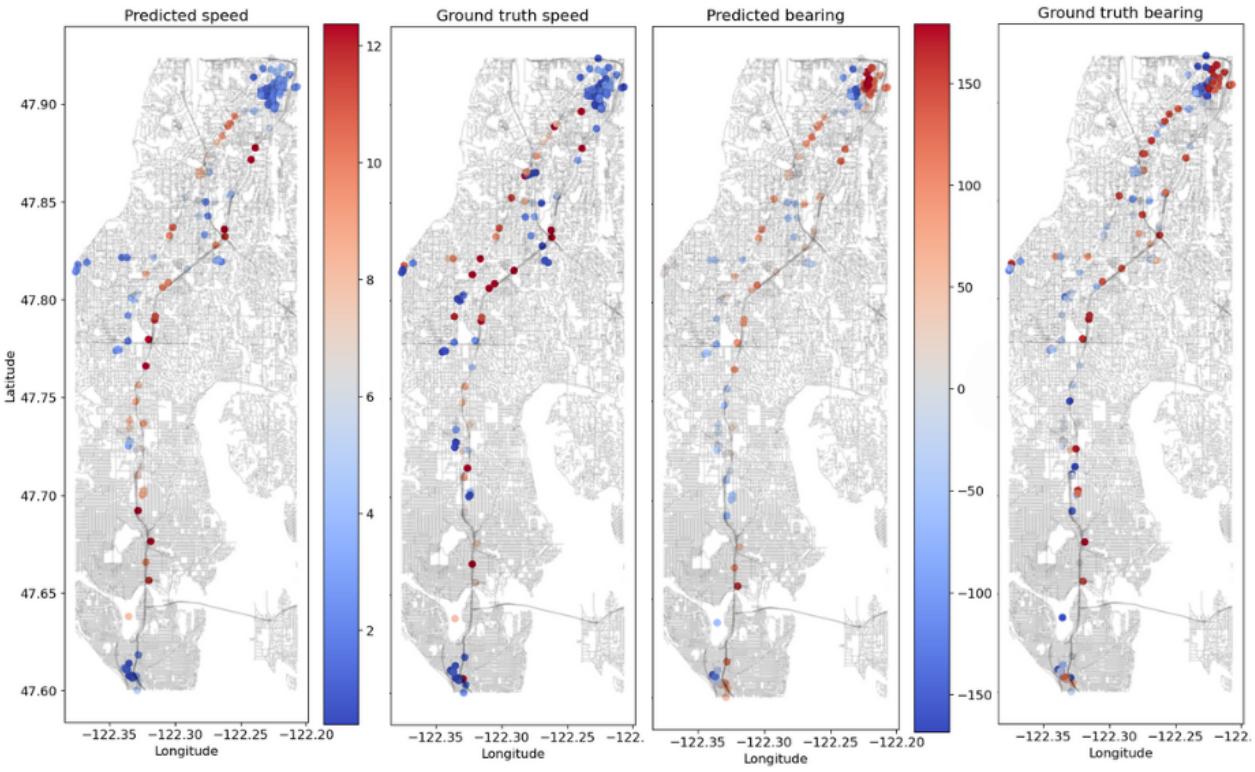


Figure 3.6: Speed and Bearing Predictions and True Observations for a Spectus User

$$
\begin{aligned}
k_p = \eta_{RQ,1}\left(1+\frac{|\mathbf{x}_i-\mathbf{x}_g|^2}{2\alpha_1 l_{RQ,1}^2}\right)^{-\alpha_1} \\
+\eta_{MAT,1}\left(1+\frac{\sqrt{5}|\mathbf{x}_i-\mathbf{x}_g|}{l_{MAT,1}}+\frac{5|\mathbf{x}_i-\mathbf{x}_g|^2}{3l_{MAT,1}^2}\right)\exp\left(-\frac{\sqrt{5}|\mathbf{x}_i-\mathbf{x}_g|}{l_{MAT,1}}\right)+\eta_{RQ,2}\left(1+\frac{|\mathbf{x}_i-\mathbf{x}_g|^2}{2\alpha_2 l_{RQ,2}^2}\right)^{-\alpha_2} \\
+\eta_{MAT,2}\left(1+\frac{\sqrt{5}|\mathbf{x}_i-\mathbf{x}_g|}{l_{MAT,2}}+\frac{5|\mathbf{x}_i-\mathbf{x}_g|^2}{3l_{MAT,2}^2}\right)\exp\left(-\frac{\sqrt{5}|\mathbf{x}_i-\mathbf{x}_g|}{l_{MAT,2}}\right).
\end{aligned}
\tag{3.20}
$$

3.4.3 Passively-collected Mobile Dataset

We randomly select a subset of 33 users with at least 10,000 observations and 3 months of data (from first data point until last). Within the selected 33, we only retain data from the month with the most observations. We minimize Equation 3.10 using the mean and covariance structures in Equation 3.9 to evaluate the performance of the physics-regularized GP formulation against two alternative formulations: One using only physical variables, and

another using only temporal variables. We note that the former is unrealistic in practice, as variables like speed and bearing have to be derived from coordinates (which are assumed missing in our tests). However, we show the results as they illuminate the motivation for integrating the two types of variables.

We use the root mean square error (RMSE) and the mean standardized log loss (MSLL) as our performance metrics—the former is the root of the squared residual between the mean prediction and the target, averaged over the test set; the latter is a probabilistic measure that takes into account the uncertainty of the predictions. The MSLL incorporates the logarithmic transformation of the predicted and observed values, normalized by the associated uncertainty. This allows for a more nuanced assessment of the model's performance, as it captures the relative differences between predicted and observed values in a probabilistic manner. We use these metrics to compare our predictions to various testing sets, allowing us to comment on the model's generalizability and the impact of the physics-regularization on bias and variance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.21}$$

$$MSLL = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\log(\hat{y}_i + 1) - \log(y_i + 1)}{\sigma_i}\right)^2. \tag{3.22}$$

where y_i and $\hat{y}_i$ are respectively the true and predicted values of the dependent variable at index i and σ_i is the predicted standard deviation.

We remove consecutive chunks of data to test the model's accuracy in imputing time series, progressively increasing the size of the training set in each iteration while keeping the size of the test set constant. Specifically, we divided the dataset into three sequential folds for training and testing, ensuring chronological order is maintained. The first split (S1) uses the first and the second weeks of the month as training and testing sets, respectively; the second set (S2) then uses the first two weeks to train and the third week to test. Finally, the third set (S3) uses the first three weeks to train and the last week to test. This approach simulates a real-world scenario where a model is trained on an increasing amount of historical data and tested on the subsequent period. It also allows us to observe the performance of the model as it is exposed to more data over time.

Table 3.2 shows the median performance of each model across the three splits. The Physics-regularized GP outperforms the individual physics-only and temporal-only models

Table 3.2: Models and median test set metrics

Models	**Metrics**								
	RMSE			**MSLL**			**RT**		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
Physical GP	0.105	0.128	0.130	**0.767**	0.816	0.855	**6.69**	**13.9**	**30.3**
Temporal GP	0.107	0.119	0.116	0.885	0.808	0.874	7.99	17.4	38.5
Physics-reg. GP	**0.103**	**0.119**	**0.113**	0.882	**0.779**	**0.849**	11.5	22.8	50.2

Note: S1, S2, and S3 represent three different time-series splits.

in each testing chunk with respect to RMSE and MSLL. The increase in RMSE from S1 to S2 for all models and from S2 to S3 for the Temporal GP indicates a decrease in predictive accuracy as the training set size increases, possibly due to overfitting or increased model complexity. However, the Physics-regularized GP maintains consistently low RMSE values across all splits, demonstrating better generalization as more data is included. The steady performance improvement in this model suggests that the integration of both physical and temporal inputs helps to counteract overfitting and improve predictive accuracy as the dataset grows.

The MSLL is lower for the Physics-regularized GP in two out of the three splits, indicating not only better prediction accuracy but also better calibrated uncertainty estimates. The consistent performance of the Physics-regularized GP across the splits in terms of MSLL suggests that it remains reliable and well-calibrated regardless of the amount of training data. Although the Physics-regularized GP estimates more input dimensions simultaneously than the other two models, its Kronecker product structure keeps the increase in training time relatively modest. The physical-only model has the fastest runtime, but the added computational cost of the Physics-regularized GP is justified by its improvements in both prediction and uncertainty calibration.

3.4.4 GeoLife Dataset

To test the generalizability of our framework across multiple modes and populations, we conduct an experiment on Microsoft Asia's GeoLife dataset (Zheng et al., 2008, 2010, 2009), specifically using the subset of all GeoLife users who had mode labels available. The essence of this experiment is to assess how well our model can generate unobserved trips based on a training set of similar trips. Rather than a time-series-based experimental design where

the training and testing sets are temporally sequential, we use a stratified sampling strategy (shown within Algorithm 7 in Appendix B.6) applied to all of one user's trips. That is, we cluster trips based on their start and end locations, and sample out of the set of similar trips to build the training set. We do this for two reasons: (1) in GeoLife, users may appear in the dataset over the course of years and data continuity is low (i.e., we do not see trips from the same user in consecutive days), making a time-series split more difficult; (2) Ensuring that the training set of trips is similar to the testing set allows us to better isolate the effect of physics-regularization, as a pair of start/end locations can have multiple routes (with presumably different underlying physics).

We assess the proposed exact GP formulation (PIMTGP) against three alternatives: a sparse GP model (SparsePIMTGP), an exact temporal GP using multiple kernel learning but without physics-regularization (TempMTGP), and a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). Unlike probabilistic models, LSTM networks do not inherently provide measures of uncertainty. To facilitate a comparison between LSTM predictions and that of the GPs, we conduct Monte Carlo (MC) simulations with a dropout rate of 0.5 across 70 simulations during inference. This approach introduces randomness during inference to approximate the model estimation uncertainty of the LSTM. It is important to note that this does not equate to the prediction uncertainty provided by the GP models, which is derived from their probabilistic nature. Therefore, while MC dropout helps in assessing the robustness of the LSTM predictions, the comparison in probabilistic measures between LSTMs and GPs should be interpreted with caution as they reflect different aspects of uncertainty. We use the `Torch` implementation of LSTM with a mean squared error loss, optimized via Adam (Kingma and Ba, 2014) as with all other models. We choose hyperparameters via cross-validation and limit the number of MC simulations to 70 to maintain scalability.

The key idea of the sparse GP model is that we can learn a memory-efficient representation of the $n \times n$ covariance matrix using c inducing points, which capture most of the variation in the dataset (where $c << n$) (Snelson and Ghahramani, 2005). The benchmark formulation uses a variational approximation where the predictive distribution is given by

$$p(\mathbf{f}(\mathbf{x}_*)) = \int_o p(f(\mathbf{x}_*|\mathbf{o}))q(\mathbf{o})d\mathbf{o} \tag{3.23}$$

where $\mathbf{o}$ represents the function values at the c inducing points, $q(\mathbf{o}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$. Here,

$\mathbf{m} \in \mathbb{R}^c$ and $\mathbf{S} \in \mathbb{R}^{c \times c}$ are learnable parameters. Its multitask extension leverages the linear model of coregionalization (LMC) which assumes that each task j is the linear combination of some latent functions $\mathbf{g}(\cdot) = [g^{(1)}(\cdot), \ldots, g^{(Q)}(\cdot)]$ and

$$f_j(\mathbf{X}) = \sum_{v=1}^{Q} a^{(v)} g^{(v)}(\mathbf{X}), \tag{3.24}$$

where $a^{(v)}$ are learnable parameters.

Figure 3.7 shows the performance of each model across seven evaluation metrics and three modes of travel–bike, walk, and bus. RMSE and MSLL are defined as before. The mean absolute error (MAE) and the median absolute deviation (MAD) are defined as follows,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{3.25}$$

$$MAD = \text{median}(|y_i - \hat{y}_i|). \tag{3.26}$$

We also assess model performance with metrics specific to curves (or trajectories). These metrics better capture differences in predicted and actual route choice, an important variable in travel behavior research. The curve length (CL) measures the arc-length distance along the curve from the origin, capturing total travel distance in a given trip. Meanwhile, dynamic time warping (DTW) incorporates timestamps of trajectories when considering sequence alignment, capturing differences in speed of travel (Ugurel et al., 2024a). A lower DTW value indicates a closer match whereas a higher value suggests a greater disparity between the compared sequences. Finally, we compare the training runtime (RT) of all methods.

From Figure 3.7, we observe a few things: PIMTGP tends to outperform the other models in the classical error metrics (RMSE, MAE, and MAD) across all modes. In the probabilistic MSLL metric, we observe a clear distinction between GP-based models and the LSTM network. The MKL-equipped TempMTGP seems to be competitive with PIMTGP in this probabilistic measure, highlighting the importance of using Multiple Kernel Learning. In the trajectory comparison metrics (CL and DTW), the outcomes are fairly similar, with TempMTGP and PIMTGP tending to perform slightly better. Finally, taking a look at the runtimes paints a clear picture–the more complex models take longer to train but also tend to perform better at individual data generation. It is also worthy to note that the SparsePIMTGP has similar runtimes to the LSTM model while maintaining the core of our

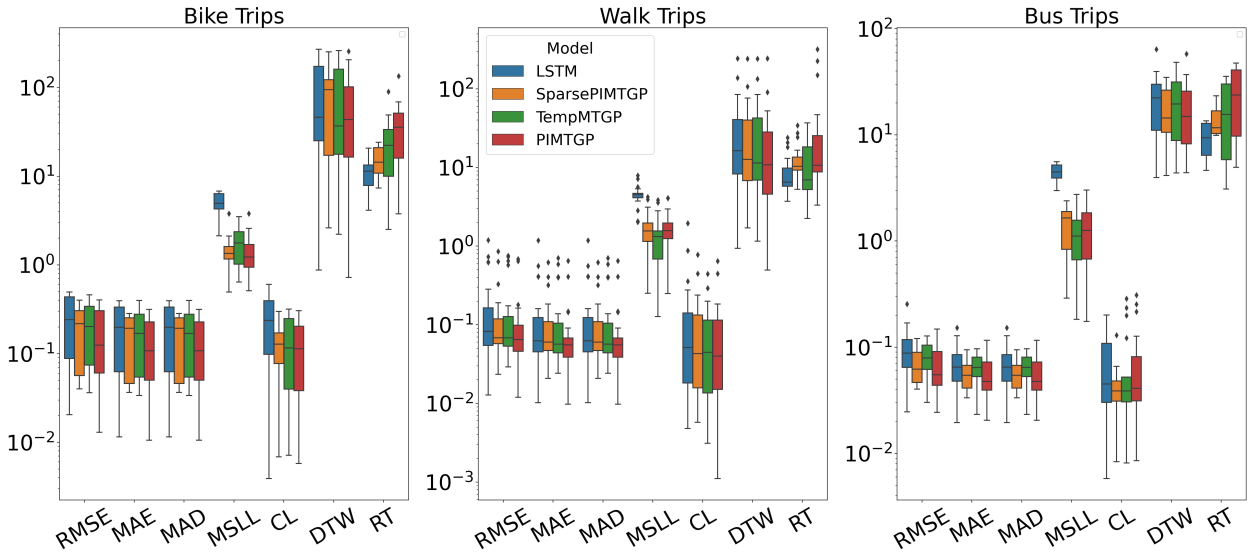


Figure 3.7: Evaluation metrics for all models being compared. Y-axis is shown in log-scale. The lower the better for all metrics.

framework, opening a path to future scalable implementations. We believe this is a promising research avenue.

3.5 Discussion

To realize the potential of creating synthetic population for large scale simulations of individual mobility patterns, we need to develop a flexible and scalable model that can capture individual heterogeneity and also adapt to changes over time. These two properties: flexible and scalable, require nonparametric models whose forms can easily adapt. In this paper, we propose a physics-regularized multi-task Gaussian Process (GP) model for learning kernel structures that characterize individual movement patterns. Though there are other choices for nonparametric models such as deep neural networks, GP models are another promising method that offers several benefits over deep learning-based models. First, GPs are generally more interpretable than deep learning architectures, as choices of basic kernels are based on the nature of the phenomenon of interest (in this context, human mobility patterns) and their hyperparameters such as lengthscale are explainable (they directly tell us the periodicity of the underlying function). A related second point is that kernel methods are more mathematically mature than their deep learning counterparts. Especially in exact GPs, all calculations are done in

closed-form—this signals a potential for theoretical bounds on generalization. Additionally, the probabilistic framework of GPs facilitates robust uncertainty quantification, motivating the use of different metrics to capture this added benefit (Mohamed et al., 2022).

As noted earlier, the study is motivated by an emerging application of passively-generated mobile data that will have great potential when realized, a problem that can hamper the great potential, and the fact that few have looked into the problem at hand. That emerging application is synthetic population generation for large-scale simulations of cities for various needs such as state monitoring, policy formulation and evaluations, and building digital twins. Existing methods such as activity-based models or probability models are inflexible, unable to capture heterogeneity exhibited in individual mobility patterns and adapt to changing patterns over time. The reality now and in the future is that the quality of data is degrading (substantial sparsity or missingness associated with the data) with rising awareness to preserve privacy on the part of both data vendors and consumers. To fulfill the great potential that passively-generated mobile data has for synthetic population generation and to address the increasing issue of sparsity, an understanding of the underlying data generation process behind the observed individual mobile data is needed. This is the key theme of our paper: the development of a flexible modeling framework that can adapt to heterogeneity at the individual level.

We summarize our contributions as follows: We propose a conditional-generative GP framework to generate synthetic individual mobile data that integrates physical knowledge and provably replicates observed travel patterns. To account for heterogeneity at the individual level, we develop a data-driven multiple kernel learning approach to determine the most suitable spatiotemporal kernel for each user. The learned multiple kernel GP models can be viewed as mechanisms to uncover underlying data generation processes that result in the observed big mobile data, which in turn can be used to generate synthetic data mimicking the properties of collected data.

Moving forward, our research opens several avenues for advancement. Future studies can aim to explore alternative formulations for the MKL objective function that allow for the prioritization of different mobility metrics in the synthetic data generation process, tailoring the model to specific applications. Furthermore, new methods in one-step ahead forecasting strategies (i.e., predicting the next data point in a time series based on the observed values up to the current time point) infused with our framework could provide real-time insights and predictions, enhancing the model's applicability in dynamic environments. Finally, to

address the computational burden often associated with GP models, studies should further investigate approximation techniques (as we do with the SparsePIMTGP) as well as algebraic methods, such as exploiting the grid structure in inputs, to streamline the model's efficiency without sacrificing accuracy. These approaches will be key to unlocking the production-level (i.e., scalable) potential of such methods.

Chapter 4

ON PREDICTING SOCIODEMOGRAPHICS FROM MOBILITY SIGNALS

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).

4.1 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 Koppelman, 1994; Lee et al., 2007; Lu and Pas, 1999; McGuckin and Murakami, 1999b; 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., 2016b). 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., 2024b). Metropolitan planning organizations (MPOs) see potential in using them to stitch origin-destination matrices (Alexander et al., 2015a; Iqbal et al., 2014a), site electric-vehicle chargers (Yang et al., 2017a), 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., 2024c). 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. This issue has been studied from a variety of angles, including travel behavior (Auld et al., 2015; Zhang et al., 2024a), 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 interpretablely associated with sociodemographic attributes and that they substantially increase the predictive power of imputation models beyond classical and spatiotemporal features (defined in Section 4.4.1).

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 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 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 chapter is organized as follows: Section 4.2 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 4.3 outlines the datasets we leverage, while Section 4.4 details our methodological approach. Section 4.5 presents our experimental design and the numerical results. We conclude with a discussion of our findings, limitations, and ideas for future work in Section 4.6.

4.2 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.

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., 2024a). 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.

4.3 Datasets

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

Table 4.1: 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)

Category	Variable	2017	2019	2023
Wave	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%**)
HH Income	Under $25,000	410 (**8.07%**)	314 (**6.63%**)	354 (**7.07%**)
	$25k–49,999	642 (**12.6%**)	588 (**12.4%**)	533 (**10.6%**)
	$50k–74,999	744 (**14.6%**)	715 (**15.1%**)	646 (**12.9%**)
	$75k–99,999	728 (**14.3%**)	657 (**13.9%**)	531 (**10.6%**)
	$100k+	2,558 (**50.3%**)	2,460 (**51.9%**)	2,943 (**58.8%**)
Children in HH	0	3,538 (**69.6%**)	3,297 (**69.6%**)	3,371 (**67.3%**)
	1	702 (**13.8%**)	586 (**12.4%**)	604 (**12.1%**)
	2	712 (**14.0%**)	580 (**12.2%**)	777 (**15.1%**)
	3+	130 (**2.57%**)	250 (**5.28%**)	277 (**5.53%**)

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 4.1 highlights relevant descriptive statistics from our post-processed version of this dataset.

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 modeling.

Although our motivation is to enable sociodemographic 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.

4.4 Methodology

Section 4.4 details our methodology. 4.4.1 formalizes the mobility-graph representation and defines the activity- and trip-level covariates used in prediction. In 4.4.2, we outline definitions and metrics that allow us to operationalize uncertainty quantification in our context. Finally, 4.4.3 describes the theory behind the MT approach as well as the specific neural architecture we leverage.

4.4.1 Characterizing travel behavior with mobility graphs

Daily activity chains are encoded as directed graphs $G = (V, E)$, where 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*, which describe how frequently, evenly, and in what sequence specific activities are pursued; and (2) *trip (edge) features*, which summarize mode diversity (i.e., tendency to use different travel modes), co-travel composition (i.e., share of trips taken alone vs. with others), and daily travel motifs (i.e., minimal, recurring subgraphs that capture

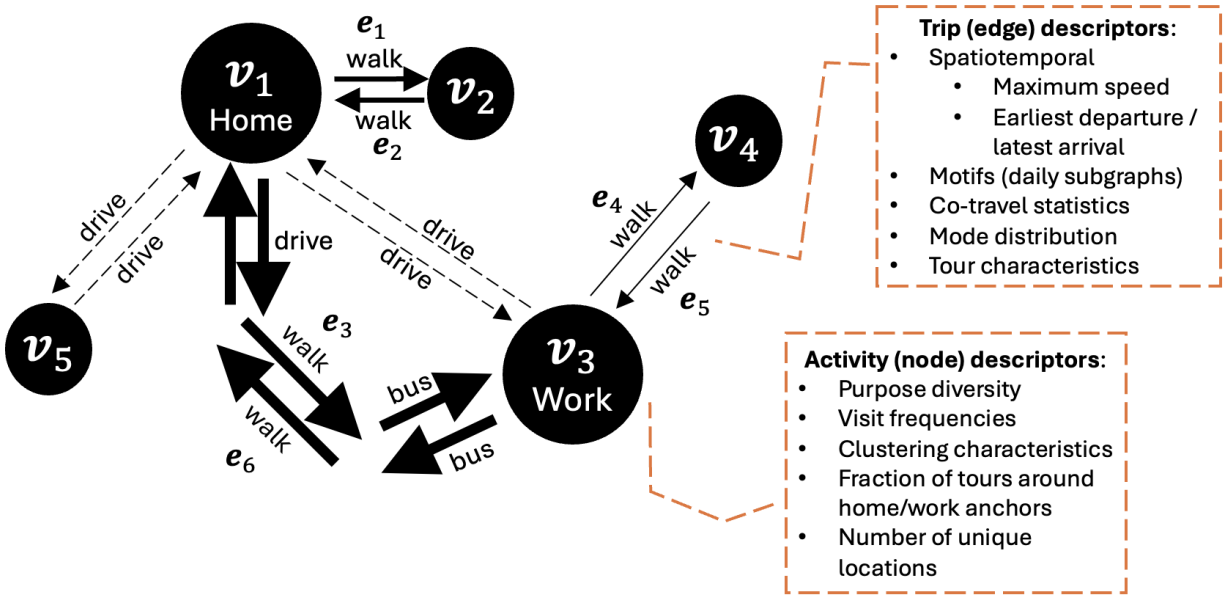


Figure 4.1: Example daily mobility graph where the edges are chronologically numbered. Width of trip arrows corresponds to frequency of visits over observation period. Dashed arrows denote known trips that are not observed on this day.

the daily structure of trip sequences). We give precise definitions for metrics to quantify the above in Section 4.4.1.2 (see Figures 4.1–4.3).

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 4.5.1). 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.

4.4.1.1 Activity (node) features

Let $\mathbf{x} = (x_1, \ldots, x_N)$ be the frequency vector of an individual's N distinct origin-destination (OD) purpose pairs, where each pair connects two activity purposes (e.g., going home, to

the gym, to 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):

$$H_t = -\sum_{i=1}^{N} \frac{x_i}{T} \log_2 \left(\frac{x_i}{T} \right), \tag{4.1}$$

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, $x_{(1)} \leq \ldots \leq x_{(N)}$, 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:

$$G = 1 + \frac{1}{N} - \frac{2}{NT} \sum_{k=1}^{N} C_k. \tag{4.2}$$

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).

To quantify global cohesion in the mobility graph, we compute the *global clustering coefficient* C_{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 $\tau_{\triangle}$ denote the number of closed triplets (triangles) and $\tau_{\wedge}$ the number of connected triplets (wedges). Then:

$$C_{\text{glob}} = \frac{3\tau_{\triangle}}{\tau_{\wedge}}. \tag{4.3}$$

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 the overall cohesion of the network, 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:

$$\bar{c} = \frac{1}{N} \sum_{v=1}^{N} \frac{2t_v}{k_v(k_v - 1)} \tag{4.4}$$

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.

4.4.1.2 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_{trips} and n_{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_{rush}; defined as 7–9 AM or 4–6 PM) and the fraction taken on weekends (f_{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* $\rightarrow$ *home* $\rightarrow$ *store* becomes *home* $\rightarrow$ 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 4.2). 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 $f_k = m_j/M$ together with the motif entropy $H_m = -\sum_j \frac{m_j}{M} \log_2(\frac{m_j}{M})$ 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 $\mathbb{1}[mm]$ be the indicator that a given tour involves at least two distinct transport

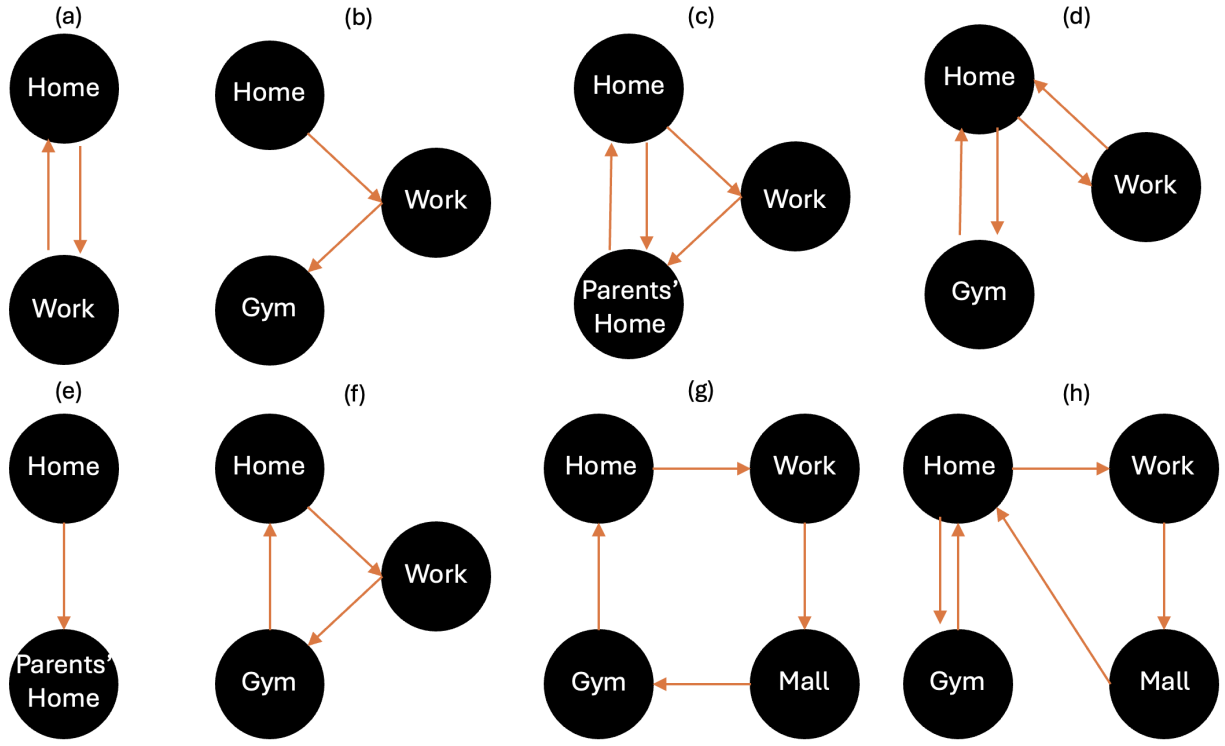


Figure 4.2: Illustration of daily travel motifs (Schneider et al., 2013; Wu et al., 2019); (a) Out-and-back; (b) Chain; (c) Cycle-chain; (d, h) Double-cycle; (e) Single-no-return; (f, g) Single-Cycle

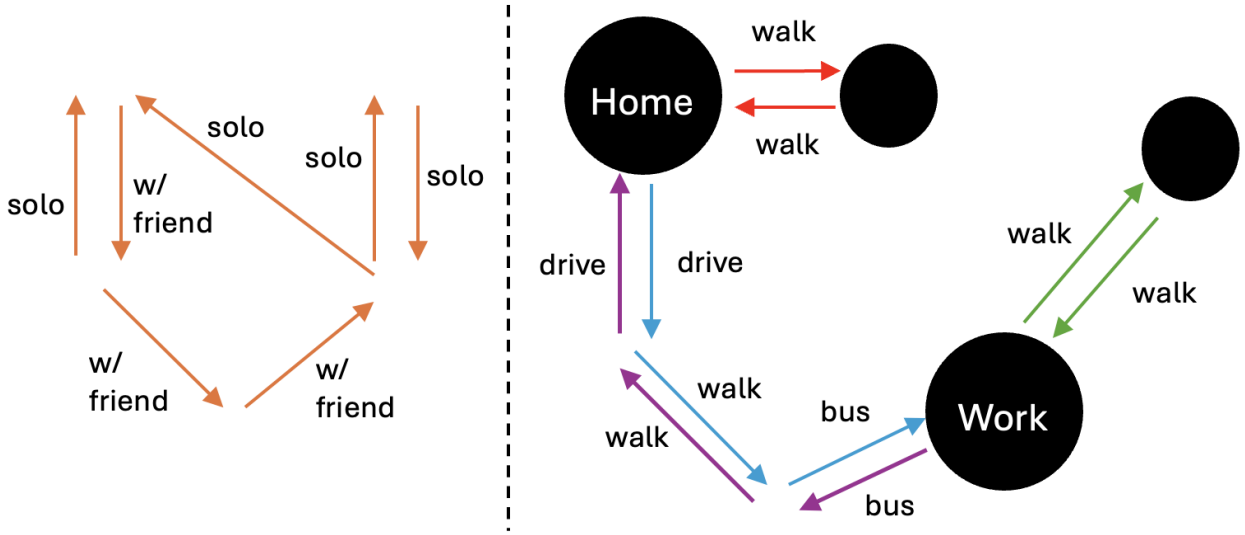


Figure 4.3: Illustration of selected metrics; (left) example travel day. In this case, $n_{\text{trips}} = 7$ and $f_{\text{comp}} = 3/7$; (right) each color denotes a tour to/from the anchors. In this case, $n_{\text{tour}} = 4$ and $f_{\text{mm}} = 2/4$.

modes. The *multi-modal fraction* is then

$$f_{mm} = \frac{\mathbb{1}[mm]}{n_{\text{tour}}} \tag{4.5}$$

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 4.5).

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

$$f_{\text{solo}} = \frac{n_{\text{solo}}}{n_{\text{trips}}}, \quad f_{\text{hh}} = \frac{n_{\text{hh}}}{n_{\text{trips}}}, \quad f_{\text{nonhh}} = \frac{n_{\text{nonhh}}}{n_{\text{trips}}}, \tag{4.6}$$

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 4.3.

4.4.2 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.

4.4.2.1 Definitions

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

$$\Pr(\hat{Y} = Y \mid \hat{P} = p) = p \quad \text{for all } p \in [0, 1], \tag{4.7}$$

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.

4.4.2.2 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". However, it evaluates predictions at a single decision rule (i.e., choose the argmax), and is therefore sensitive to class imbalance and agnostic to the quality of the full probability vector. 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.

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 $\{(x_i, y_i)\}_{i=1}^N$ as inputs and K classes, a probabilistic classifier specifies a categorical distribution q, where each sample has predicted class probabilities $\hat{\mathbf{p}}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})$ satisfying $\sum_k \hat{p}_{ik} = 1$. The likelihood of the observed label y_i under the model for input x_i is the scalar $q(y_i \mid x_i) = \hat{p}_{i,y_i}$. Then, the NLL averages the negative log of those probabilities:

$$\text{NLL} = -\frac{1}{N}\sum_{i=1}^{N} \log q(y_i \mid x_i) = -\frac{1}{N}\sum_{i=1}^{N} \log \hat{p}_{i,y_i}. \tag{4.8}$$

NLL is minimized when the predicted probabilities match the true conditional distribution (Gneiting and Raftery, 2007). Intuitively, NLL rewards placing 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 $\hat{y}_i = \arg\max_k \hat{p}_{ik}$ be the predicted label, and let the confidence of the prediction be the highest predicted probability: $\hat{p}_i = \max_k \hat{p}_{ik}$. 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 $I_m = \left(\frac{(m-1)}{M}, \frac{m}{M}\right]$. The accuracy and average confidence within bin B_m are:

$$\text{acc}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \tag{4.9}$$

$$\text{conf}(B_m) = \frac{1}{|B_m|}\sum_{i \in B_m} \hat{p}_i, \tag{4.10}$$

where y_i is the true 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):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{4.11}$$

The difference between accuracy and confidence for a given bin is called the *calibration gap*. Thus, a perfectly calibrated model attains ECE = 0. In our experiments, we pre-specify $M = 15$ and use equal-width bins, following common practice in the calibration literature (Naeini et al., 2015). This number provides a practical balance between resolution and statistical stability: too few bins obscure fine-grained miscalibration, while too many yield noisy estimates due to small sample counts per bin. Equal-width binning ensures that confidence intervals are evenly spaced and facilitates comparability across models and tasks.

One related diagnostic tool we leverage is the *reliability diagram* (e.g., shown in Figure 4.7), which visually represents model calibration. These diagrams plot expected sample accuracy (Eq. 4.9) as a function of confidence (Eq. 4.10). If the model is perfectly calibrated, the diagram should trace the identity line. Points below the diagonal indicate over-confidence, while those above indicate under-confidence. Note that reliability diagrams do not display the proportion of samples in each bin, and thus cannot be used to estimate how many samples are calibrated.

4.4.3 Multitask Learning

Multitask 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 $t \in \{1, \dots, T\}$, where each task provides samples $(X, y_t) \sim \mathcal{D}_t$ and a task-specific loss ℓ_t. We learn a shared representation $h = g(X; \theta)$ and task-specific predictors $f_t(h; \phi_t)$ by minimizing the weighted multi-task risk:

$$\min_{\theta, \{\phi_t\}} \sum_{t=1}^{T} w_t \, \mathbb{E}_{(X, y_t) \sim \mathcal{D}_t} \left[\ell_t \left(f_t(g(X; \theta); \phi_t), y_t \right) \right] + \Omega(\theta, \phi_t), \tag{4.12}$$

where w_t are task weights, θ are the shared parameters of the representation h, ϕ_t are the task-specific parameters of head f_t, and Ω is a standard parameter regularization term. In practice, we optimize the empirical version using mini-batches sampled from the joint dataset and backpropagate the sum of per-task losses; the shared parameters θ receive gradients from all tasks and thereby encode common structure.

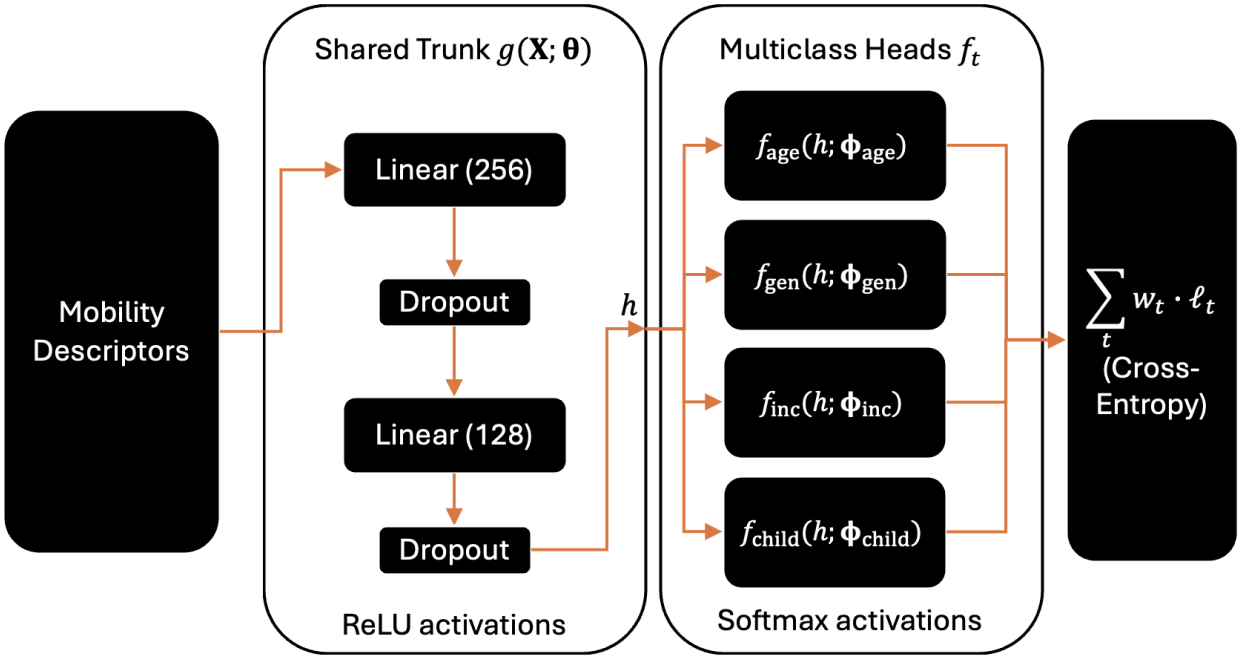


Figure 4.4: Shared-trunk multitask architecture. Mobility descriptors are mapped to a shared representation $h = g(X;\theta)$ by a two-layer feed-forward network with ReLU activations and dropout. Four task-specific heads f_t produce class probabilities via softmax. Training minimizes the cross-entropy loss $\sum_t w_t \cdot \ell_t$ where we set the weights to be equal.

In this study, we adopt hard parameter sharing in a feed-forward network (shown in Figure 4.4). 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 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, where we use equal weights for each task.

The mobility descriptors given in Section 4.4.1 may capture latent routines that are jointly informative for several sociodemographic attributes. Thus, sharing this representation across the learning tasks pools statistical evidence, improving sample efficiency for sparsely labeled

or imbalanced tasks, and regularizing 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.

4.5 Experiments

We showcase our results. In 4.5.1, we analyze correlations between our descriptors and sociodemographic targets, which fit well-known tropes in the literature. 4.5.2 assesses the extent to which our features improve the predictability of demographics and model calibration. In 4.5.3, we demonstrate the value of multitask learning, which improves outcomes in data scarce regimes or on test sets that differ from the training set.

4.5.1 Linkages between mobility descriptors and sociodemographics

Figure 4.5 shows the largest magnitude correlations between selected sociodemographics and our feature set. Age is most strongly related to the share of school-purpose travel ($\rho \approx -0.50$) and the drive-alone fraction ($\rho \approx +0.30$), 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.

Table 4.2 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

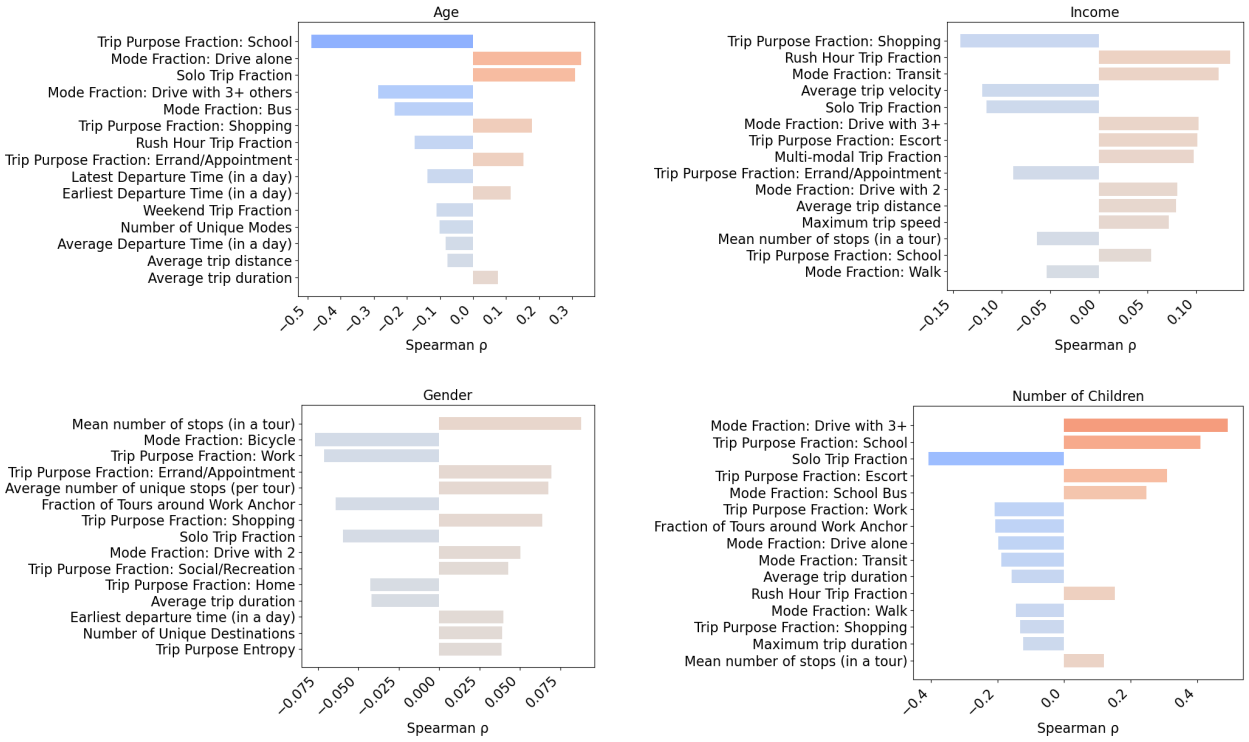


Figure 4.5: Largest magnitude Spearman rank correlations (all $p < 0.001$) between mobility descriptors and demographics. (top left): age; (top right): household income; (bottom left): gender; (bottom right): number of children. Bars to the left indicate negative associations; to the right, positive.

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.

4.5.2 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 4.4.1.

Table 4.2: 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.]

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 *overall* split, we created a 70/10/20 train/validation/test partition on the combined waves of the PSRC survey. On the *cross-temporal* split, we trained and validated on the 2017 and 2019 waves and tested on the 2023 wave, a period with documented shifts in mobility behavior due to the aftermath of COVID-19 (see differences in wave composition in Table 4.1). 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 4.6 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 C.1 (Tables C.1-C.8), and show that the proposed descriptors broadly improve separability and likelihood across models. However, the most comprehensive feature set (+CT) does not always yield the best performance—approximately 44% (14 out of 32) of the top metric values (shown in red) come from simpler sets. Performance trends are largely stable across model types, suggesting that the observed improvements stem primarily from the covariates rather than model-specific effects. That no single model dominates across all tasks and metrics further underscores the absence of a one-size-fits-all solution: practitioners must select models and feature sets based on their specific performance goals.

For the multi-task DNN shown in Figure 4.6, 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 C.1-C.8, 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.

Finally, the proposed descriptors generally improve both class separability (AUROC) and

model fit (NLL). Under the overall 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 4.7 highlights that, under several settings (e.g., Age in the overall 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).

4.5.3 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 C.1 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. To avoid overfitting as training data size got small, we used only the classical and spatiotemporal features.

Figure 4.8 displays our findings. On the overall split (top four rows of Figure 4.8), the MT network and and its STVs achieve comparable top-1 accuracies and AUROC at full data. As the training fraction shrinks, the relative patterns diverge by task. Age, gender, and number of children exhibit a modest MT advantage in accuracy and AUROC at intermediate and low fractions, indicating effective transfer of shared structure when labels are scarce. By contrast,

HH income shows mixed patterns, particularly at the smallest fractions, suggesting that shared representation can sometimes blur task-specific distinctions (i.e., negative transfer). In these cases, the single-task networks appear better able to specialize when the shared structure among targets is weak.

The uncertainty metrics favor the multitask learning-based model. NLL remains lower for MT than STVs as data become scarce, and ECE is lower for MT for three of the four fractions over each sociodemographic target. This indicates that even when discrimination gains are inconsistent, the shared representation can still regularize learning by preventing overconfidence and smoothing probability estimates.

Under the cross-temporal generalization setting (bottom four rows of Figure 4.8), the MT network is again competitive with STVs in accuracy and AUROC across most fractions, though the advantages are uneven. Gains are most apparent for HH income, number of children, and to a lesser extent, age, while gender shows little benefit or mild negative transfer. The improvements in NLL are more pronounced than in the overall split, suggesting that multitask learning helps regularize against overconfidence under distribution shift. By pooling representational strength across targets, the 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). This stabilizing effect is especially useful in real-world deployment, where shifts in behavior over time (e.g., due to pandemics, fuel prices, or policy changes) can erode model reliability.

4.6 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.

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., 2024b) and later adapted to sociodemographic inference.

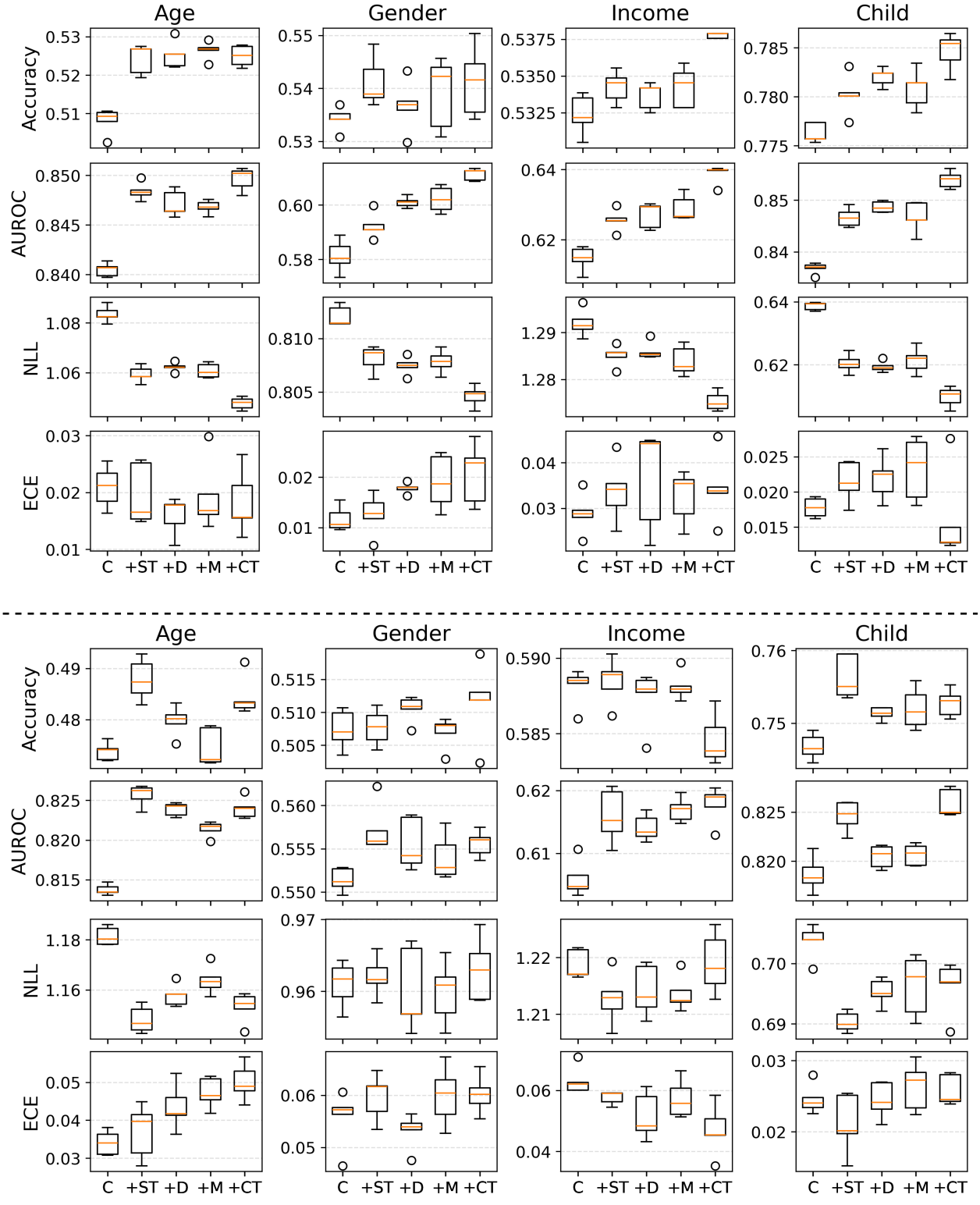


Figure 4.6: Marginal changes in top-1 accuracy (higher = better), AUROC (higher = better), NLL (lower = better), ECE (lower = better) as more features are added. (top four rows) Overall split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data.

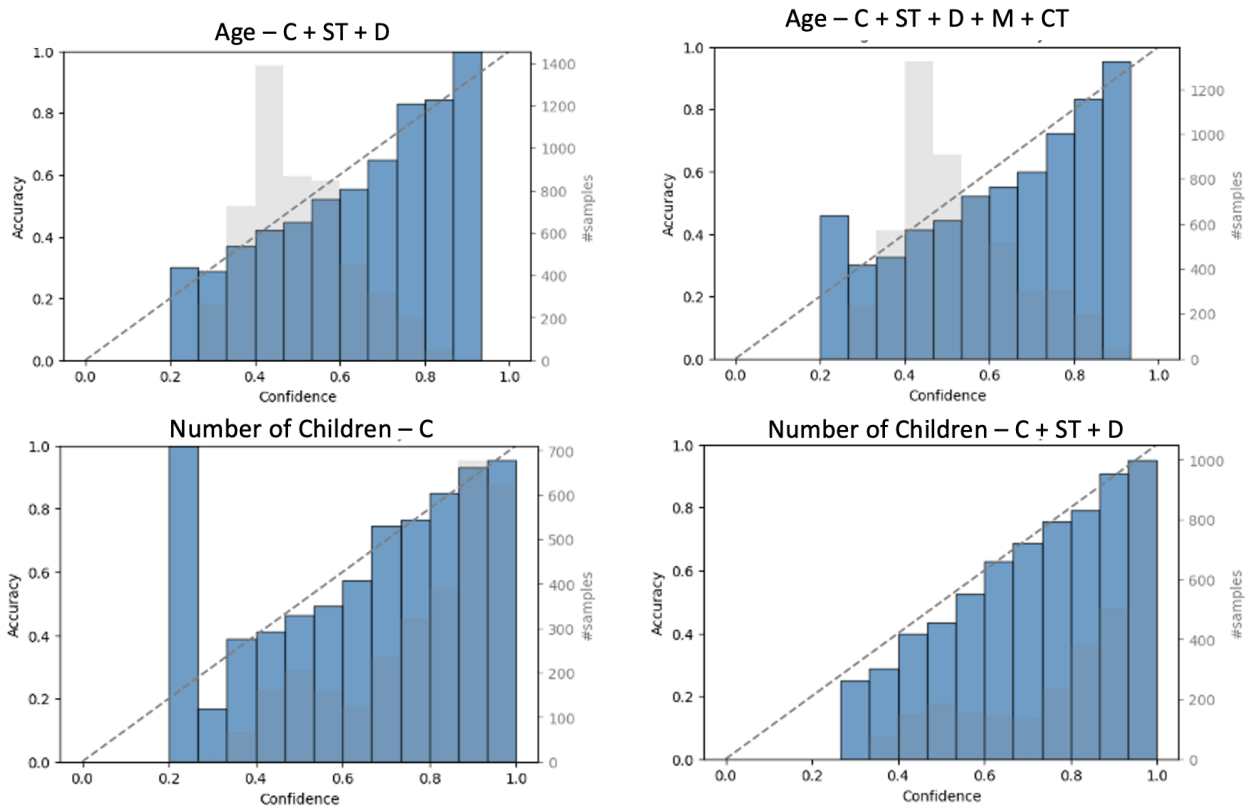


Figure 4.7: Reliability diagrams for representative settings. (TOP) Overall split, Age task with C+ST+D features (left) and All features (right). (BOTTOM) Cross-temporal split, Number of Children task with C (left) and C+ST+D (right). Bars show empirical accuracy within 15 equal-width confidence bins; 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.

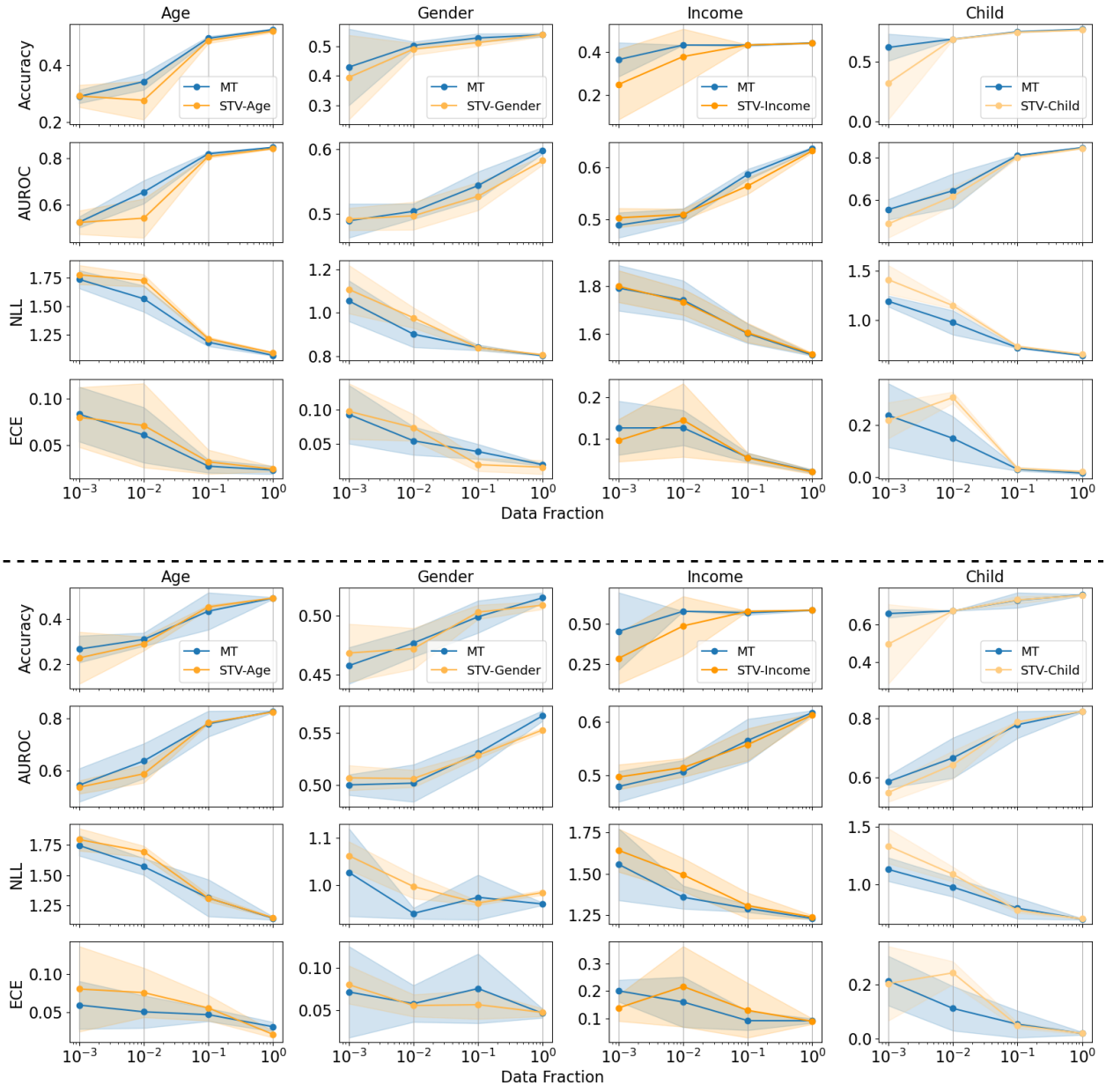


Figure 4.8: Performance of the MT variant (in blue) compared to ST variants (in orange) at different fractions of training data. (top four rows) Overall split; (bottom four rows) training with the 2017/2019 data and testing on the 2023 data.

Chapter 5

MPO'S USES OF AND NEEDS FOR BIG DATA

Big data products offer a new paradigm to understand and analyze human mobility patterns, a primary interest of long-range transportation planners. However, questions remain about how widely these datasets are used in practice and what factors limit their broader adoption in planning workflows and decision-making. In this chapter, we first provide a comprehensive review of the literature on the use of big data in metropolitan planning, with a focus on practical applications, perceptions, and institutional barriers to adoption. Then, we present the perspectives of more than 50 planners from MPOs across the United States, collected at a workshop in May 2024. While we found a range of use cases, there was also a tendency to focus on a narrow set of applications. *Transparency*, *regulation*, and *legitimacy* emerged as the primary factors influencing adoption decisions.

5.1 Introduction

Long-range transportation planners (LRTPs) are tasked with supporting decision-makers in shaping multimillion-dollar investments in future mobility. In recent years, the increased availability and commercialization of big data have resulted in the rise of mobility-analytics vendors who offer a suite of products, hereafter referred to as big data products (BDPs), designed to help planning agencies make more informed decisions. These companies aggregate passively collected data from multiple sources (such as smartphones, in-vehicle GPS trackers, connected-vehicle data, etc.), to provide insights into fine-grained human mobility behavior. Their data products typically include key information on individuals' movements over time, capturing patterns such as travel routes, trip frequencies, and timing, which can be invaluable for understanding mobility trends and planning future infrastructure investments.

While BDPs offer a new way to understand and analyze mobility patterns, we know little about their actual use in transportation planning. It is unclear how widely these datasets or products are being utilized by planners or to what extent they are influencing decision-making processes. Moreover, the full potential of these datasets, especially in areas such as forecasting, equity analysis, and resilience planning, remains largely unexplored. Understanding the gaps

in usage and application is key to unlocking their value for long-term transportation planning.

This chapter is motivated by two recurring themes identified through ad-hoc conversations with various practitioners. These themes are: (A) the use of BDPs has been limited to a narrow range of applications and they are not the preferred solution to many problems faced by LRTPs; and (B) only a small number of MPOs (primarily ones with sufficient resources) use BDPs. Thus, this chapter addresses the following questions:

1. How do planners perceive big data products offered by mobility analytics companies?
2. What factors influence an MPO's decision to adopt big data products and services?
3. What are the various use cases for big data in regional transportation planning?

5.2 Literature Review

The interest MPOs have in leveraging big data for transportation planning has generated considerable discussion (in the form of studies, reports, and memorandums). Here, we first outline common use cases noted in the literature. We then examine the key barriers to adoption, focusing on how planners perceive big data products and the institutional, technical, and organizational factors that shape their uptake.

5.2.1 Use Cases

A primary use of big data is to calibrate and validate travel demand models. Many MPOs now incorporate archived traffic speed/travel-time data and volume estimates from private vendors to improve their regional models (Bauer et al., 2014). For example, the Maricopa Association of Governments (Phoenix area) has purchased probe-based speed data (e.g., products offered by INRIX) for several years and uses the National Management Research Dataset, a national traffic speed dataset provided by the FHWA, to validate model travel times and congestion patterns (FHWA, 2020). These passive data allow agencies to check that model outputs (e.g. travel speeds, delays) reflect observed conditions, leading to model improvements. The Capital District Transportation Committee (Albany, NY MPO) likewise compared model-predicted speeds and delays with an archived traffic sensor dataset and discovered the model had been underestimating off-peak and non-recurring delay (FHWA, 2020). Thus, insights from big data help refine model parameters and ensure forecasts are grounded in real-world conditions

Another key application is using big data to fill gaps in traffic counts and OD data for planning studies. Some MPOs face declining coverage or frequency of traditional traffic counts (i.e., due to limited budgets or disruptions like the pandemic) and have turned to private "Big Data" vendors to estimate measures like Annual Average Daily Traffic (AADT) on road segments. In one case, the Tulsa, OK MPO obtained crowdsourced traffic volume estimates from a vendor to populate their roadway network map when local agencies stopped collecting counts, and validated these estimates against the MPO's own data (StreetLight, 2025). Big data derived from mobile devices are also used to generate OD matrices and travel flows by time of day and mode. Several MPOs have experimented with big data products to derive regional trip distributions or corridor-specific OD patterns, which can supplement or reduce reliance on infrequent household travel surveys (KimleyHorn, 2021). These OD datasets have been applied for corridor studies, sub-area analyses, and to understand regional travel trends (e.g. inter-county commuting flows) that would be expensive to capture via surveys. However, as discussed later, agencies often use these new data cautiously, usually as a "second source" to compare against traditional survey-based OD data or model outputs.

LRTPs also see potential for big data in analyzing emerging transportation modes and trends that are not well covered by traditional data. Big data from smartphones and app-based services could theoretically reveal the magnitude and patterns of rideshare demand. Indeed, many MPOs aspire to use big data to understand the impacts of transportation network companies (TNCs), yet implementation has lagged. Interviews with large MPOs found that few had obtained or analyzed TNC trip data despite interest (KimleyHorn, 2021). Only a couple of agencies (e.g. New York Metro and Chicago regions) were exploring data-sharing agreements to acquire TNC trip records through local governments, and even those had no concrete analysis completed due to lack of internal expertise to manage the data (KimleyHorn, 2021).

Big data is also emerging as a tool to evaluate transportation accessibility, although efforts are relatively nascent. Because passively collected datasets can be disaggregated spatially and temporally, they hold promise for analyzing differences in travel times, mobility patterns, and service levels across different communities. For instance, location-based data can be used to do distributional comparisons across various criteria, including those based on spatial factors (e.g., the location of a proposed project in relation to the community it affects) as well as user-based factors (e.g., the number of users of the project that will benefit) (Boyd et al., 2024). Nonetheless, researchers caution that these data pose limitations, particularly

concerns over representation and bias, as not all population groups are equally captured by smartphones or connected vehicles (Richardson, 2021; Wang et al., 2025).

5.2.2 Factors Influencing Adoption

Though use cases are plenty, adoption among MPOs remains limited. Many agencies cite a combination of financial, organizational, and technical barriers. The following subsections examine key factors shaping adoption decisions, including concerns about data quality, resource constraints, staffing capacity, and leadership support. As one NCHRP workshop attendee put it bluntly, "Our big data issues are straightforward: we don't have the technology, money, or the skills." (Pecheux et al., 2020).

A central barrier to big data adoption among MPOs is concern over data quality, transparency, and methodological validity. Most vendors rely on proprietary algorithms to infer trips, scale samples, or impute missing information, yet rarely disclose these processes in detail. Elements like sampling timeframes, sample bias corrections, or definitions of trip purpose often diverge from the standards used in regional models (KimleyHorn, 2021). As a result, many MPOs avoid relying on these data for core planning tasks outside of limited domains like congestion monitoring.

The situation is exacerbated by proprietary database structures and software platforms, which limit interoperability and reinforce dependence on specific vendors (Pecheux et al., 2020). These restrictions make it difficult for agencies to conduct independent validation or integrate the data into existing workflows. Because the raw inputs are transformed through opaque processes, MPOs cannot easily adjust for known biases in the way they might with traditional data sources like travel surveys (Singh et al., 2022). Furthermore, differences in methodology across vendors or changes over time can undermine the comparability of key performance measures, particularly in long-range planning contexts (Bauer et al., 2014).

Cost is one of the most frequently cited barriers to the adoption of big data among MPOs. High upfront and ongoing subscription fees can be prohibitive, especially for smaller agencies operating under tight budgets (KimleyHorn, 2021). MPOs must assess whether the perceived benefits of big data justify diverting resources from other essential planning tasks. In many cases, the cost-benefit calculus is unfavorable, especially when planners are uncertain how much of the data will ultimately be useful for investment decision-making (Bauer et al., 2014).

Some agencies have attempted creative workarounds. For example, one MPO used federal Transportation Management Area (TMA) funds to subsidize vendor subscriptions. Yet even

when funding is available, the lack of transparent pricing from vendors can deter adoption. Opaque service-based fee structures make it difficult to anticipate long-term costs or compare alternatives (KimleyHorn, 2021).

Perceptions of cost are further shaped by technical and staffing needs. Cloud-based platforms may be viewed as more expensive than local storage options, partly because they often require hiring or contracting with specialized personnel (i.e., data analysts and engineers) to securely manage data acquisition and processing (Pecheux et al., 2020; Singh et al., 2022). Moreover, big data platforms often provide far more information than agencies can immediately use. Filtering that data down into actionable insights requires significant staff time and effort, adding another layer of hidden cost (Bauer et al., 2014).

Furthermore, the successful use of big data within MPOs depends not only on acquiring the data itself, but also on having the staff capacity to manage, interpret, and apply it effectively. Implementing big data typically requires new information technology systems, staff training, and sometimes even revised hiring strategies (KimleyHorn, 2021). Generally, however, smaller transportation agencies lack the dedicated personnel needed to carry out these tasks. Specialized roles such as data engineers, database administrators, or GIS analysts may be absent, or distributed across partner organizations without centralized oversight (Richardson, 2021).

Agencies also report difficulty recruiting and retaining staff with experience in big data management. Personnel trained in traditional data systems are more widely available and better understood by leadership, while staff familiar with modern analytics pipelines, cloud-based platforms, or large-scale data integration are scarcer and harder to retain. Furthermore, siloed data cultures and internal concerns over data security or potential misuse often inhibit both intra-agency collaboration and external data sharing, limiting the agency's ability to adopt modern data practices (Pecheux et al., 2020).

Training and knowledge sharing also matter. Some MPOs have learned from peer agencies or attended workshops (e.g., FHWA's training on using archived operations data FHWA 2020) which build confidence to adopt big data. Nonetheless, in many agencies, the lack of staff time and resources remains a core constraint. Without sustained investment in staffing and training, and without leadership recognition of the importance of data governance, agencies are unlikely to move beyond pilot projects or basic uses of big data.

Finally, leadership support plays a pivotal role in whether big data tools are adopted within MPOs. If executives or board members do not understand or trust the value of big

data, planning staff often struggle to move initiatives forward. A national review noted that "leadership often does not fully understand the value of big data," which creates institutional hesitation (Pecheux et al., 2020). In many cases, skepticism stems from doubts about data quality, limited familiarity with new technologies, or uncertainty about how big data contributes to planning outcomes (Richardson, 2021).

Without a visible champion for data sharing at the executive level, it can be difficult for staff to secure the resources and flexibility needed to experiment with or integrate new data products. Even when analysts make the case for cloud-based platforms or advanced analytics tools, they often fail to demonstrate an immediate return on investment to decision-makers, especially in agencies that prioritize cost-efficiency or regulatory compliance (Pecheux et al., 2020).

Indeed, literature on organizational psychology may provide helpful guidance in this context. Teams adopt new practices more readily when leaders create psychological safety (i.e., a shared belief that it is safe to take interpersonal risks) because staff can experiment, surface errors, and learn without fear of blame. This enables careful piloting of unfamiliar data products (Edmondson, 1999). Agencies also need absorptive capacity, the ability to recognize the value of external knowledge, assimilate it, and apply it. Leadership investment in related skills and routines strengthens this capacity (Cohen and Levinthal, 1990). Finally, technology-acceptance research shows that executive messaging and support affect user beliefs about performance gains, required effort, social norms, and facilitating conditions, which together shape adoption intentions and use (Venkatesh et al., 2003).

5.3 Methods

We conducted an interactive workshop at the Association of Metropolitan Planning Organizations (AMPO) Planning Tools & Training Symposium in Albuquerque, New Mexico, in May 2024. The workshop was attended by over 50 participants, all of whom held roles related to LRTPs in some capacity.

We used Mentimeter to facilitate real-time engagement and gather background information. Then, based on the answers to the question, "What best describes your role in your MPO?" (Table 5.1), we divided the room into three roughly even-sized groups and conducted focus group discussions tailored to each group's role and expertise.

Figure 5.1: Bar plot of answers mentioned in response to "In which area does your MPO operate?"

5.4 *Findings*

5.4.1 Descriptive Statistics of Participants

The workshop saw participation from MPO staff from across every geographical region in the United States, with the Northeast and the Pacific Northwest (PNW) having the most participants (see Figure 5.1). When asked about their MPOs' experience with BDPs, most of the participants hinted at some proficiency through in-house usage or via consultants—only 17 percent of the respondents had no prior hands-on experience with BDPs (see Table 5.1). The prior exposure of workshop participants suggests their insights would be more informed and valuable to the discussion.

We also asked the respondents about their role within the MPO as shown in Table 5.1. They were almost evenly divided between technical staff who primarily work with data (33%), managers who primarily supervise people, projects, or consultants (23%), and staff who do both (37%). Hearing from both technical and managerial staff is crucial, as their perspectives on the use of BDPs may differ. Technical staff are typically more engaged with the practical, day-to-day data challenges, while managerial staff are more focused on broader strategic concerns such as resource allocation and decision-making processes. Both viewpoints offer complementary insights into how BDPs can be integrated into MPO operations.

We note that the goal of the workshop was not necessarily to attain a representative

Table 5.1: Distribution of answers to exploratory questions about the audience

Answer	**Percentage***
"What experience does your MPO have with big data?"	
None.	7%
We're learning about it but don't have hands-on experience.	10%
We've experimented with it through trial subscriptions or purchases.	17%
We've used it in-house or through consultants to support work programs or planning/policy decisions.	66%
"What best describes your role in your MPO?"	
I primarily manage people, projects, or consultants; focus on the big picture; and do NOT handle data.	23%
I primarily work with data to support my MPO's work programs and projects.	33%
I perform some management, consider the big picture, AND directly process/analyze data.	37%
I perform other roles that little to do with data.	7%

**rounded to nearest integer*

sample of MPOs, but rather to present a wide range of perspectives that exist.

5.4.2 Perceptions of Big Data Products

Most planners view big data as complementary to traditional data collection methods—neither a complete replacement nor without value. During the focus group discussions, we found some differences in how managers perceived BDPs compared to that of technical MPO staff. Managers perceived big data as a tool for updating functional classification of various roadways and doing related analyses, such as deriving mode splits, normalizing crash rates, and

Figure 5.2: Word cloud of answers to "What word comes to mind when you think of "big data"? (98 responses)

assessing land use impacts and project effectiveness. In contrast, technical staff highlighted the role of BDPs in data validation and completion, serving as proxies for incomplete datasets and validating travel demand models. Figure 5.2 illustrates the most common associations planners have with big data. Adjectives such as complex, overwhelming, and expensive were frequently mentioned, along with nouns like cell phones, privacy, AI, machine learning, and passive data.

5.4.3 Factors Influencing MPO Adoption

When asked, "What would help you gain confidence or have trust in using big data?", transparency stood out as the most significant factor influencing planning organizations' adoption of big data, as participants emphasized the need for clear information on how data is collected, processed, and validated (see Figure 5.3). Legitimacy was the second most common concern, involving the ability to assess BDP quality via reviews, proven experiences, and

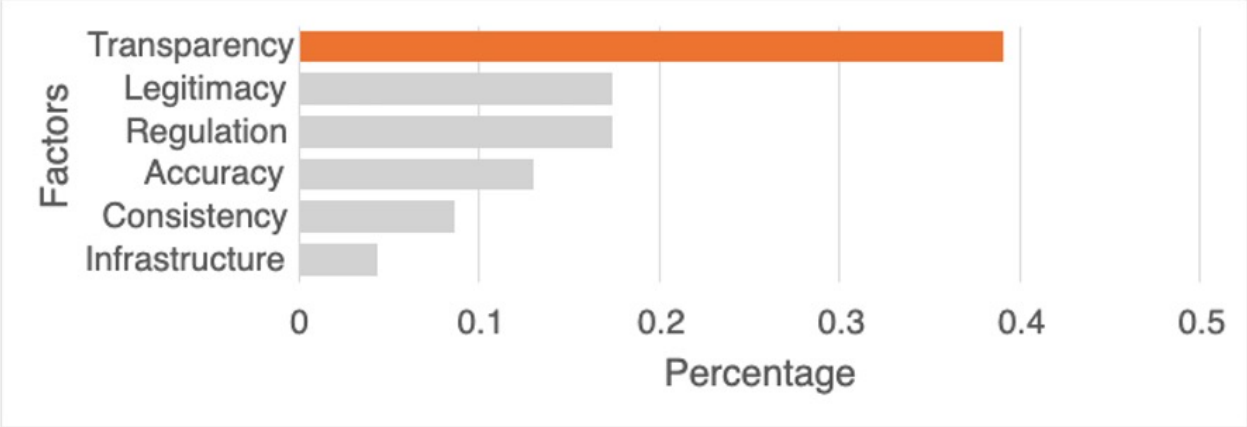


Figure 5.3: Bar plot of factors mentioned in response to "What would help you gain confidence or have trust in using big data?"

academic studies (thereby reducing risk). Regulation ranked third, referring to data quality guidance from transportation authorities and decision makers with vested interest.

As a follow-up, we asked participants what they would like to know about BDPs that they do not already know. The responses included questions regarding the quality of the data like "How accurate is it?" and "Is the accuracy consistent over time?", questions that expand on transparency from the supplier side like "What steps were taken between the raw and processed datasets?", as well as operational-level questions about the process of adopting BDPs like "Will costs change over time?" and "How do we 'keep' the data we buy after our contract ends?".

5.4.4 Use Cases

We found a wide range of BDP use cases among MPOs, as detailed in Table 5.2. However, the majority of LRTPs focus on a small number of BDP applications. The most frequently mentioned benefits of partnering with mobility-analytics companies were the ability to easily derive peak traffic volumes for low functional classification roads, understand turn counts at intersections, and measure pedestrian and cyclist activity. This seemed to agree with our hypothesis that many MPOs who adopt BDPs only utilize them for a narrow scope of problems.

Among the use cases shown in Table 5.2, we found little-to-no prior literature on using big data to study work-housing imbalances and job accessibility. Other equity-related applications,

such as identifying locations of high noise and vehicle emissions on highway networks and assessing community resiliency, were similarly under-explored. More research in these areas could provide valuable insights to address transportation inequities and improve public health and community resilience.

Table 5.2: Types of questions MPO staffers have solved or plan on solving, categorized

Category	**Short Description**	**Long Description**
Equity-related	Quantifying work-housing imbalance and job accessibility (Zhang et al., 2017; Zhou et al., 2018)	Analyzing the distribution of jobs relative to housing to ensure equitable access to employment opportunities.
	Identifying locations of high noise and vehicle emissions on highway network (Lan et al., 2020)	Locating areas with high levels of noise and emissions to address environmental justice concerns.
	Assessing community resilience (Sarker et al., 2020)	Assessing the transportation network's ability to withstand and recover from adverse events such as natural disasters.
	Origin-destination studies (Alexander et al., 2015b; Iqbal et al., 2014b; Liu et al., 2022b)	Studying travel patterns between different origins and destinations to understand mobility flows.
	Mode choice analysis (Gong et al., 2012; Phithakkitnukoon et al., 2017; Yang et al., 2022)	Analyzing the factors influencing the selection of different modes of transport (e.g., car, public transit).
Safety-related	Safety planning and project identification (Ambros et al., 2024)	Identifying and prioritizing safety projects to reduce accidents and enhance road safety.
	Crash hotspot analysis (Xie et al., 2017)	Pinpointing locations with high accident rates to implement targeted safety measures.
	Mobile device usage while driving (Ahlström et al., 2020; Khurana and Goel, 2020)	Monitoring mobile device usage to identify distracted driving behaviors and develop interventions.
	Lane change analysis (Park et al., 2018)	Analyzing lane change behaviors to improve road design and traffic flow management.

Planning-related	Complete Street studies (Bian et al., 2023)	Planning and designing streets that accommodate all users, including pedestrians, cyclists, and motorists.
	Visitor analysis (Miah et al., 2017; Reif and Schmücker, 2020)	Understanding the travel behaviors and patterns of visitors to optimize transportation services for tourism.
	Validation of observed travel behavior data (Chen et al., 2016b; Liu et al., 2014; Uğurel et al., 2024)	Comparing big data insights with observed travel behaviors to ensure accuracy and reliability.
	EV charging demand estimation (Yang et al., 2017b)	Forecasting the demand for electric vehicle charging stations to support the growing number of EVs.
	Transit planning and routing (Hadjidimitriou et al., 2021; Lu et al., 2021)	Designing and optimizing public transit routes to improve efficiency and coverage.
	Bicycle and pedestrian trip planning (Lee and Sener, 2020; Yu et al., 2020)	Tracking bicycle and pedestrian trips to support infrastructure improvements and promote active transportation.
	Freight movement tracking (Akter et al., 2023)	Monitoring the movement of goods to enhance freight logistics and reduce congestion.
	Land use changes (Yin et al., 2021)	Assessing the impact of land use changes on transportation networks and planning accordingly.
Operational-related	Land border crossing monitoring (Sakhare et al., 2024)	Providing real-time information on wait times and traffic conditions at land border crossings.
	Through trip volume estimation (Huntsinger and Ward, 2015)	Measuring the volume of trips passing through a region without stopping to better understand transit patterns.
	Turn count estimation (Kan et al., 2019)	Counting vehicle turns at intersections to optimize signal timing and improve traffic flow.
	Demand estimation during special events (Pereira et al., 2015)	Analyzing transportation demand and patterns during special events to enhance planning and resource allocation.

Chapter 6

CONCLUSION, DISCUSSION, AND FUTURE WORK

This dissertation examines how behaviorally informed machine learning methods can improve the reliability of mobility data and, in turn, support more transparent and defensible transportation planning decisions. Chapter 2 demonstrates that both short- and long-duration gaps in GPS traces can be addressed using a Bayesian learning framework that explicitly captures the temporal periodicities underlying human travel patterns. Chapter 3 extends this idea by constructing individualized function spaces, enabling the generation of synthetic mobility data that more faithfully replicates observed behavior. Chapter 4 then investigates the challenges of predicting sociodemographic attributes from mobility signals, showing that limited sample sizes and distribution shifts across time and geography impose inherent performance ceilings and underscore the need for uncertainty quantification and alternative learning strategies such as multi-task learning. Finally, Chapter 5 draws on a qualitative workshop with long-range transportation planners to document practical use cases, points of friction, and the institutional and organizational barriers that shape the adoption of big data products in real planning workflows.

This chapter situates these contributions within the broader landscape of transportation planning and engineering. It examines data governance and transparency challenges that arise when relying on proprietary pipelines, the growing mismatch between vendor practices and the scientific evidence surrounding imputation, uncertainty, and representativeness, and the implications of PCM based products for travel demand modeling and forecasting. It also highlights the increasing importance of ML literacy for planners and engineers as big data tools become embedded in everyday workflows. The chapter concludes by outlining several avenues for future research, including the development of foundation models for human mobility, methods to expand the effective amount of training data, privacy preserving computational frameworks, and the institutional infrastructure needed to support responsible and scalable adoption of big data in public sector planning.

6.1 Discussion

A central implication of this work is the need for stronger data governance frameworks that support transparency without compromising privacy. The individualized models proposed in Chapter 3 can be shared through privacy-preserving protocols that avoid exposing raw traces while still enabling reproducibility and third-party auditing. Establishing such protocols would give agencies more confidence in relying on these tools and would facilitate comparative evaluations across regions. In the longer term, the field may benefit from a national mobility data baseline: a shared, standardized, and continually updated data pool that public agencies and researchers can build upon. This idea, analogous to the National Weather Service's mandate, would lower entry barriers, reduce dependence on proprietary pipelines, and create an open benchmark for evaluating vendor products.

The findings in this dissertation also point to a growing mismatch between vendor practices and the scientific evidence on how mobility data should be handled. Many commercial providers treat imputation as a purely computational task. In contrast, Chapters 2 and 3 show that transforming raw GPS traces into behaviorally coherent trajectories is far from straightforward. It requires explicit assumptions about trip-making rhythms, anchor point structures, daily cycles, and route feasibility. When these assumptions are hidden behind proprietary code, agencies effectively lose control over the behavioral content of the datasets they rely on. Similarly, vendors often distribute aggregate traffic volumes with no associated uncertainty and no demographic auditing, masking representativeness biases that can distort downstream analyses. The methods developed here demonstrate that uncertainty and demographic skew are not nuisances to suppress but properties that must be quantified, communicated, and carried forward through planning workflows. Ignoring uncertainty can encourage a false sense of precision and can skew policy decisions toward overconfident interpretations of noisy signals.

Chapter 4 speaks directly to the issue of representativeness. It shows that accurate prediction of sociodemographic attributes such as age, gender, income, and household structure is constrained by small survey samples and distribution shifts across time and geography. There is a ceiling on achievable accuracy even under careful model design, which means that inferred labels should be interpreted probabilistically rather than as ground truth. At the same time, these models provide a practical tool for diagnosing who is under represented in PCM datasets. By comparing the predicted sociodemographic composition of PCM users to survey or census benchmarks, agencies can identify systematic blind spots and design

weighting schemes, targeted outreach, or complementary data collections. In this sense, sociodemographic inference is less a silver bullet fix for bias than a diagnostic layer that makes representativeness issues visible and actionable.

These methodological advances also have direct implications for travel demand modeling and forecasting. More complete and behaviorally coherent trajectories, as developed in Chapters 2 and 3, can strengthen activity based models by improving how daily rhythms, tour structures, and multi stop chains are represented and calibrated. PCM derived traces and synthetic data can be aggregated into OD flows that capture temporal and spatial heterogeneity at resolutions that traditional surveys cannot support. This enables a wider range of applications, including anomaly detection, special event demand forecasting, and quasi experimental evaluations of network changes or policy interventions using before and after PCM signals. Incorporating uncertainty from imputation and sociodemographic inference into these models would allow long range forecasts and project appraisals to reflect the true confidence bounds of the underlying data, rather than treating datasets as fixed and error free.

Finally, these discussions and some of the results we show in Chapter 5 suggest that ML literacy must become a core competency within transportation agencies. The goal is not for planners or engineers to train new neural networks, but to critically interrogate the assumptions, data transformations, confidence scores, and representativeness characteristics of the products they procure. As PCM-based tools become more influential in shaping planning analyses and downstream investments, the profession needs a workforce capable of evaluating models not only on accuracy but on transparency, behavioral validity, and uncertainty quantification.

6.2 *Future Work*

This dissertation opens several promising directions for advancing behaviorally informed mobility modeling and its integration into planning practice. In the era of large language models, a natural next step is the development of foundation models for human mobility. Emerging large multimodal transformers can jointly encode mobility traces, points of interest, text-rich contextual data, and sociodemographic attributes, offering a flexible backbone for a wide range of downstream tasks (Wu et al., 2024b; Zhang et al., 2024b). Training such models on continental-scale datasets and fine-tuning them for specific metropolitan regions would provide a shared representational layer that captures both universal and region-specific mobility structure. Embedding privacy-preserving mechanisms, such as differential privacy during training or federated learning across data providers, will be essential to ensure that these models are deployable in environments where data sensitivity and institutional constraints remain significant.

At the modeling level, a key opportunity lies in the integration of PCM-enhanced datasets with public-sector ABMs. Many existing ABMs rely on synthetic populations built from infrequent surveys or censuses (Brinckerhoff, 2010), which age quickly and struggle to reflect nuanced telecommuting patterns or emerging modes of travel. Calibration is often limited by sparse validation data and by aggregate count measures that hide within-day heterogeneity. PCM-derived trip data can help address these gaps by capturing temporal patterns, tour structures, and neighborhood-level variability that surveys may miss (Tozluoğlu et al., 2025). Future work should develop methods for fusing PCM with HTS-based populations, updating or reweighting agents as conditions evolve, and incorporating PCM-based OD estimates into scenario analysis and network design. An important piece of this agenda will be explicit uncertainty propagation, so that ABM forecasts reflect both structural model uncertainty and the noise inherent in high-frequency mobility data.

The practical deployment of these models raises important questions around privacy-preserving computation. Federated learning offers a pathway for training individualized imputation or inference models without centralizing raw traces (Ezequiel et al., 2022), and on-device imputation and feature extraction can reduce privacy risks by ensuring that only aggregated or derived quantities leave the device (Qi et al., 2017). These approaches, however, come with tradeoffs (Charles and Konečný, 2021). Centralized training is often simpler to implement, can be more data efficient, and may achieve higher accuracy when data are highly unbalanced across participants. Distributed frameworks introduce communication overhead

and coordination challenges when data are not independent across clients. They also face cybersecurity risks such as model inversion and poisoning attacks, which can undermine both privacy and data integrity (Wang et al., 2024). Future work will need to clarify when federated or on-device approaches offer meaningful privacy gains over well-governed centralized systems, and how to design architectures, legal agreements, and monitoring tools that balance competing interests (i.e., utility vs. privacy) in public-sector planning contexts.

Another question that warrants deeper investigation is how to overcome the fundamental sample size limitations of rich behavioral datasets such as household travel surveys. As shown in Chapter 4, predictive performance tends to have a moderate upper bound when trained solely on HTS data, reflecting both the small size of these surveys and their limited coverage of the joint mobility-sociodemographic space. Semi-supervised learning (SSL) methods, specifically *self-training*, offer a principled path forward by expanding the effective training set through the incorporation of pseudo-labeled observations (Lee, 2013; Van Rooyen and Williamson, 2018). Building calibrated, uncertainty-aware SSL pipelines could meaningfully enhance our ability to enrich PCM datasets with respondent-like attributes, thereby improving downstream planning analyses without requiring substantially larger surveys.

Finally, an obvious research direction involves the institutional infrastructure surrounding big data products. As PCM datasets become embedded in planning workflows, independent audits of vendor products, much like financial or algorithmic audits in other sectors (Costanza-Chock et al., 2022; Liu et al., 2022a), will be critical. These audits should evaluate data completeness, imputation logic, uncertainty communication, and demographic representativeness, providing agencies with an evidence-based basis for procurement and use. The methodological tools introduced in this dissertation offer building blocks for such a framework.

BIBLIOGRAPHY

Abkowitz, M. D. (1981). An analysis of the commuter departure time decision. *Transportation* 10, 283–297

Ahlström, C., Wachtmeister, J., Nyman, M., Nordenström, A., and Kircher, K. (2020). Using smartphone logging to gain insight about phone use in traffic. *Cognition, Technology & Work* 22, 181–191

Ak, C., Ergönül, Ö., and Gönen, M. (2018). Structured gaussian processes with twin multiple kernel learning. In *Asian Conference on Machine Learning (ACML)*. 65–80

Akter, T., Hernandez, S., and Camargo, P. V. (2023). Freight Operational Characteristics Mined from Anonymous Mobile Sensor Data. *Transportation Research Record* 2677, 236–247. doi:10.1177/03611981231158639. Publisher: SAGE Publications Inc

Alberini, A., Burra, L. T., Cirillo, C., and Shen, C. (2021). Counting vehicle miles traveled: What can we learn from the nhts? *Transportation research part D: transport and environment* 98, 102984

Alessandretti, L., Aslak, U., and Lehmann, S. (2020). The scales of human mobility. *Nature* 587, 402–407

Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour* 2, 485–491. doi:10.1038/s41562-018-0364-x. Publisher: Nature Publishing Group

Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015a). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58, 240–250. doi:10.1016/j.trc.2015.02.018

Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015b). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58, 240–250. doi:10.1016/j.trc.2015.02.018

Alvarez, M. A., Luengo, D., and Lawrence, N. D. (2013). Linear latent force models using gaussian processes. *IEEE transactions on pattern analysis and machine intelligence* 35, 2693–2705

Ambros, J., Elgner, J., Valentova, V., Bak, R., and Kiec, M. (2024). Proactive safety assessment of urban through-roads based on GPS data. *Archives of Transport* 69, 113–125. doi:10.61089/aot2024.gpa7v104. Number: 1

Apple (2021). *Apple*

Auld, J., Mohammadian, A. K., Oliveira, M. S., Wolf, J., and Bachman, W. (2015). Demographic Characterization of Anonymous Trace Travel Data. *Transportation Research Record* 2526, 19–28. doi:10.3141/2526-03. Publisher: SAGE Publications Inc

Ban, X. J., Chen, C., Wang, F., Wang, J., Zhang, Y., et al. (2018). *Promises of data from emerging technologies for transportation applications: Puget sound region case study.* Tech. rep., United States. Federal Highway Administration

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., et al. (2018). Human mobility: Models and applications. *Physics Reports* 734, 1–74

Barla, A., Odone, F., and Verri, A. (2003). Histogram intersection kernel for image classification. In *Proceedings 2003 international conference on image processing (Cat. No. 03CH37429)* (IEEE), vol. 3, III–513

Batista, S. F., Cantelmo, G., Menéndez, M., and Antoniou, C. (2022). A gaussian sampling heuristic estimation model for developing synthetic trip sets. *Computer-Aided Civil and Infrastructure Engineering* 37, 93–109

Bauer, J., Pack, M., Giragosian, A., Kehoe, N., Evans, J., Voorhies, K., et al. (2014). *Use of Archived Operations Data in Transportation Planning.* Tech. rep., Saxton Transportation Operations Laboratory

Baxter, J. (2000). A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research* 12, 149–198. doi:10.1613/jair.731. ArXiv:1106.0245 [cs]

Bayarma, A., Kitamura, R., and Susilo, Y. O. (2007). Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. *Transportation Research Record* 2021, 55–63

Ben-Moshe, B., Elkin, E., Levi, H., and Weissman, A. (2011). Improving accuracy of gnss devices in urban canyons. In *CCCG*. 511–515

Berke, A., Doorley, R., Larson, K., and Moro, E. (2022). Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (New York, NY, USA: Association for Computing Machinery), SAC '22, 964–967. doi:10.1145/3477314.3507230

Bhat, C. R. and Koppelman, F. S. (1994). A structural and empirical model of subsistence activity behavior and income. *Transportation* 21, 71–89. doi:10.1007/BF01119635

Bhat, C. R. and Koppelman, F. S. (1999). *Activity-Based Modeling of Travel Demand* (Boston, MA: Springer US). 35–61. doi:10.1007/978-1-4615-5203-1_3

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

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task gaussian process prediction. *Advances in neural information processing systems* 20

Boyd, T., Dinehart, T., and Williams, K. (2024). *Uses and Limitations of Big Data for Evaluating Transportation Equity*. Tech. rep., National Institute for Congestion Reduction, Tampa, FL. doi:10.5038/cutr-nicr-y3-1-11

[Dataset] Brinckerhoff, P. (2010). Design and Development Plan for the MAG CT-RAMP Activity-Based Model (ABM)

Cao, W., Wu, Z., Wang, D., Li, J., and Wu, H. (2016). Automatic user identification method across heterogeneous mobility data sources. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. 978–989. doi:10.1109/ICDE.2016.7498306

Caruana, R. (1997). Multitask Learning. *Machine Learning* , 41–75doi:10.1023/A:1007379606734

Charles, Z. and Konečný, J. (2021). Convergence and Accuracy Trade-Offs in Federated Learning and Meta-Learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (PMLR), 2575–2583. ISSN: 2640-3498

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

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

Chen, C., Wang, R., Bansal, P., Chen, L., Ugurel, E., Zhang, Y., et al. (2026). From biases to opportunities: leveraging location-based-service (lbs) data for next-generation transportation planning. *Transportation Research Part C: Emerging Technologies* 182, 105416

Chen, Y., Hosseini, B., Owhadi, H., and Stuart, A. M. (2021). Solving and learning nonlinear pdes with gaussian processes. *Journal of Computational Physics* 447, 110668

Cipra, T., Trujillo, J., and Robio, A. (1995). Holt-winters method with missing observations. *Management Science* 41, 174–178

Cohen, W. M. and Levinthal, D. A. (1990). Absorptive Capacity: A New Perspective on Learning and Innovation. *Administrative Science Quarterly* 35, 128–152. doi:10.2307/2393553. Publisher: [Sage Publications, Inc., Johnson Graduate School of Management, Cornell University]

Costanza-Chock, S., Raji, I. D., and Buolamwini, J. (2022). Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F. (2022). Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing* 92, 88

Daisy, N. S., Millward, H., and Liu, L. (2018). Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. *Journal of Transport Geography* 69, 150–162

Deng, T., Zhang, K., and Shen, Z.-J. M. (2021). A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *Journal of Management Science and Engineering* 6, 125–134

Ding, S., Huang, H., Zhao, T., and 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)*. 1–9. doi: 10.1109/ICCCN.2019.8847051. ISSN: 2637-9430

Doi, S., Mizuno, T., and Fujiwara, N. (2021). Estimation of socioeconomic attributes from location information. *Journal of Computational Social Science* 4, 187–205. doi: 10.1007/s42001-020-00073-w. PMID: 32838050 PMCID: PMC7271143

Dorfman, R. (1979). A Formula for the Gini Coefficient. *The Review of Economics and Statistics* 61, 146–149. doi:10.2307/1924845. Publisher: The MIT Press

Douglas, D. H. and Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization* 10, 112–122

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes*. Ph.D. thesis, University of Cambridge

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., and Zoubin, G. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning* (PMLR), 1166–1174

Eagle, N. and Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral ecology and sociobiology* 63, 1057–1066

Edmondson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44, 350–383. doi:10.2307/2666999. Publisher: SAGE Publications Inc

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

Ezequiel, C. E. J., Gjoreski, M., and Langheinrich, M. (2022). Federated Learning for Privacy-Aware Human Mobility Modeling. *Frontiers in Artificial Intelligence* 5. doi:10.3389/frai.2022.867046. Publisher: Frontiers

Feng, J., Yang, Z., Xu, F., Yu, H., Wang, M., and Li, Y. (2020). Learning to Simulate Human Mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA: Association for Computing Machinery), KDD '20, 3426–3433. doi:10.1145/3394486.3412862

[Dataset] FHWA (2020). Applying Archived Operations Data in Transportation Planning

[Dataset] Fiedler, D., Čáp, M., Nykl, J., Žilecký, P., and Schaefer, M. (2019). Map Matching Algorithm for Large-scale Datasets. ArXiv:1910.05312 [cs, eess] version: 1

Frias-Martinez, V., Soguero, C., and Frias-Martinez, E. (2012). Estimation of urban commuting patterns using cellphone network data. In *Proceedings of the ACM SIGKDD international workshop on urban computing*. 9–16

Gammelli, D., Peled, I., Rodrigues, F., Pacino, D., Kurtaran, H. A., and Pereira, F. C. (2020). Estimating latent demand of shared mobility through censored gaussian processes. *Transportation Research Part C: Emerging Technologies* 120, 102775

Gao, L., Huang, H., Ye, J., and 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. doi:10.1016/j.trc.2024.104550

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems* 31

Gibbs, M. N. (1998). *Bayesian Gaussian processes for regression and classification*. Ph.D. thesis, Citeseer

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102, 359–378. doi:10.1198/016214506000001437. Publisher: Informa UK Limited

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research* 12, 2211–2268

Gong, H., Chen, C., Bialostozky, E., and Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems* 36, 131–139. doi:10.1016/j.compenvurbsys.2011.05.003

Gong, X., Huang, Z., Wang, Y., Wu, L., and Liu, Y. (2020). High-performance spatiotemporal trajectory matching across heterogeneous data sources. *Future Generation Computer Systems* 105, 148–161

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature* 453, 779–782

Hadjidimitriou, N. S., Lippi, M., and Mamei, M. (2021). A Data Driven Approach to Match Demand and Supply for Public Transport Planning. *IEEE Transactions on Intelligent Transportation Systems* 22, 6384–6394. doi:10.1109/TITS.2020.2991834. Conference Name: IEEE Transactions on Intelligent Transportation Systems

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

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

Hanson, S. and Huff, J. O. (1988). Repetition and day-to-day variability in individual travel patterns: Implications for classification. *Behavioural Modelling in Geography and Planning, Croom Helm, London* , 368–398

Hao, Q., Chen, L., Xu, F., and Li, Y. (2020). Understanding the urban pandemic spreading of covid-19 with real world mobility data. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3485–3492

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi:10.1038/s41586-020-2649-2

Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics (New York, NY: Springer New York). DOI: 10.1007/978-0-387-21606-5 ISSN: 0172-7397

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., and Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences* 19, 46–54

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation* 9, 1735–1780

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* 9, 90–95. doi:10.1109/MCSE.2007.55

Huntsinger, L. F. and Ward, K. (2015). Using Mobile Phone Location Data to Develop External Trip Models. *Transportation Research Record* 2499, 25–32. doi:10.3141/2499-04. Publisher: SAGE Publications Inc

Huo, J., Cox, C. D., Seaver, W. L., Robinson, R. B., and Jiang, Y. (2010). Application of two-directional time series models to replace missing data. *Journal of Environmental Engineering* 136, 435–443

Idé, T. and Kato, S. (2009). Travel-time prediction using gaussian process regression: A trajectory-based approach. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (SIAM), 1185–1196

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

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

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

Jiang, L., Chen, C.-X., and Chen, C. (2023). L2mm: learning to map matching with deep models for low-quality gps trajectory data. *ACM Transactions on Knowledge Discovery from Data* 17, 1–25

Kaiser, M. (2008). Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics* 10, 083042. doi:10.1088/1367-2630/10/8/083042

Kan, Z., Tang, L., Kwan, M.-P., Ren, C., Liu, D., and Li, Q. (2019). Traffic congestion analysis at the turn level using Taxis' GPS trajectory data. *Computers, Environment and Urban Systems* 74, 229–243. doi:10.1016/j.compenvurbsys.2018.11.007

Kapp, A., Hansmeyer, J., and Mihaljević, H. (2023). Generative Models for Synthetic Urban Mobility Data: A Systematic Literature Review. *ACM Comput. Surv.* 56, 93:1–93:37. doi:10.1145/3610224

Khurana, R. and Goel, M. (2020). Eyes on the Road: Detecting Phone Usage by Drivers Using On-Device Cameras. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA: Association for Computing Machinery), CHI '20, 1–11. doi:10.1145/3313831.3376822

Kim, D., Park, K., Park, Y., and Ahn, J.-H. (2019). Willingness to provide personal information: Perspective of privacy calculus in iot services. *Computers in Human Behavior* 92, 273–281

Kim, E.-K. and MacEachren, A. (2014). An index for characterizing spatial bursts of movements: A case study with geo-located twitter data. In *GIScience 2014 Workshop on Analysis of Movement Data* (Citeseer)

KimleyHorn (2021). *Framework for Evaluating Big Data in Regional Travel and Mobility Analyses*. Tech. rep.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation* 15, 9–34

Kitamura, R. and Kermanshah, M. (1983). Identifying time and history dependencies of activity choice. *Transportation Research Record* 944, 22–30

Kitamura, R. and Van Der Hoorn, T. (1987). Regularity and irreversibility of weekly travel behavior. *Transportation* 14, 227–251

Kohn, R. and Ansley, C. F. (1986). Estimation, prediction, and interpolation for arima models with missing data. *Journal of the American statistical Association* 81, 751–761

Kondor, D., Hashemian, B., de Montjoye, Y.-A., and Ratti, C. (2020). Towards Matching User Mobility Traces in Large-Scale Datasets. *IEEE Transactions on Big Data* 6, 714–726. doi:10.1109/TBDATA.2018.2871693

Kroesen, M. (2014). Modeling the behavioral determinants of travel behavior: An application of latent transition analysis. *Transportation Research Part A: Policy and Practice* 65, 56–67. doi:10.1016/j.tra.2014.04.010

Kulkarni, V. and Garbinato, B. (2017). Generating synthetic mobility traffic using RNNs. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery* (New York, NY, USA: Association for Computing Machinery), GeoAI '17, 1–4. doi:10.1145/3149808.3149809

Lan, Z., He, C., and Cai, M. (2020). Urban road traffic noise spatiotemporal distribution mapping using multisource data. *Transportation Research Part D: Transport and Environment* 82, 102323. doi:10.1016/j.trd.2020.102323. ADS Bibcode: 2020TRPD...8202323L

Lapidus, L. and Pinder, G. F. (1999). *Numerical solution of partial differential equations in science and engineering* (John Wiley & Sons)

Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (IEEE), vol. 1, 87–94

Le, T. V., Oentaryo, R., Liu, S., and Lau, H. C. (2016). Local gaussian processes for efficient fine-grained traffic speed prediction. *IEEE Transactions on Big Data* 3, 194–207

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539. Publisher: Nature Publishing Group

Lee, D.-H. (2013). Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*

Lee, K. and Sener, I. N. (2020). Emerging data for pedestrian and bicycle monitoring: Sources and applications. *Transportation Research Interdisciplinary Perspectives* 4, 100095. doi:10.1016/j.trip.2020.100095

Lee, M. and McNally, M. G. (2006). An empirical investigation on the dynamic processes of activity scheduling and trip chaining. *Transportation* 33, 553–565

Lee, M. S. and McNally, M. G. (2003). On the structure of weekly activity/travel patterns. *Transportation Research Part A: Policy and Practice* 37, 823–839

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

Li, M., Wang, K., and Nurul Habib, K. (2025). Investigating recall & proxy bias in household travel surveys under the core-satellite fusion paradigm. *Transportation* doi: 10.1007/s11116-025-10654-1

Li, Z., Ning, H., Jing, F., and 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. doi:10.1371/journal.pone.0294430. Publisher: Public Library of Science

Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., and Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications* 41, 6174–6189. doi:10.1016/j.eswa.2014.03.054

Liu, G. and Onnela, J.-P. (2021). Bidirectional imputation of spatial gps trajectories with missingness using sparse online gaussian process. *Journal of the American Medical Informatics Association* 28, 1777–1784

Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., and Oakden-Rayner, L. (2022a). The medical algorithmic audit. *The Lancet Digital Health* 4, e384–e397

Liu, Z., Liu, Z., and Fu, X. (2022b). Dynamic Origin-Destination Flow Prediction Using Spatial-Temporal Graph Convolution Network With Mobile Phone Data. *IEEE Intelligent Transportation Systems Magazine* 14, 147–161. doi:10.1109/MITS.2021.3082397. Conference Name: IEEE Intelligent Transportation Systems Magazine

Liu, Z., Lyu, C., Huo, J., Wang, S., and Chen, J. (2022c). Gaussian process regression for transportation system estimation and prediction problems: The deformation and a hat kernel. *IEEE Transactions on Intelligent Transportation Systems* 23, 22331–22342

Lu, K., Liu, J., Zhou, X., and Han, B. (2021). A Review of Big Data Applications in Urban Transit Systems. *IEEE Transactions on Intelligent Transportation Systems* 22, 2535–2552. doi:10.1109/TITS.2020.2973365. Conference Name: IEEE Transactions on Intelligent Transportation Systems

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

McCool, D., Lugtig, P., and Schouten, B. (2022). Maximum interpolable gap length in missing smartphone-based gps mobility data. *Transportation* , 1–31

McGuckin, N. and Murakami, E. (1999a). Examining trip-chaining behavior: Comparison of travel by men and women. *Transportation Research Record* 1693, 79–85. doi:10.3141/1693-12

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

Merikhipour, M., Khanmohammadidoustani, S., and Abbasi, M. (2024). Transportation mode detection through spatial attention-based transductive long short-term memory and off-policy feature selection. *Expert Systems With Applications* 267. doi:10.1016/j.eswa.2024.126196. [Online; accessed 2024-12-25]

Merrill, N. H., Atkinson, S. F., Mulvaney, K. K., Mazzotta, M. J., and Bousquin, J. (2020). Using data derived from cellular phone locations to estimate visitation to natural areas: An application to water recreation in new england, usa. *PloS one* 15, e0231863

Miah, S. J., Vu, H. Q., Gammack, J., and McGrath, M. (2017). A Big Data Analytics Method for Tourist Behaviour Analysis. *Information & Management* 54, 771–785. doi:10.1016/j.im.2016.11.011

Mohamed, A., Zhu, D., Vu, W., Elhoseiny, M., and Claudel, C. (2022). Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *European Conference on Computer Vision* (Springer), 463–479

Mokhtarian, P. L. and 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. doi:10.1016/j.tra.2003.12.004

Morris, D. H., Rossine, F. W., Plotkin, J. B., and Levin, S. A. (2021). Optimal, near-optimal, and robust epidemic control. *Communications Physics* 4, 78

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion* , 69–84

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence* 29. doi:10.1609/aaai.v29i1.9602. Number: 1

Nasernejad, P., Sayed, T., and Alsaleh, R. (2021). Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous gaussian process inverse reinforcement learning (gp-irl) approach. *Accident Analysis & Prevention* 161, 106355

Nevin, J. W., Vaquero-Caballero, F., Ives, D. J., and Savory, S. J. (2021). Physics-informed gaussian process regression for optical fiber communication systems. *Journal of Lightwave Technology* 39, 6833–6844

Nishii, K., Kondo, K., and Kitamura, R. (1988). An empirical analysis of trip chaining behavior

Ong, C., Williamson, R. C., and Smola, A. (2002). Hyperkernels. *Advances in neural information processing systems* 15

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

Pappalardo, L. and Simini, F. (2018). Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery* 32, 787–829. doi:10.1007/s10618-017-0548-4

Pappalardo, L., Simini, F., Barlacchi, G., and Pellungrini, R. (2019). scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data. *arXiv preprint arXiv:1907.07062*

Park, H., Oh, C., Moon, J., and Kim, S. (2018). Development of a lane change risk index using vehicle trajectory data. *Accident Analysis & Prevention* 110, 1–8. doi:10.1016/j.aap.2017.10.015

Pecheux, K. K., Pecheux, B. B., and Ledbetter, G. (2020). *Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making* (Washington, D.C.: Transportation Research Board). doi:10.17226/25965

Pendyala, R. M., Kitamura, R., Chen, C., and Pas, E. I. (1997). An activity-based microsimulation analysis of transportation control measures. *Transport Policy* 4, 183–192. doi:https://doi.org/10.1016/S0967-070X(97)00005-X

Pereira, F. C., Rodrigues, F., and Ben-Akiva, M. (2015). Using Data From the Web to Predict Public Transport Arrivals Under Special Events Scenarios. *Journal of Intelligent Transportation Systems* 19, 273–288. doi:10.1080/15472450.2013.868284. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15472450.2013.868284

Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., and Bento, C. (2017). Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science* 6, 11. doi:10.1140/epjds/s13688-017-0108-6

Qi, B., Kang, L., and Banerjee, S. (2017). A vehicle-based edge computing platform for transit and human mobility analytics. In *Proceedings of the Second ACM/IEEE Symposium on Edge Computing* (New York, NY, USA: Association for Computing Machinery), SEC '17, 1–14. doi:10.1145/3132211.3134446

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378, 686–707

Rasmussen, C. and Ghahramani, Z. (2000). Occam' s Razor. In *Advances in Neural Information Processing Systems* (MIT Press), vol. 13

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning (Cambridge, Mass: MIT Press). OCLC: ocm61285753

Razavi, R., Xue, G., and Akpan, I. J. (2024). Predicting Sociodemographic Attributes from Mobile Usage Patterns: Applications and Privacy Implications. *Big Data* 12, 213–228. doi:10.1089/big.2022.0182. Publisher: Mary Ann Liebert, Inc., publishers

Reif, J. and Schmücker, D. (2020). Exploring new ways of visitor tracking using big data sources: Opportunities and limits of passive mobile data for tourism. *Journal of Destination Marketing & Management* 18, 100481. doi:10.1016/j.jdmm.2020.100481

Ren, H., Ruan, S., Li, Y., Bao, J., Meng, C., Li, R., et al. (2021). Mtrajrec: Map-constrained trajectory recovery via seq2seq multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1410–1419

Richardson, A. J., Ampt, E. S., and Meyburg, A. H. (1996). Nonresponse issues in household travel surveys. In *Transportation Research Board* (National Research Council), vol. 10, 79–114

Richardson, H. (2021). Finding a Win-Win: Planning and Data-Sharing Partnerships between Governments and Public Land Management Agencies

Rittel, H. W. and Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences* 4, 155–169

Rodrigues, F., Henrickson, K., and Pereira, F. C. (2018). Multi-output gaussian processes for crowdsourced traffic data imputation. *IEEE Transactions on Intelligent Transportation Systems* 20, 594–603

Sakhare, R. S., Desai, J., Saldivar-Carranza, E. D., and Bullock, D. M. (2024). Methodology for Monitoring Border Crossing Delays with Connected Vehicle Data: United States and Mexico Land Crossings Case Study. *Future Transportation* 4, 107–129. doi:10.3390/futuretransp4010007. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute

Sarker, M. N. I., Peng, Y., Yiran, C., and Shouse, R. C. (2020). Disaster resilience through big data: Way to environmental sustainability. *International Journal of Disaster Risk Reduction* 51, 101769. doi:10.1016/j.ijdrr.2020.101769

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10, 20130246. doi:10.1098/rsif.2013.0246. Publisher: Royal Society

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press)

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* , 461–464

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Sheller, M. and Urry, J. (2006). The New Mobilities Paradigm. *Environment and Planning A: Economy and Space* 38, 207–226. doi:10.1068/a37268. Publisher: SAGE Publications Ltd

Shin, Y., Darbon, J., and Karniadakis, G. E. (2020). On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type pdes. *arXiv preprint arXiv:2004.01806*

Singh, G., Sivaraman, V., and Hard, E. (2022). *State of Emerging Mobility Big Data Sources and its Applications — Task 1: Evaluate Mobility Datasets.* Tech. rep., Texas A&M Transportation Institute

Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems* 18

Solomon, A., Bar, A., Yanai, C., Shapira, B., and Rokach, L. (2018). Predict Demographic Information Using Word2vec on Spatial Trajectories. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (New York, NY, USA: Association for Computing Machinery), UMAP '18, 331–339. doi:10.1145/3209219.3209224. [Online; accessed 2025-06-20]

Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature physics* 6, 818–823

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science* 327, 1018–1021

[Dataset] Spectus (2022). Sensitive Points of Interest Policy

Steentoft, A., Lee, B.-S., and Schläpfer, M. (2024). Quantifying the uncertainty of mobility flow predictions using gaussian processes. *Transportation* 51, 2301–2322

Stopher, P. R., Kockelman, K., Greaves, S. P., and Clifford, E. (2008). Reducing Burden and Sample Sizes in Multiday Household Travel Surveys. *Transportation Research Record* 2064, 12–18. doi:10.3141/2064-03. Publisher: SAGE Publications Inc

Storm, P. J., Mandjes, M., and van Arem, B. (2022). Efficient evaluation of stochastic traffic flow models using gaussian process approximation. *Transportation research part B: methodological* 164, 126–144

[Dataset] StreetLight (2025). Tulsa MPO Fills Traffic Count Gaps - AADT.pdf

Sulis, P., Manley, E., Zhong, C., and Batty, M. (2018). Using mobility data as proxy for measuring urban vitality. *Journal of Spatial Information Science* 2018, 137–162. doi: 10.5311/JOSIS.2018.16.384

Sun, H., Yang, C., Deng, L., Zhou, F., Huang, F., and Zheng, K. (2021). Periodicmove: Shift-aware human mobility recovery with graph neural network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1734–1743

Teixeira, D. d. C., Almeida, J. M., and Viana, A. C. (2021). On estimating the predictability of human mobility: the role of routine. *EPJ Data Science* 10, 49

Timmermans, H. and Arentze, T. A. (2011). Transport models and urban planning practice: Experiences with albatross. *Transport Reviews* 31, 199–207. doi:10.1080/01441647.2010.518292

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics* (PMLR), 567–574

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography* 46, 234–240

Tozluoğlu, Ç., Liao, Y., and Sprei, F. (2025). Mobile phone application data for activity plan generation. *Transportation* , 1–33

Ugurel, E. (2023). Time-varying transition matrices with multi-task gaussian processes. *arXiv preprint arXiv:2306.11772*

Ugurel, E., Guan, X., Wang, Y., Huang, S., Wang, Q. R., and Chen, C. (2024a). Correcting missingness in passively-generated mobile data with multi-task gaussian processes. *Transportation Research Part C* 161

Ugurel, E., Wu, X., Wang, R., Lee, B. H. Y., and Chen, C. (2024b). Metropolitan Planning Organizations' Uses of and Needs for Big Data. *Findings* doi:10.32866/001c.127143. Publisher: Findings Press

Uğurel, E., Huang, S., and Chen, C. (2024). Learning to generate synthetic human mobility data: A physics-regularized Gaussian process approach based on multiple kernel learning. *Transportation Research Part B: Methodological* 189, 103064. doi:10.1016/j.trb.2024.103064

Van Rooyen, B. and Williamson, R. C. (2018). A theory of learning with corrupted labels. *Journal of Machine Learning Research* 18, 1–50

van Rossum, G. (1995). *Python tutorial*. Tech. Rep. CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam

Venkatesh, Morris, Davis, and Davis (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 425. doi:10.2307/30036540. Publisher: JSTOR

Wallace, B., Barnes, J., and Rutherford, G. S. (2000). Evaluating the effects of traveler and trip characteristics on trip chaining, with implications for transportation demand management strategies. *Transportation Research Record* 1718, 97–106. doi:10.3141/1718-13

Wang, F. and Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies* 87, 58–74

Wang, F., Wang, J., Zhang, Y., Chen, C., and Ban, X. J. (2021a). *Travelers' Adaptive Behaviors in Response to Seattle's Alaskan Way Viaduct Replacement*. Tech. rep.

Wang, F., Wang, X., and Ban, X. J. (2024). Data poisoning attacks in intelligent transportation systems: A survey. *Transportation Research Part C: Emerging Technologies* 165, 104750. doi:10.1016/j.trc.2024.104750

Wang, S., Wang, H., and Perdikaris, P. (2021b). On the eigenvector bias of fourier feature networks: From regression to solving multi-scale pdes with physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering* 384, 113938

Wang, S., Yu, X., and Perdikaris, P. (2022a). When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics* 449, 110768

Wang, Y., Guan, X., Ugurel, E., Chen, C., Huang, S., and 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. doi:10.1016/j.jtrangeo.2024.104108

Wang, Z., Xing, W., Kirby, R., and Zhe, S. (2022b). Physics informed deep kernel learning. In *International Conference on Artificial Intelligence and Statistics* (PMLR), 1206–1218

Wang, Z., Zhang, S., and James, J. (2020). Reconstruction of missing trajectory data: a deep learning approach. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)* (IEEE), 1–6

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, eds. Stéfan van der Walt and Jarrod Millman. 56 – 61. doi:10.25080/Majora-92bf1922-00a

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., and Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface* 10, 20120986. doi:10.1098/rsif.2012.0986

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., and Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with gis data. *PloS one* 10, e0133630

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning* (PMLR), 1067–1075

Wilson, A. G. (2014). *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. Ph.D. thesis, University of Cambridge Cambridge, UK

Wu, F., Cheng, Z., Chen, H., Qiu, Z., and Sun, L. (2024a). Traffic state estimation from vehicle trajectories with anisotropic gaussian processes. *Transportation Research Part C: Emerging Technologies* 163, 104646

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., et al. (2019). Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems* 77, 101368. doi:10.1016/j.compenvurbsys.2019.101368

Wu, X., He, H., Wang, Y., and Wang, Q. (2024b). Pretrained mobility transformer: A foundation model for human mobility. *arXiv preprint arXiv:2406.02578*

Wu, X., Wang, Y., Ugurel, E., Chen, C., Huang, S., and Wang, Q. R. (2024c). Location-based service (lbs) data quality metrics and effects on mobility inference. *arXiv preprint arXiv:2411.16595*

Xie, K., Ozbay, K., Kurkcu, A., and Yang, H. (2017). Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. *Risk Analysis* 37, 1459–1476. doi:10.1111/risa.12785. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12785

Xie, Y., Zhao, K., Sun, Y., and Chen, D. (2010). Gaussian processes for short-term traffic volume forecasting. *Transportation Research Record* 2165, 69–78

Xu, F., Lin, Z., Xia, T., Guo, D., and 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. doi:10.1145/3411807

Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C., and Zhang, L. (2022). A data-driven travel mode share estimation framework based on mobile device location data. *Transportation* 49, 1339–1383. doi:10.1007/s11116-021-10214-3

Yang, T., Xu, X., Guo, Q., Zhang, L., and Sun, H. (2017a). Ev charging behaviour analysis and modelling based on mobile crowdsensing data. *IET Generation, Transmission & Distribution* 11, 1683–1691. doi:10.1049/iet-gtd.2016.1200. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-gtd.2016.1200

Yang, T., Xu, X., Guo, Q., Zhang, L., and Sun, H. (2017b). EV charging behaviour analysis and modelling based on mobile crowdsensing data. *IET Generation, Transmission & Distribution* 11, 1683–1691. doi:10.1049/iet-gtd.2016.1200. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/iet-gtd.2016.1200

Yang, X., Barajas-Solano, D., Tartakovsky, G., and Tartakovsky, A. M. (2019). Physics-informed cokriging: A gaussian-process-regression-based multifidelity method for data-model convergence. *Journal of Computational Physics* 395, 410–431

Yin, J., Dong, J., Hamm, N. A. S., Li, Z., Wang, J., Xing, H., et al. (2021). Integrating remote sensing and geospatial big data for urban land use mapping: A review. *International Journal of Applied Earth Observation and Geoinformation* 103, 102514. doi:10.1016/j.jag.2021.102514

Yu, Q., Zhang, H., Li, W., Sui, Y., Song, X., Yang, D., et al. (2020). Mobile phone data in urban bicycle-sharing: Market-oriented sub-area division and spatial analysis on emission reduction potentials. *Journal of Cleaner Production* 254, 119974. doi:10.1016/j.jclepro.2020.119974

Yuan, Y. and Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In *Geographic Information Science: 7th International Conference, GIScience 2012, Columbus, OH, USA, September 18-21, 2012. Proceedings 7* (Springer), 354–367

Yuan, Y., Zhang, Z., Yang, X. T., and Zhe, S. (2021). Macroscopic traffic flow modeling with physics regularized gaussian process: A new insight into machine learning applications in transportation. *Transportation Research Part B: Methodological* 146, 88–110

Zeidan, A., Lagerspetz, E., Zhao, K., Nurmi, P., Tarkoma, S., and Vo, H. T. (2020). GeoMatch: Efficient Large-scale Map Matching on Apache Spark. *ACM/IMS Transactions on Data Science* 1, 1–30. doi:10.1145/3402904

Zhang, B., Rasouli, S., and Feng, T. (2024a). Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach. *Travel Behaviour and Society* 37, 100843. doi:10.1016/j.tbs.2024.100843

Zhang, K., Pang, Y., Zhang, Y., and Sekimoto, Y. (2024b). MobGLM: A Large Language Model for Synthetic Human Mobility Generation. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems* (New York, NY, USA: Association for Computing Machinery), SIGSPATIAL '24, 629–632. doi:10.1145/3678717.3691311

Zhang, P., Zhou, J., and Zhang, T. (2017). Quantifying and visualizing jobs-housing balance with big data: A case study of Shanghai. *Cities* 66, 10–22. doi:10.1016/j.cities.2017.03.004

Zhao, Y., Pawlak, J., and Sivakumar, A. (2022). Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach. *Transportation Research Part B: Methodological* 155, 101–134. doi:10.1016/j.trb.2021.11.004

Zheng, Y., Li, Q., Chen, Y., Xie, X., and Ma, W.-Y. (2008). Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*. 312–321

Zheng, Y., Xie, X., Ma, W.-Y., et al. (2010). Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* 33, 32–39

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*. 791–800

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and 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* (New York, NY, USA: Association for Computing Machinery), WSDM '15, 295–304. doi:10.1145/2684822.2685287. [Online; accessed 2025-06-20]

Zhou, H., Wang, H., Zhou, Y., Luo, X., Tang, Y., Xue, L., et al. (2020). Demystifying diehard android apps. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 187–198

Zhou, X., Yeh, A. G. O., and Yue, Y. (2018). Spatial variation of self-containment and jobs-housing balance in Shenzhen using cellphone big data. *Journal of Transport Geography* 68, 102–108. doi:10.1016/j.jtrangeo.2017.12.006

Zhu, Y., Ye, Y., Wu, Y., Zhao, X., and Yu, J. (2023). SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. *Advances in Neural Information Processing Systems* 36, 22961–22977

Zhu, Z., Xu, M., Di, Y., and Yang, H. (2022). Fitting spatial-temporal data via a physics regularized multi-output grid gaussian process: case studies of a bike-sharing system. *IEEE Transactions on Intelligent Transportation Systems* 23, 21090–21101

Álvarez, M., Luengo, D., and Lawrence, N. D. (2009). Latent force models. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, eds. D. van Dyk and M. Welling (Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR), vol. 5 of *Proceedings of Machine Learning Research*, 9–16

Appendices

Appendix 1

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 Initialization Strategy & Parameter Optimization

The optimization problem of a GP is non-convex. Therefore, kernel parameter initialization can help avoid bad local optima, leading to better model estimation and more accurate prediction results. We achieve this in two ways. First, if the gap is short enough, we take advantage of training points near the gap to infer trip characteristics. These include metrics like average velocity, total distance traveled, trip duration, and stop rate, among others. Our first set of experiments suggest that trips with differing characteristics tend to land at different optimal parameters. Therefore, we suggest initializing the lengthscale parameter of continuous temporal dimensions (Unix time in our case) according to the characteristics of the trip with missing data. If using a monotonically-increasing temporal dimension as an input, one alternative we have identified is to initialize the lengthscale parameter with the average length of gap between observations in the training data. Ak et al. (2018) employed this strategy without any optimization. We however do further optimize the parameter beyond this initial heuristic using the Adaptive Moment Estimation algorithm (Kingma and Ba, 2014), a preferred choice among various machine learning frameworks due to its computational efficiency and ability to manage sparse gradients on noisy data. We initialize binary variables (i.e., one-of-k encoded categorical variables) with a lengthscale of 1. This is largely done to avoid model misspecification during the optimization stage as lengthscale parameters are constrained to be nonnegative. In our implementation, a small amount of noise (called jitter) is added to the diagonal of the covariance matrix. This is done to ensure numerical stability when performing matrix operations like inversion and decomposition. Simultaneously, however, jitter can lead to parameter values dipping below 0 at the very first iteration of the optimization algorithm[1].

[1]If the learning rate is small, the first iteration of the optimization algorithm may result in a value that is very close to 0, and the associated jitter may cause the parameter to flip signs.

A.2 Algorithm to Determine Training Data for Experiments in 2.6.2

To choose the training dataset in 2.6.2, we used Algorithm 2, which simulates gaps in trajectory data based on the current temporal occupancy with respect to the bin length being tested. The following is a step-by-step breakdown of that algorithm:

1. **Binning the Trajectory Data**: The algorithm starts by creating an array of integers called bins that span from $\mathbf{t}(u, 1)$ to $\mathbf{t}(u, N)$ with a specified step size τ. This step is effectively dividing the entire temporal range of the trajectory data into bins of equal length.

2. **Mapping Data Points to Bins**: The *bins_dict* is then created using dictionary comprehension to map each data point to the relevant bin. This means that each data point in the trajectory is associated with a specific bin based on its timestamp.

3. **Selecting Non-Empty Bins**: Another dictionary, *non_empty_bins_dict*, is created to store only the non-empty bins. This step involves filtering out bins that have data points associated with them.

4. **Setting Target Occupancy**: The algorithm then sets a target occupancy level ($target\_ocp$) by randomly selecting a value between 0 and the current temporal occupancy (q_curr). The selected value is used to determine the desired temporal occupancy.

5. **Gap Simulation Loop**: The algorithm enters a loop to simulate gaps in the trajectory data until the temporal occupancy falls below the target level ($target\_ocp$). Within this loop:

   - It randomly chooses a bin from the non-empty bins using *np.random.choice*.
   - All data points associated with the chosen bin are removed from the original data, and the *bins_dict* and *non_empty_bins_dict* are updated accordingly.

6. **Calculating New Temporal Occupancy**: After the loop, the algorithm calculates the new temporal occupancy (new_ocp) with the gapped data.

7. **Return Output**: Finally, the algorithm returns the gapped trajectory data, as well as the new temporal occupancy (q_τ).

A.3 Detailed Imputation Accuracy Metrics for Robustness Experiments

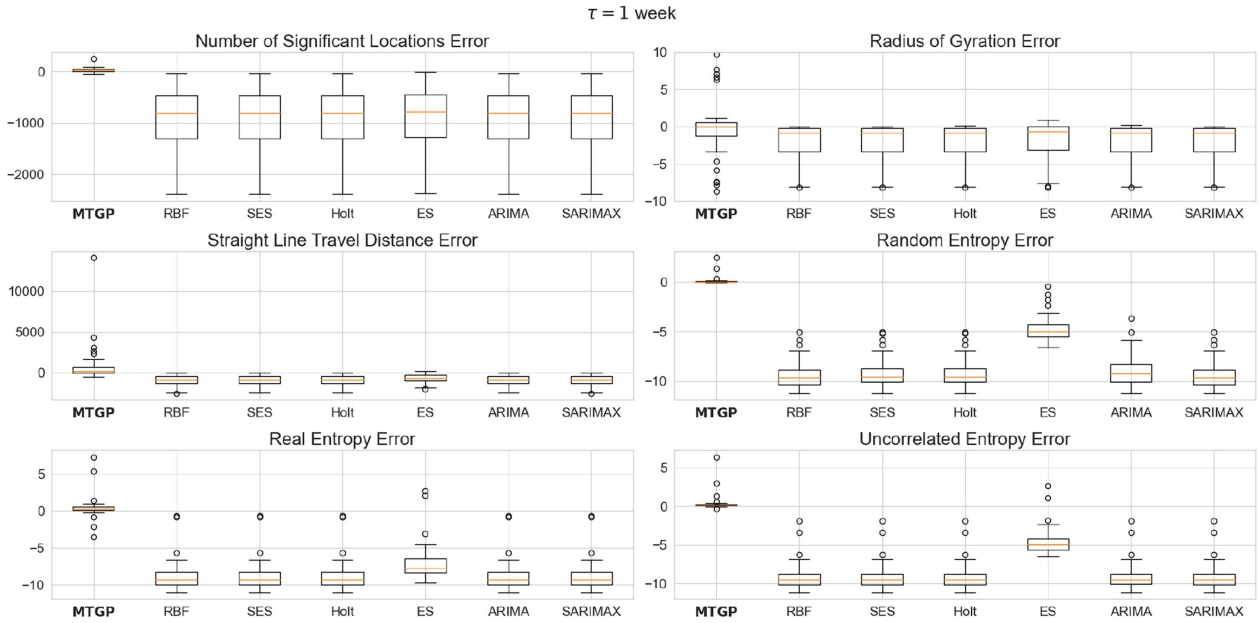


Figure A.1: Mobility error metrics for $\tau = 1$ week

A.4 Benchmark Methods and Model Parameters

A.5 Max Speed Threshold Sensitivity Analysis

To determine the optimal choice for a max speed threshold during the data preprocessing stage, we created a kernel density estimation (KDE) plot to visualize the number of observations eliminated by different speed limits. The KDE plot shows that the 200 km/h speed limit captures all of the points that the 300 km/h speed limit captures and some more. Because the speeds we calculate are based on "as the crow flies" distances, it is better to be more conservative in this limit—therefore, the additional points captured by the 200 km/h limit are most likely erroneous and have actual speeds much higher than 200 km/h (which is already high). This suggests that the 200 km/h speed limit is a better limit for eliminating outliers and reducing noise in the dataset. The 100 km/h speed limit, on the other hand, eliminates too many observations and reduces the data quality. Therefore, we chose the 200 km/h speed limit as the optimal threshold for our model (Figure A.15).

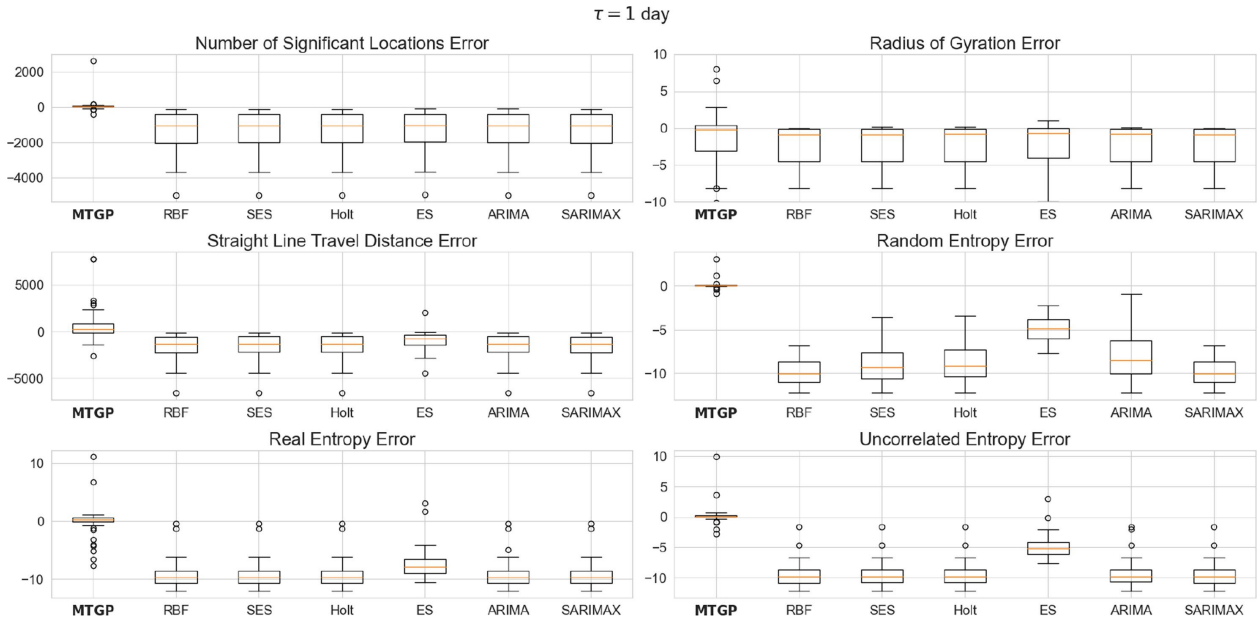


Figure A.2: Mobility error metrics for $\tau = 1$ day

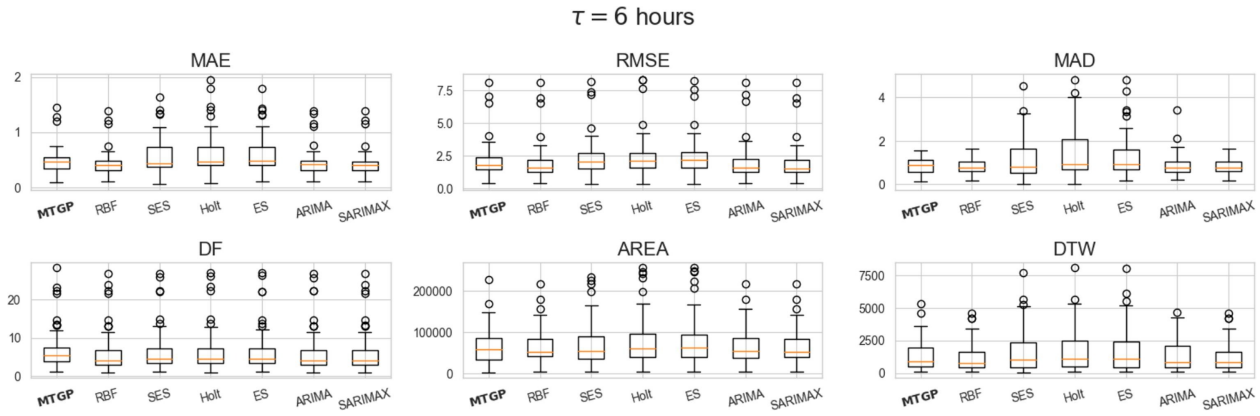


Figure A.3: Mobility error metrics for $\tau = 6$ hours

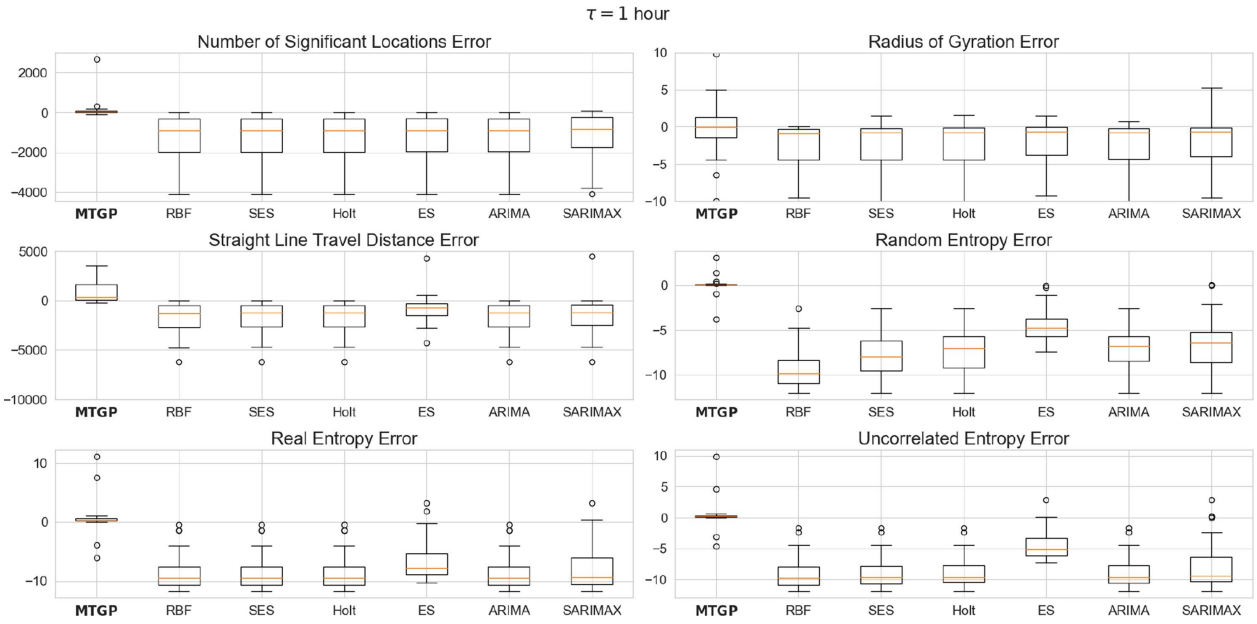


Figure A.4: Mobility error metrics for $\tau = 1$ hour

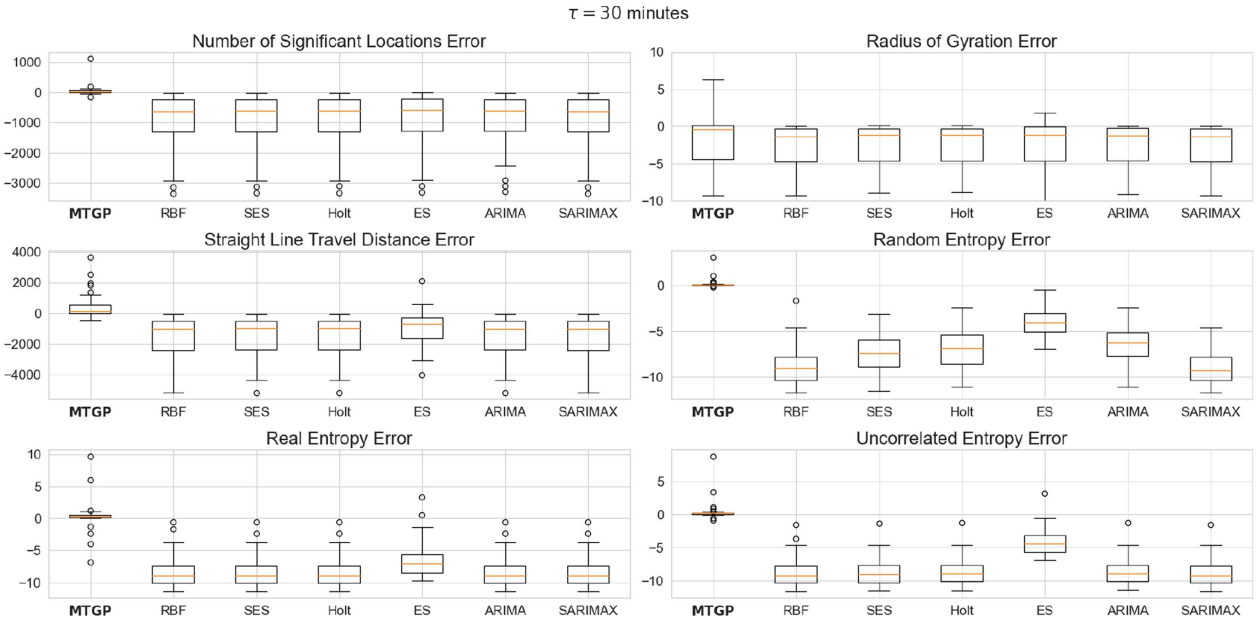


Figure A.5: Mobility error metrics for $\tau = 30$ minutes

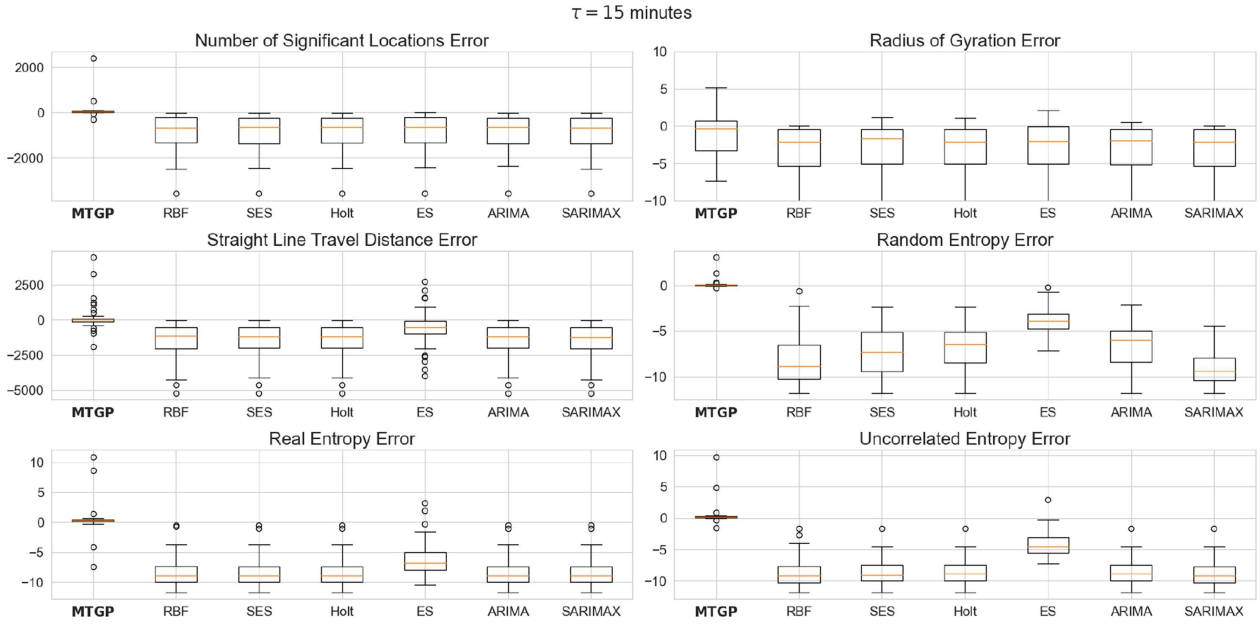


Figure A.6: Mobility error metrics for $\tau = 15$ minutes

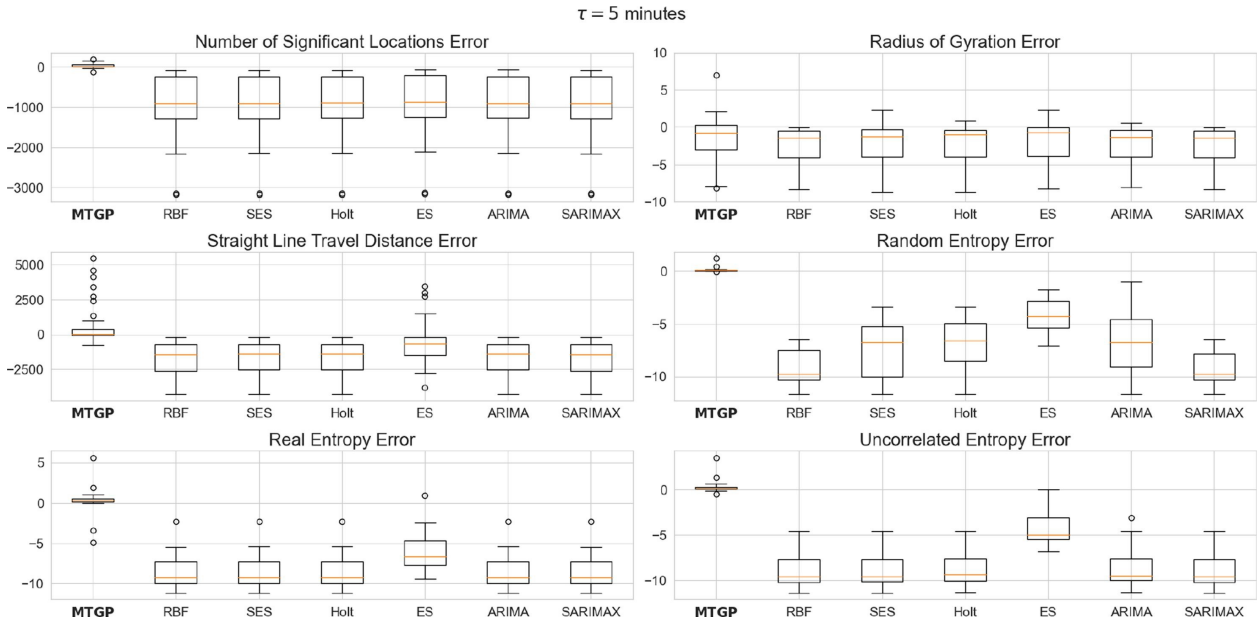


Figure A.7: Mobility error metrics for $\tau = 5$ minutes

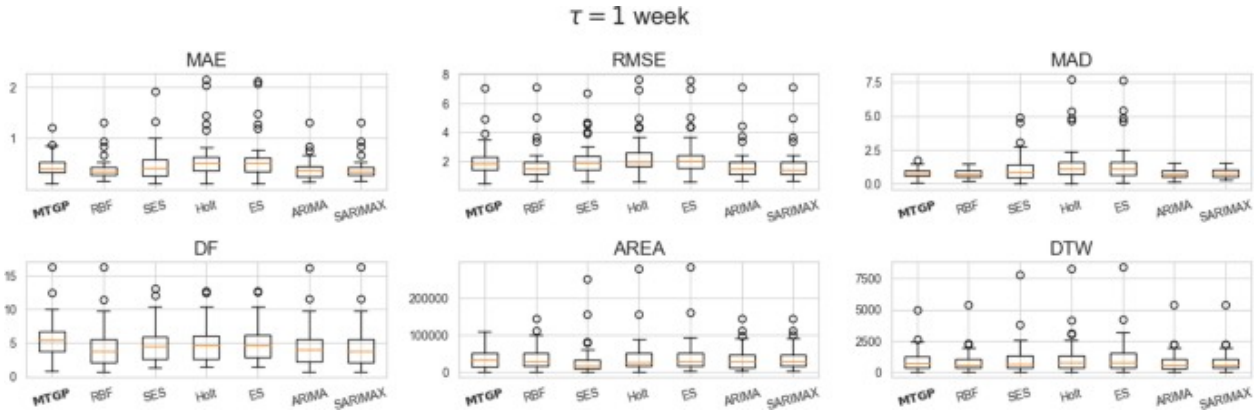


Figure A.8: Classical error metrics for $\tau = 1$ week

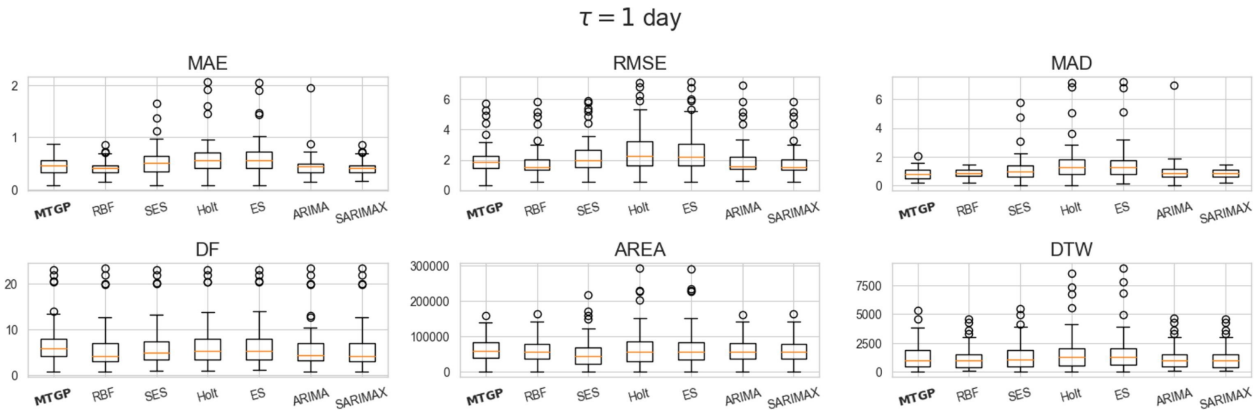


Figure A.9: Classical error metrics for $\tau = 1$ day

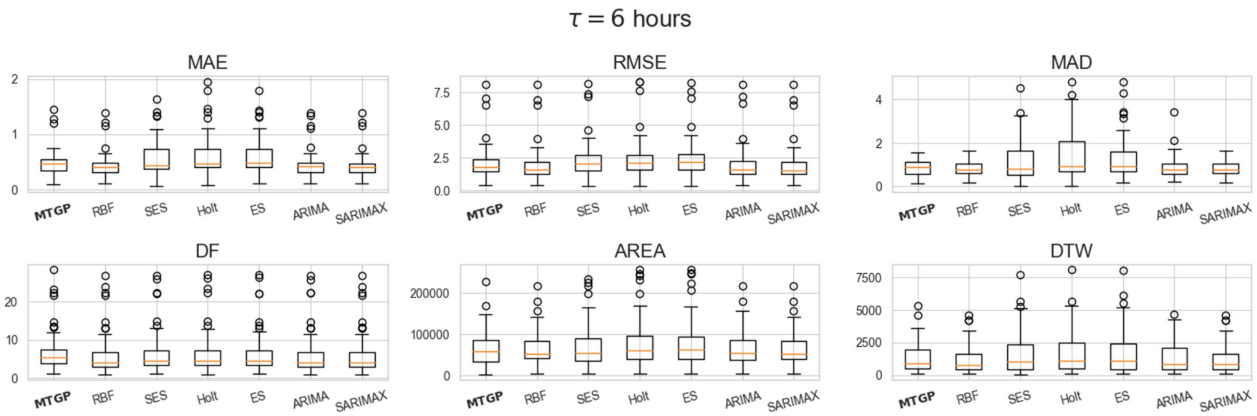


Figure A.10: Classical error metrics for $\tau = 6$ hours

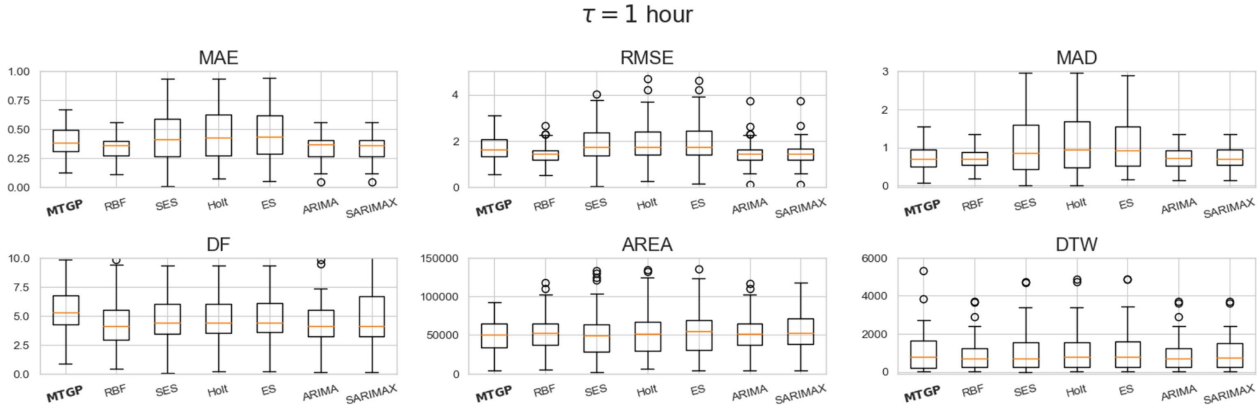


Figure A.11: Classical error metrics for $\tau = 1$ hour

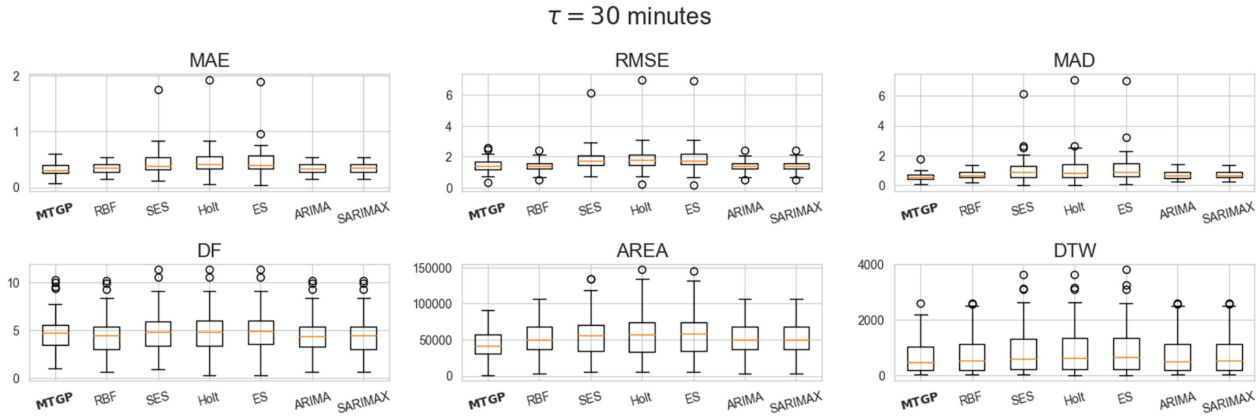


Figure A.12: Classical error metrics for $\tau = 30$ minutes

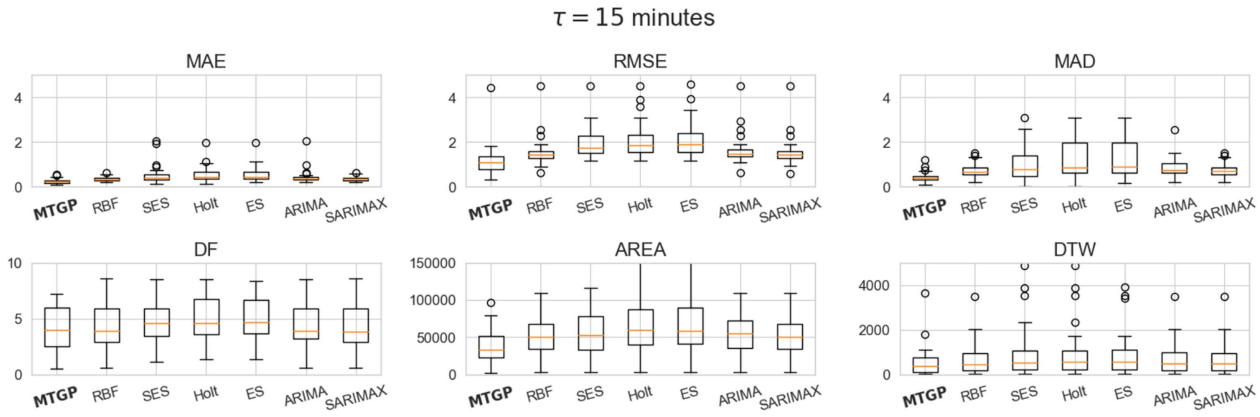


Figure A.13: Classical error metrics for $\tau = 15$ minutes

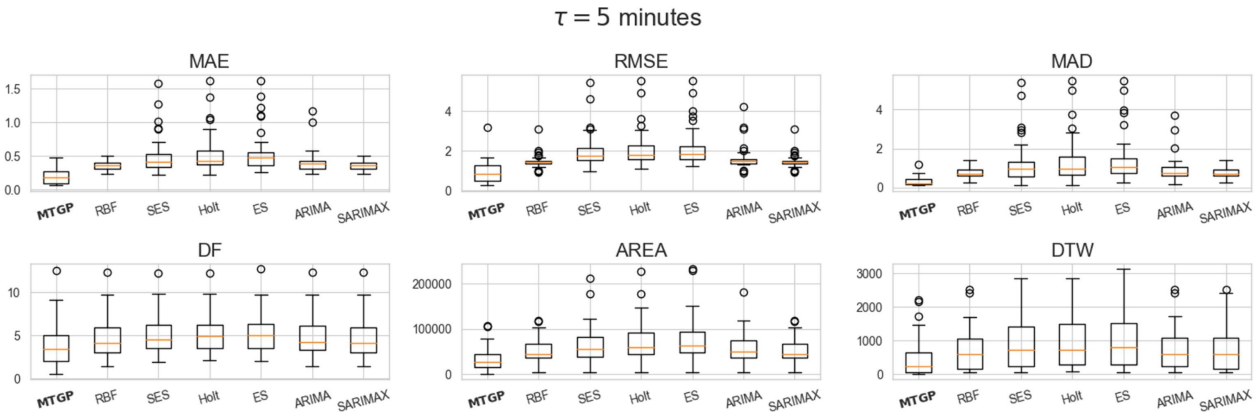


Figure A.14: Classical error metrics for $\tau = 5$ minutes

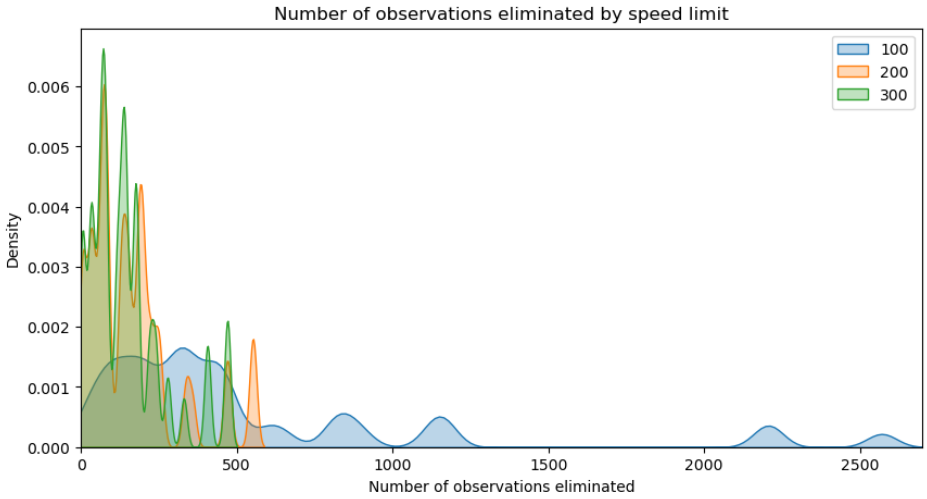


Figure A.15: The number of observations eliminated by varying speed limits across the users' data analyzed in Section 2.6

Table A.1: Parameters of benchmark methods.

Method	Parameter Count	Parameters	Notes
SES	1	α	Controls the weight given to the most recent observation when estimating the baseline of the time series.
Holt–Winters	3	α	Determines the weight given to the recent change in the level of the data.
		β_s	Determines the weight given to the recent change in the level of the data.
		β_d	Determines the dampening of the trend.
ES	4	α	Same as above.
		β	Determines the weight given to the recent change in the level of the data.
		γ	Controls the weight assigned to the seasonal component.
		m	Represents the length of the seasonal cycle (number of seasonal periods).
ARIMA	5–10	p	Number of lagged observations used to model the current value.
		d	Number of differencing operations to make the series stationary.
		q	Number of lagged forecast errors used to model the time series.
SARIMAX	8–20	p, d, q	Same as above.
		P	Number of seasonal autoregressive terms.
		D	Number of seasonal differencing operations.
		Q	Number of seasonal moving average terms.
		s	Determines periodicity (number of periods in season).
MTGP–RBF	16	ℓ	Controls extrapolation distance; one lengthscale per input dimension (14 total).
		η	Weight of a kernel component (output variance).
		ε	Homoscedastic Gaussian noise.
MTGP (ours)	21	ℓ, η, ε	Same as above (two per dimension).
		p	Determines the distance between repetitions of the function (two total).
		α	Controls the weighting of large-scale and small-scale variations (two total).

Table A.2: Parameters used during robustness experiments.

Method	**Parameter**	**Value**
Preprocessing	Max Speed (km/h)	200
	Spatial Radius for Compression (km)	0.3
	Uncertainty Filter (m)	100
Adam	Learning Rate for Adam	0.3
	Max Number of Training Iterations	150
Periodic Kernels	Initial Period Length for $K_{\text{PER},1}$	1440
	Initial Period Length for $K_{\text{PER},2}$	10080
Rational Quadratic Kernels	Initial Lengthscale for t_u	$\frac{1}{2N}\sum_{i=1}^{N} t_{u,i} - t_{u,i-1}$
	Initial Lengthscale for $[t_{s_1}, \ldots, t_{s_m}]$	1
Search Range for ARIMA	Order of the AR, p	[0, 3]
	Order of the differencing, d	[0, 2]
	Order of the MA, q	[0, 3]
Search Range for SARIMAX	Order of the AR, p	[0, 3]
	Order of the differencing, d	[0, 2]
	Order of the MA, q	[0, 3]
	Order of the seasonal AR, P	[0, 3]
	Order of the seasonal differencing, D	[0, 2]
	Order of the seasonal MA, Q	[0, 3]
	Order of the periodicity, s	24

Appendix 2

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

B.1 List of Notation Used

Table B.1 lists every variable and its definition used in Chapter 3.

B.2 Dataset and Software Implementation

For Sections 3.4.1 through 3.4.3, we employ privacy-protected, passively-generated GPS data from Spectus, a U.S.-based data solution provider specializing in geospatial analytics. The dataset contains discrete GPS points of 2,000 anonymous, opted-in individuals in the Greater Seattle Area between December 2019 and July 2020. The location data recorded includes timestamps, unique device identifiers, latitudes and longitudes, and a measure of data precision (i.e., a spatial radius for which the provider has 95% confidence in the reported coordinates). For the experiment in Section 3.4.3, we sequentially increased the number of training points in each split while keeping the number of testing points constant, as shown in Figure B.1.

For the experiment in Section 3.4.4, we use Microsoft Asia's GeoLife dataset, a comprehensive collection of location-based data gathered from GPS-enabled devices. This dataset spans from April 2007 to April 2012, amassed through the voluntary participation of users who carried GPS-enabled devices, such as smartphones or dedicated GPS trackers. These devices continuously recorded the users' location coordinates, capturing their movements and activities throughout the day. Besides being publically available, a huge benefit of this data source is that data points are more evenly distributed in time (as GPS-deviced were pinged regularly) giving a more accurate estimation of individuals' route choices. Our experiments specifically used the subset of all GeoLife users who had mode labels available. We discarded training sets with less than 100 points (due to numerical instabilities in matrix operations) and more than $2,000$ points (due to time constraints). We only considered trips made via buses, bikes, and by walking, as the other modes had too few trips to attain meaningful results.

Table B.1: Variable dictionary

	General Notation
λ	Longitude
ϕ	Latitude
v	Velocity
β	Bearing
Subscripts i and g	Indices for observations
Subscript j	Index for task
Subscript u	Index for input dimension
	GP Notation
$\mathbf{Y}$	Matrix of outputs
$\mathbf{T}$	Matrix of temporal variables
$\mathbf{P}$	Matrix of physical variables
$\mathbf{X}$	Matrix of inputs including $\mathbf{T}$ and $\mathbf{P}$
ϵ	Gaussian white noise
d	Number of input dimensions
n	Number of training data points
m	Number of inferred data points
c	Number of inducing points
k	Covariance function of a GP
$\mathbf{K}$	Covariance matrix
$\mathbf{K}^f$	Inter-task covariance matrix
$\mathbf{k}$	Vector of covariances
$\otimes$	Kronecker product
$\boldsymbol{\mu_*}$	Posterior mean function of the multi-task GP
$\boldsymbol{\sigma^2_*}$	Posterior variance of the multi-task GP
$\boldsymbol{\mu_{**}}$	Posterior mean function of the physics-regularized multi-task GP
$\boldsymbol{\sigma^2_{**}}$	Posterior variance of the physics-regularized multi-task GP
Θ	Set of GP parameters
l_k	Lengthscale parameter of kernel k
α	Scale mixture parameter of an RQ kernel
p	Period length parameter for a PER kernel
	MKL Notation
B	Set of base kernels
M	Number of MKL branches
A	Set of algebraic operations
η_h	Weight of kernel component h

bold denotes vectors.
Capital Bold denotes matrices.

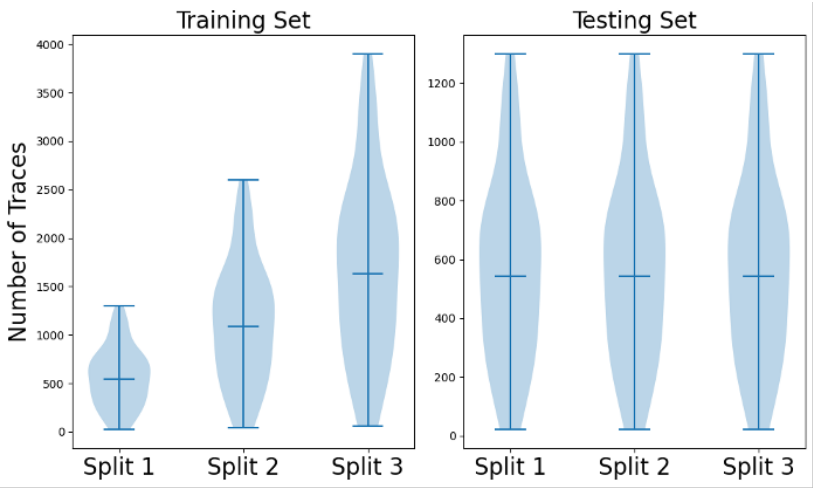
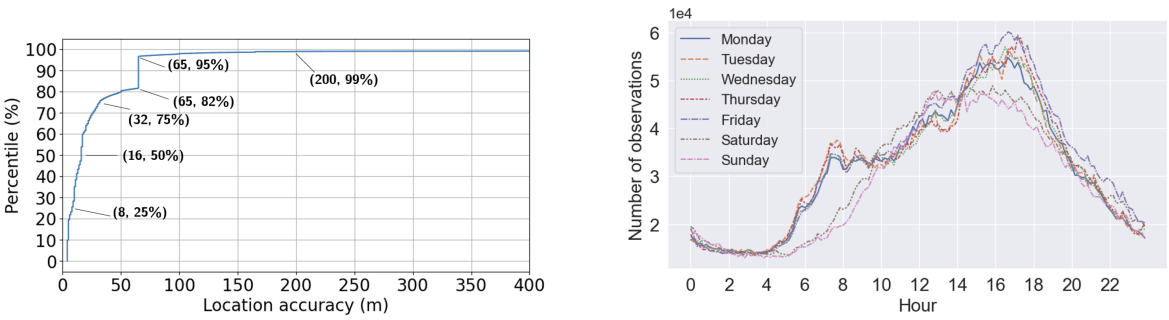


Figure B.1: Violin plots of training and testing set sizes across our passively-collected mobile data experiments

We implemented our methodology in the PYTHON programming language (van Rossum, 1995). We used GPYTORCH (Gardner et al., 2018) to increase efficiency and speed up matrix inversions within our GP framework. Furthermore, we used PANDAS (Wes McKinney, 2010) to read and wrangle the mobile dataset from which we derived the GP models, NUMPY (Harris et al., 2020) for a variety of array operations, and SCIKIT-MOBILITY (Pappalardo et al., 2019) to conduct data cleaning and preprocessing. Last but not least, we used MATPLOTLIB (Hunter, 2007) to produce all visualizations of our results.

B.3 Descriptive Statistics

Most data points exhibit a high level of precision, with approximately 95% of all traces located within a 65-meter radius of their true position (Figure B.2a). This represents a significant improvement compared to a previous study conducted in 2018 using a similar dataset, where only around 7% of observations had a precision worse than 1,000 meters (Ban et al., 2018). Regarding the temporal distribution of observations throughout the day (Figure B.2b), our findings align with that of Ban et al. (2018). We observe two distinct peaks on weekdays: one in the morning, around 7 am, and another in the evening, around 6 pm. On weekends, a single mid-day peak is evident. Moreover, the number of observations exhibits a positive trend as the hours progress, which holds true for every day.



(a) (b)

Figure B.2: Descriptive analysis of Spectus data for all 2,000 users. (a) Cumulative distribution of location accuracy. (b) Distribution of observations within each day of the week.

We find trivial differences between the entire dataset and the 33 users sampled for the experiments in Section 3.4.3, confirmed by looking at various metrics. Figure B.3 highlights the distribution of location accuracies between the two sets, showing that the experiments sample does not contain any outliers. Figure B.4 shows the distribution of observation counts by time of day (normalized by the total count of observations) between the two datasets, where we observe slight differences in the weekday distributions. Finally, Figure B.5 shows the distribution of sampling rates over time in the dataset. We observe slightly higher intervals between consecutive observations in certain weeks, but overall the two distribution look fairly similar.

B.4 Data Preprocessing and Modeling Considerations

We preprocess raw GPS traces to remove noisy data in two ways. First, we filter out rows using an upper bound on velocity (200 km/h, as the crow flies), which removes oscillations and physically-impossible jumps that may be present in raw mobile data. We further remove observations with a precision radius greater than 300 meters. The potential of these noisy points to provide valuable location information is shadowed by the problems they cause for model calibration, which disrupts the continuity of smooth trajectories.

We also reduce the size of trajectories through a compression algorithm (Algorithm 4) that aims to generate an approximated trajectory that largely retains the shape of the original trajectory. If two points are within a certain Haversine distance (we use 100 meters), we

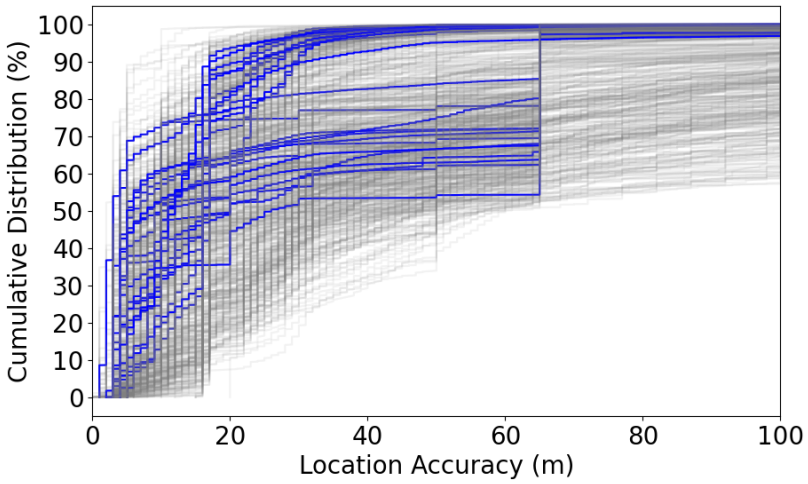


Figure B.3: Cumulative distribution of location accuracy for Spectus users in the Experiments sample (blue) and the whole dataset (grey)

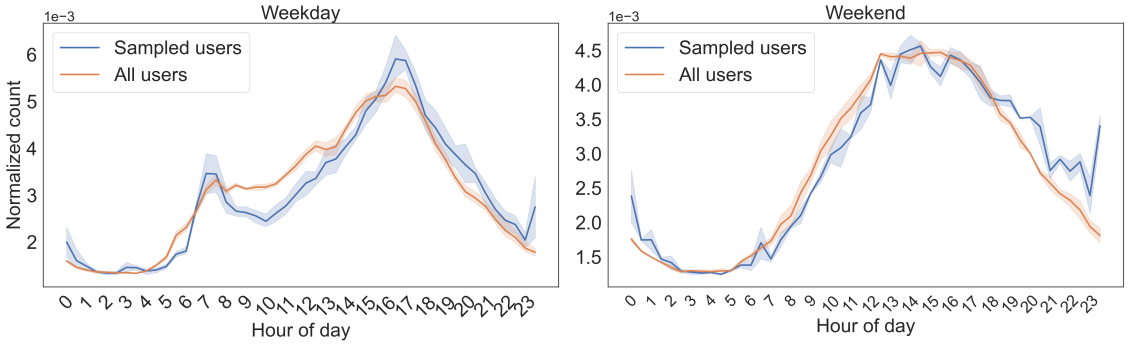


Figure B.4: Normalized observation counts by time of day for Spectus users in the Experiments sample (blue) and the whole dataset (orange)

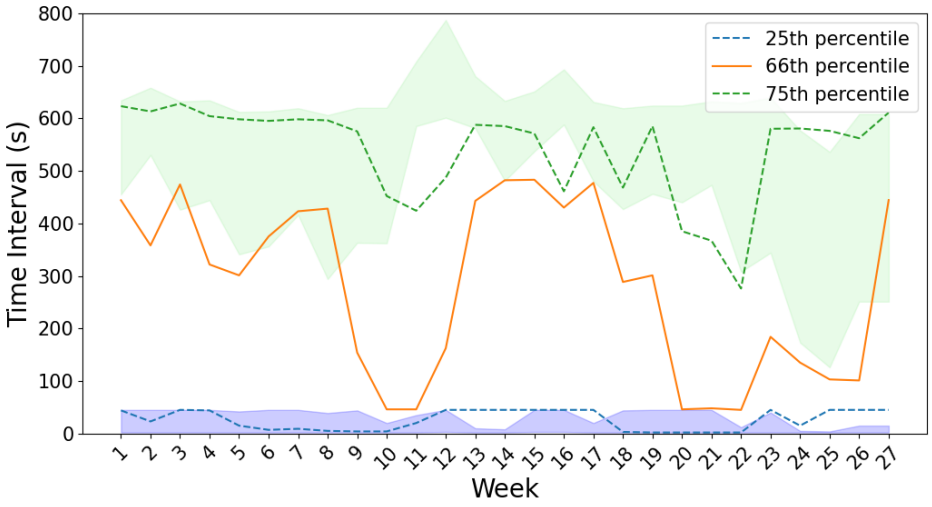


Figure B.5: Evolution of time interval between two consecutive points Week 1 to 27. Lines show the percentiles for the experiment sample, while the green and blue fillings shows the 70-80 and 15-35 percentile range for the entire dataset, respectively.

replace the two entries with their median—this is a variation of the Douglas-Peucker algorithm (Ugurel et al., 2024a). We do this to ease the computational burdens of training our GP framework, which can be burdensome for large trajectory sets. Because many GPS points are close in time and space, this allows us to significantly reduce the size of our training sample without compromising accuracy.

A practical consideration with using a greedy strategy to learn the form of the best-fitting kernel is that of computational complexity. For this, we suggest leveraging "the median trick" (Schölkopf and Smola, 2002)—instead of minimizing the negative marginal log likelihood to learn the parameters of the kernel, they can be manually set using apriori knowledge of the domain and the data. For example, the lengthscale parameter can be set as the median gap between observations in the input space. This idea gets tricky with parameters that do not relate to the smoothness of the function (i.e., the outputscale or the period length), but these too can be experimented with in a systematic way to reduce the computational burden.

B.5 Predictor Variables

Table B.2 shows the temporal and physical variables we use as predictors in our experiments. As discussed in Section 3.3.4, we leverage various calendar-based structures to encode more

Table B.2: Temporal and physical dimensions used in our experiments. Asterisk* denotes a dimension that was not employed in the GeoLife experiments.

Variable	Notation	Type	Model Inputs
Unix time (normalized)	$\mathbf{t}_u$	Continuous	$[0, 1, \ldots, n\ ]$
Seconds (after midnight) - Sine	$\mathbf{t}_{ss}$	Continuous	$[0, \ldots, 1]$
Seconds (after midnight) - Cosine	$\mathbf{t}_{sc}$	Continuous	$[0, \ldots, 1]$
Day of week	$\mathbf{t}_d$	Categorical (Binary)	[0, 1]
AM peak*	$\mathbf{t}_{am}$	Binary	[0, 1]
PM peak*	$\mathbf{t}_{pm}$	Binary	[0, 1]
Average segment velocity	$\mathbf{v}$	Continuous	$[0, \infty]$
Bearing	$\boldsymbol{\beta}$	Continuous	$[0, 360]$

regularity into the temporal GP. We normalize Unix time by equating a user's first observation to 0 and increasing it linearly thereafter. The second of the day is converted to cyclical encoding by calculating the sine and cosine values, wrapping smoothly around the boundaries (e.g., 23:59:59 and 00:00:00). Monday is the first day of the week and it is denoted with a 0. The AM peak variable is active in any observation between 6:00 AM and 10:00 AM, while the PM peak is between 3:00 PM and 7:00 PM (both in local time). The average segment velocity and bearing are calculated according to Equations 3.7 and 3.8, respectively.

B.6 Algorithms

Algorithms 3-6 are used inside Algorithm 7 (which is for the GeoLife experiments), while Algorithm 1 is also employed in the rest of the experiments.

Algorithm 3: Filter Trips

Input: Points
Output: Filtered points
Function *FilterTrips(points)*:
    Filter points outside Beijing based on longitude and latitude constraints;
    Remove trips going outside the boundary;
    Filter points if segment velocity between two consecutive points is greater than 300 km/h;
    return *Filtered points*;

Algorithm 4: Trajectory Data Compression

Input: Individual's trajectory data; spatial radius parameter r
Output: Compressed trajectory data
LenTraj ← Count the total number of observations in the raw trajectory;
Sorting: Sort by 'User ID' and 'Datetime';
Initialization: Create lists $[\phi]$ and $[\lambda]$, initializing them with ϕ_i and λ_i. Initialize $i = 0$ and $j = 1$;
for $i <$ *LenTraj* **do**
    for $j <$ *LenTraj* $+ 1$ **do**
        $d_{ij} \leftarrow$ Measure Haversine distance between (ϕ_i, λ_i) and (ϕ_j, λ_j);
        if $d_{ij} > r$ **then**
            break // End current for loop;
        end
        Add ϕ_j and λ_j to lists $[\phi]$ and $[\lambda]$;
        $j \leftarrow j + 1$;
    end
    $(\phi, \lambda) \leftarrow$ np.median($[\phi]$), np.median($[\lambda]$) // Replace each point in $[\phi]$ and $[\lambda]$ with the median point of each list;
    $i \leftarrow i + j$;
    $j \leftarrow i + 1$ // Update indices so that the compressed points are skipped in the next iteration;
end
return *Output* // The compressed trajectory data;

Algorithm 5: Elbow Method

Input: Data, max_clusters, plot
Output: Optimal number of clusters
Function *FindOptimalTripClusters(data, max_clusters)*:
    Initialize empty lists to store the inertia values and number of clusters;
    inertia_values ← [];
    num_clusters ← [];
    for *k* **to** *2* **to** *max_clusters + 1* **do**
        Apply K-means clustering;
        Compute inertia and number of clusters;
        Append inertia value and number of clusters to the lists;
    end
    Calculate differences in inertia values;
    Calculate differences ratio;
    Find the index of the elbow point;
    return *Optimal number of clusters*;

Algorithm 6: Train-Test Split

Input: DataFrame, k, random_state
Output: List of K folds
Function *TripLabelBasedKFoldSplit(df, k, random_state)*:
    Get unique trip IDs from the dataframe;
    Initialize KFold;
    List to hold the training and testing dataframes for each fold;
    Perform the k-fold split on the trip IDs;
    return *List of K folds*;

Algorithm 7: Trip Sampling

Input: Set of trips $Z_{w,e}$ for user w and mode e
Output: Three folds for training and testing
Function *TripSampling($Z_{w,e}$)*:
    FilteredTrips ← `FilterTrips(`$Z_{w,e}$, 300 *km/h*`)`;
    SimplifiedTrips ← `TrajectoryDataCompression(`*FilteredTrips,* 0.2 *km radius*`)`;
    if $|$*SimplifiedTrips*$| < 10$ **then**
        Continue to next set of trips;
    end
    else if $10 \leq |$*SimplifiedTrips*$| \leq 20$ **then**
        End;
    end
    else
        k ← `ElbowMethod(KMeansClustering(`*SimplifiedTrips*`))`;
        Clusters ← `KMeansClustering(`k, *SimplifiedTrips*`)`;
        Sort Clusters by size in descending order;
        TopTwoClusters ← First two clusters in Clusters;
        if $|$*SimplifiedTrips*$| > 20$ **then**
            Shuffle SimplifiedTrips;
            SampledTrips ← First 20 trips in SimplifiedTrips;
        end
        else if $|$*SimplifiedTrips*$| < 10$ **then**
            MoreClusters ← `KMeansClustering(`*SimplifiedTrips,* $k+1$`)`;
            SimplifiedTrips;
        end
        else
            End;
        end
    end
    `TrainTestSplit(`*SimplifiedTrips*`)`;

B.7 Handling Nonstationarity with GPs

Human mobility has various spatial and temporal scales (Alessandretti et al., 2020). Specifically, as a person moves, time determines their location at the level of minutes and seconds, while when that person is stationary at an activity location, time impacts their travel behavior at the level of hours. For modeling purposes, this implies that in a traditional time-series GP with a single temporal input, the lengthscale parameter of the GP will not be able to capture the impact of time evenly between the two states (moving and staying).

There are multiple ways to overcome this problem. The most common, which is the one we rely on in this paper, is to represent time using various calendar-based structures (e.g., seconds of the day, days of the week). This allows the GP to learn unique lengthscales across more degrees of freedom, providing the adaptive local smoothing necessary for non-stationary data. For example, while Unix time may have a small lengthscale (capturing second-level variations), other variables capturing the time of the day may have a lengthscale on the order of hours.

GPs are naturally well-suited to handle statistical nonstationarity as well. Duvenaud (2014) details various approaches to combining different kernels to attain nonstationary processes without having to detrend data. This tends to require usage of nonstationary kernels, like the Linear kernel (Equation B.1), which produces different predictions if the data were moved while the kernel parameters were kept fixed.

$$k_{LIN} = \mathbf{x}^\top \mathbf{x}' \tag{B.1}$$

To showcase GP's flexibility in handling nonstationarity, we have included an example of modeling the Mauna Loa CO2 concentration dataset using a GP with feature engineering (and without detrending). This dataset exhibits an increasing mean in the observed values over years. The usual approach to modeling this type of data using linear models is to first detrend the data (i.e., as shown here). However, we show that by creating cyclical features like months and years and leveraging the right kernel structure, this type of data can be modeled accurately using GPs (as seen in Figure B.6).

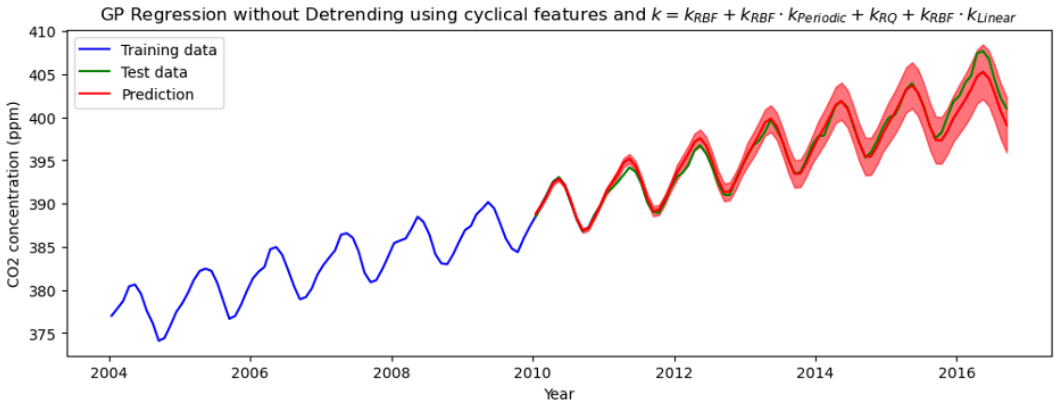


Figure B.6: GP regression using cyclical features and a composite kernel accurately captures nonstationary behavior in CO2 concentration of the Mauna Loa volcano

Appendix 3

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

C.1 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 $\{10^{-3}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$, 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 5×10^{-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 $256 \rightarrow 128$ units, while each single-task variant (STV) used $64 \rightarrow 32$ units. Since there are four STVs, their combined width $(4 \times 64, 4 \times 32)$ 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 C.1).

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

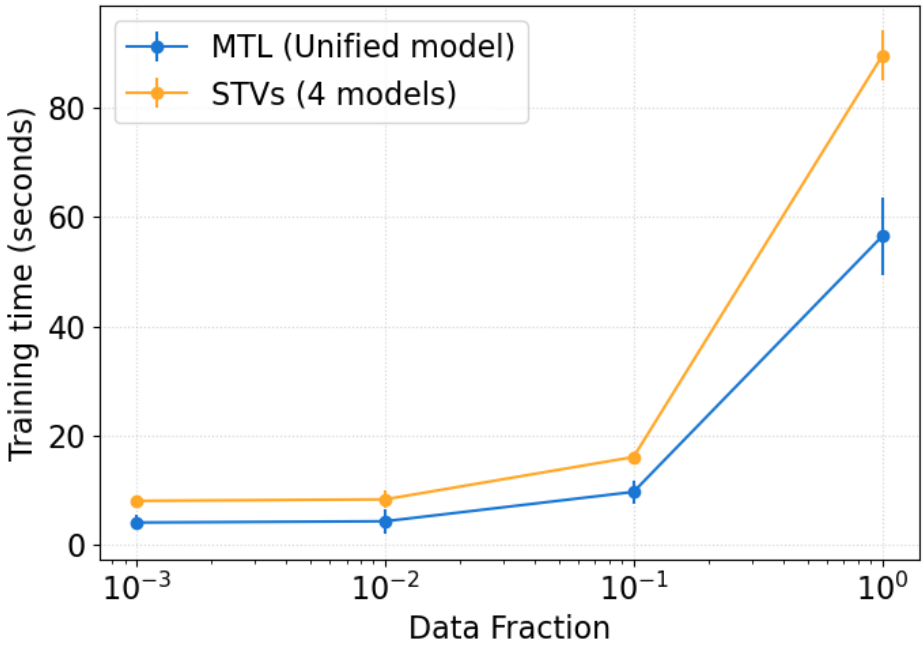


Figure C.1: Comparison of training times between the unified multi-task learning (MT) model and four separate single-task variants (STVs) across varying data fractions (log-scaled on x-axis). Despite its larger architecture, the MT model trains faster overall because its shared trunk amortizes computation across tasks, whereas STVs require four independent forward–backward passes. Error bars show the standard deviation across cross-validation folds.

Table C.1: Performance across feature sets and models for Age (overall 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

estimates for the SVC classifier, which slowed down convergence. The rest of this section contains the full results tables in C.1 until C.8.

C.1.1 Overall Split

Tables C.1-C.4 show the full model results in the overall split.

C.1.2 Cross-temporal Split

Tables C.5-C.8 show the full model results in the overall split.

Table C.2: Performance across feature sets and models for Gender (overall 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 C.3: Performance across feature sets and models for HH Income (overall 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 C.4: Performance across feature sets and models for Number of Children (overall 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**

Table C.5: Performance across feature sets and models for Age (2017–2019 train, 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**	**0.490 ± 0.001**	**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 C.6: Performance across feature sets and models for Gender (2017–2019 train, 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 C.7: Performance across feature sets and models for HH Income (2017–2019 train, 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 C.8: Performance across feature sets and models for Number of Children (2017–2019 train, 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