Multi-task, Multi-kernel Learning for Location-Based-Service (LBS) Data

Ekin Ugurel, Shuai Huang, Cynthia Chen

APSIPA ASC 2023

November 2nd, 2023

UNIVERSITY of WASHINGTON

Motivation

- **The past:** active solicitation (i.e., travel surveys)
  - Low sample sizes
  - Mixed reporting accuracy
  - Demographic info available
- **The present (and future):** passively-generated mobile data
  - Massive sample sizes
  - Found “in the wild”; data points are not generated due to any research-related processes
  - Prevalence of sparsity (large chunks of missing data)

Motivation

- Two pervasive issues:
  - As data collection practices become more transparent and user-centric, the sparsity issue only gets worse (DeGiulio et al., 2021)
  - Researchers are not able to share individual mobile data used in their studies due to privacy agreements with data providers (Gao et al., 2019; Rao et al., 2018; Sun et al., 2021; Li et al., 2023)
- The above motivates:
  1. An imputation method to correct missing data in GPS traces at various levels (Ugurel et al., under review)
  2. A generative modeling framework for individual mobile data to create synthetic datasets replicating real travel behavior (Ugurel, E., Huang, S., Chen, C., under review)

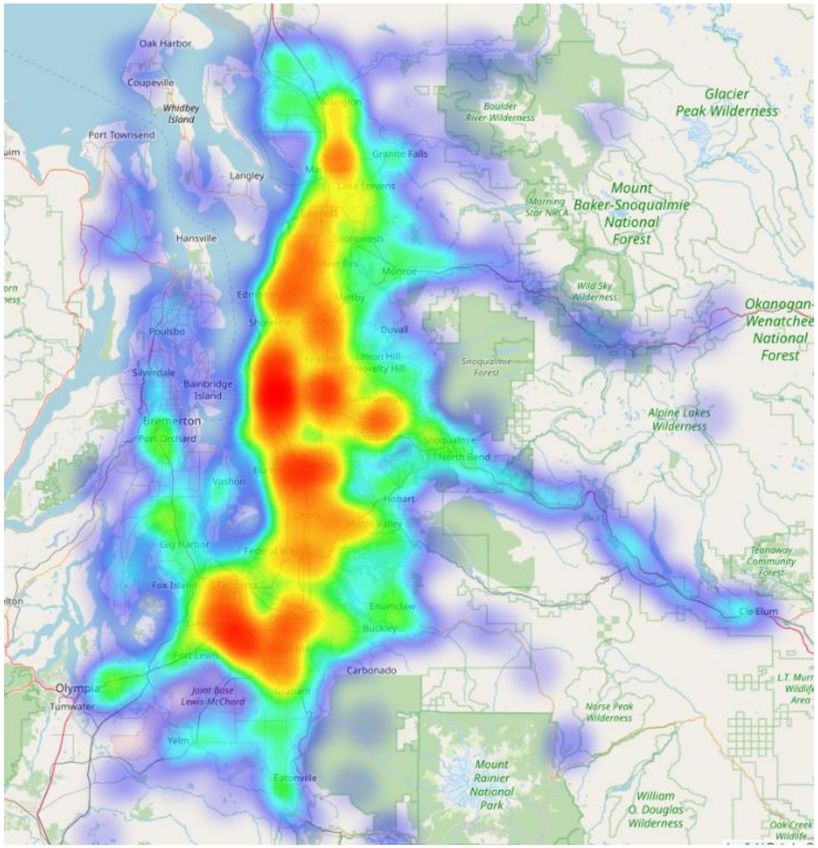
Challenges

> **Mode changes**
  – Can occur intra- or inter-trip

> **Heterogeneous human mobility behavior**
  – Varying tendencies to explore and exploit

Any method to correct missingness need to be flexible enough to capture these individual-level complexities

Spectus Dataset



Observations per user per day	
Mean	135
Standard Deviation	162
Min	1
25%	40
50%	98
75%	181
Max	9,159

(left) Heat map of a random sample of 20,000 GPS traces in the Greater Seattle Area; (right) summary GPS trace count statistics of the entire sample of 2,000 users

Research Question

> To what extent is a multi-task, multi-kernel learning framework a suitable method for correcting missingness in mobile data?

> How do we generate synthetic mobile data that replicates real individuals' travel behavior?

Multi-task Gaussian Process

The basic form of our location learning problem is

$$\mathbf{y} = f(\mathbf{X}) + \boldsymbol{\varepsilon},$$

where f specifies a systematic function of exogenous variables $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ is Gaussian white noise. We represent $\boldsymbol{y}$ through latitudes ϕ and longitudes λ

$$\mathbf{Y}^T = \begin{bmatrix} y_{1,\phi}, \ldots, y_{m,\phi} \\ y_{1,\lambda}, \ldots, y_{m,\lambda} \end{bmatrix},$$

where $y_{i,t}$ is the output for the t^{th} task on the i^{th} observation.

Given two correlated tasks, the covariance structure for the output vector can be specified as

$$\mathbf{K} = k(x_*, \mathbf{X})\mathbf{K}^f(y_\phi, y_\lambda),$$

where $\mathbf{K}^f$ is a PSD matrix containing the inter-task covariance and k is any valid PSD kernel.

Multi-task Gaussian Process

An inferred location y_* of a new input vector $\mathbf{x}_*$ conditioned on the training data is then assumed to be distributed as follows

$$y_* | \mathbf{x}_*, \mathbf{X}, Y, \sigma_y^2 \sim \mathbb{N}(y_*, \boldsymbol{\mu}_*, \boldsymbol{\sigma}_*^2),$$

$$\boldsymbol{\mu}_* = (k_t^f \otimes k_*)(\mathbf{K}^f \otimes \mathbf{K} + D \otimes \mathbf{I})^{-1} Y$$
$$\boldsymbol{\sigma}_*^2 = (k_t^f \otimes k_{**}) - (k_t^f \otimes k_*)(\mathbf{K}^f \otimes \mathbf{K} + D \otimes \mathbf{I})^{-1} (k_t^f \otimes k_*).$$

where $\otimes$ denotes the Kronecker product, k_t^f selects the t^{th} column of $\mathbf{K}^f$, $k_* = k(x_*, \mathbf{X})$ is the vector of covariance between the test point and the training set, and $k_{**} = k(x_*, x_*)$.

Finally, we minimize the negative marginal log-likelihood in determining the optimal model hyperparameters Θ

$$-\log(p(Y|\mathbf{X}, \Theta)) = \frac{1}{2}[Y^T (\mathbf{K} + \sigma_y^2 \mathbf{I})^{-1} Y + \log|\mathbf{K}| + m \log(2\pi)],$$

Kernels for Modeling Mobile Data

> **Squared Exponential (SE)**

$$K_{SE}(\mathbf{x}-\mathbf{x}') = \sigma^2 \exp\left(-\frac{1}{2\ell^2}|\mathbf{x}-\mathbf{x}'|^2\right)$$

> **Periodic (PER)**

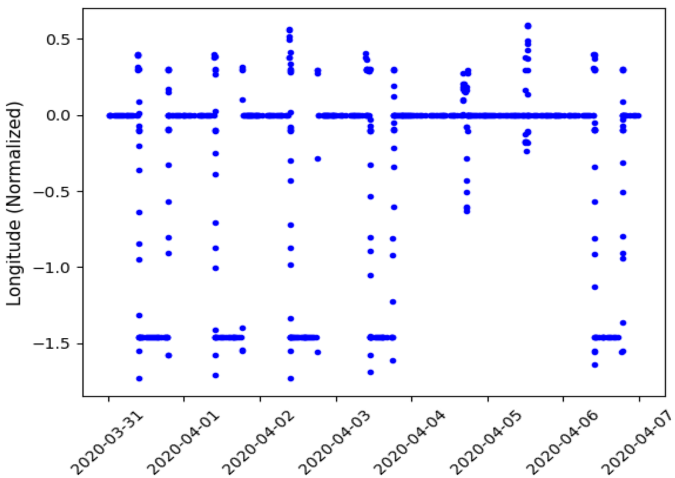
$$K_{PER}(\mathbf{x}-\mathbf{x}') = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|\mathbf{x}-\mathbf{x}'|/p)}{\ell^2}\right)$$

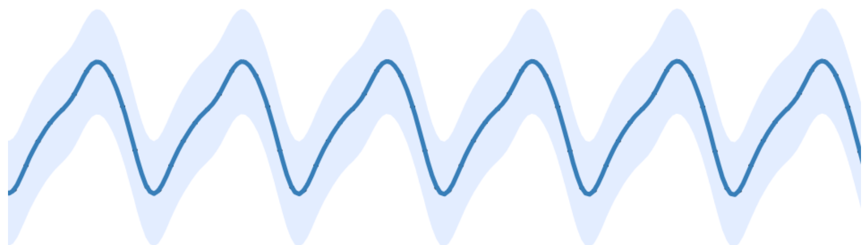
> **Rational Quadratic (RQ)**

$$K_{RQ}(\mathbf{x},\mathbf{x}') = \sigma^2 \left(1+\frac{(\mathbf{x}-\mathbf{x}')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

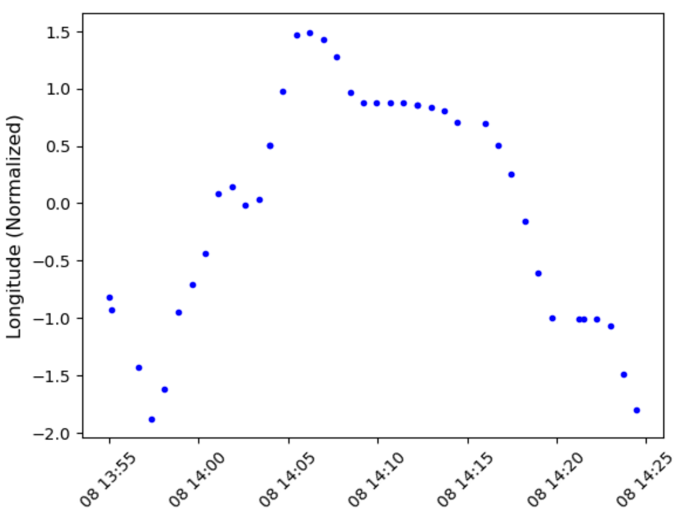
Where ℓ is a lengthscale (smoothing) parameter, σ^2 is the output variance, p is the period length, and α is the scale mixture (i.e., the relative weight of large- and small-scale variances)

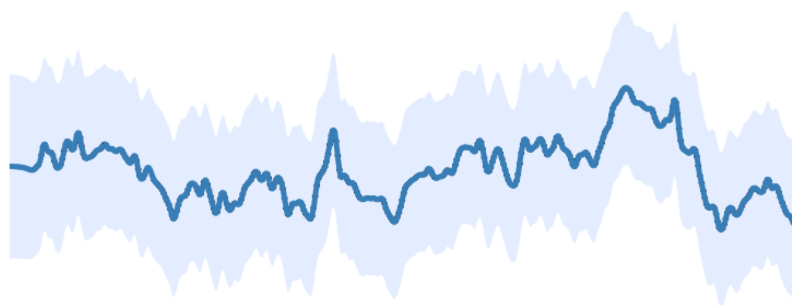
Kernels for Modeling Mobile Data





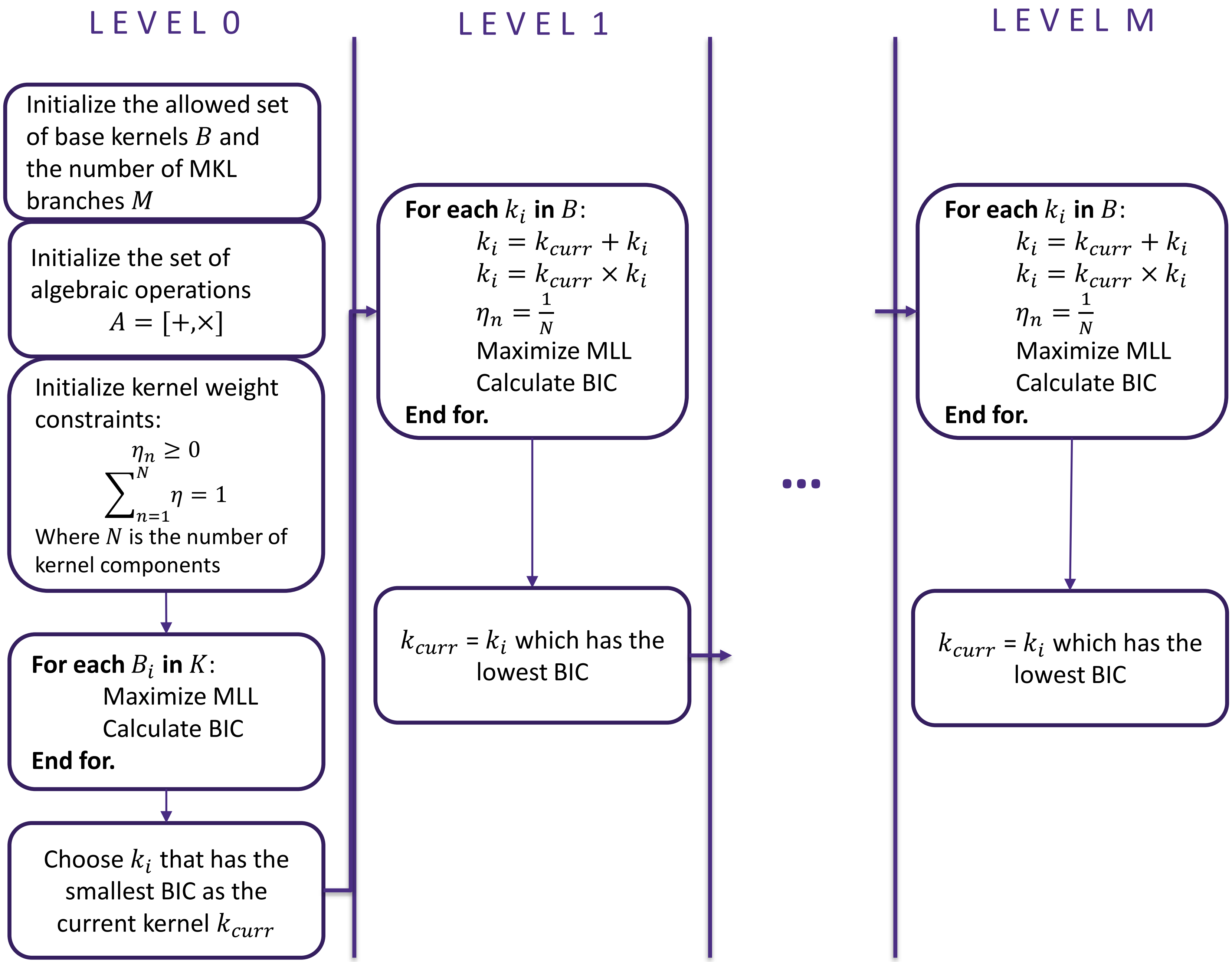
$K_{SE} \times K_{PER}$





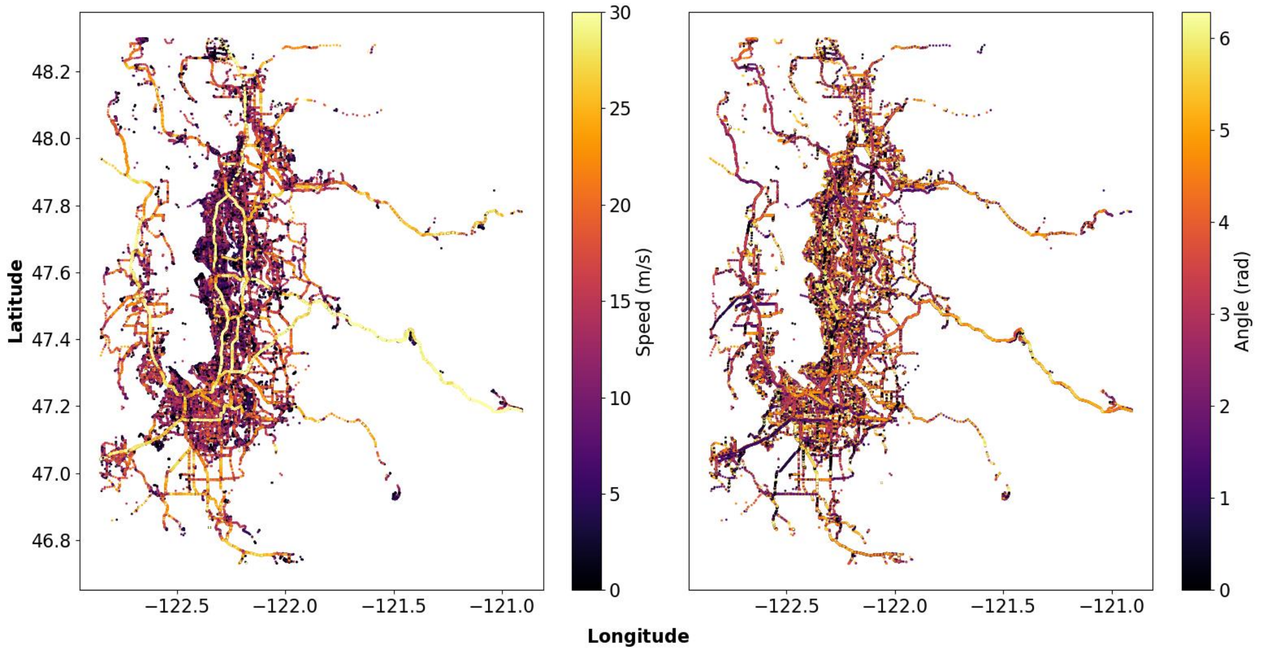
$K_{SE} \times K_{RQ}$

Greedy Multiple Kernel Learning



Physics-regularized GP

- Physical variables (i.e., instantaneous velocity, direction of travel) are functions of the transportation network
  - Speed limits, street widths, and traffic dictate how fast one can go in any given segment
  - Bodies of water or the existence of pavement dictate which direction one can travel at a given location



The Constrained Optimization Problem

We define functional constraints that reflect the limitations of human mobility within the given spatial and temporal context

$$\begin{aligned} \arg\min_{\Theta} \quad & -\log(p(\mathbf{v}, \mathbf{b} | \mathbf{X}, \Theta)) \\ s.t. \quad & v_i^*(\mathbf{x}_i) \leq v_{max} && \forall \mathbf{x}_i \in \mathbf{X} \\ & v_i^*(\mathbf{x}_i) \sim p(v | \mathbf{x}_i, \Theta) && \forall \mathbf{x}_i \in \mathbf{X} \\ & b_i^*(\mathbf{x}_i) \sim p(b | \mathbf{x}_i, \Theta) && \forall \mathbf{x}_i \in \mathbf{X}. \end{aligned}$$

However, functional constraints are hard to enforce within GPs. Instead, we enforce it on a set of constraint points $\mathbf{X}_c = \{x_c^{(u)}\}_{u=1}^{m}$

$$\begin{aligned} \arg\min_{\Theta} \quad & -\log(p(\mathbf{v}, \mathbf{b} | \mathbf{X}, \Theta)) \\ s.t. \quad & v_i(x_c^{(u)}) \leq v_{max} && \forall u = 1, \ldots, m \\ & v_i(x_c^{(u)}) \sim p(v | \mathbf{x}_i, \Theta) && \forall u = 1, \ldots, m \\ & b_i(x_c^{(u)}) \sim p(b | \mathbf{x}_i, \Theta) && \forall u = 1, \ldots, m. \end{aligned}$$

Implementation

- GPyTorch (Gardner et al., 2018)
  - Reduces the computational burden of exact GPs to $O(n^2)$.
    - Uses a modified batched version of linear conjugate gradients
- **Nonlinear optimization**
  - Adaptive Moment Estimation (Kingma and Ba, 2017)
  - Initialization is a prerequisite to avoid model misspecification

Table 1: Temporal dimensions used in our experiments

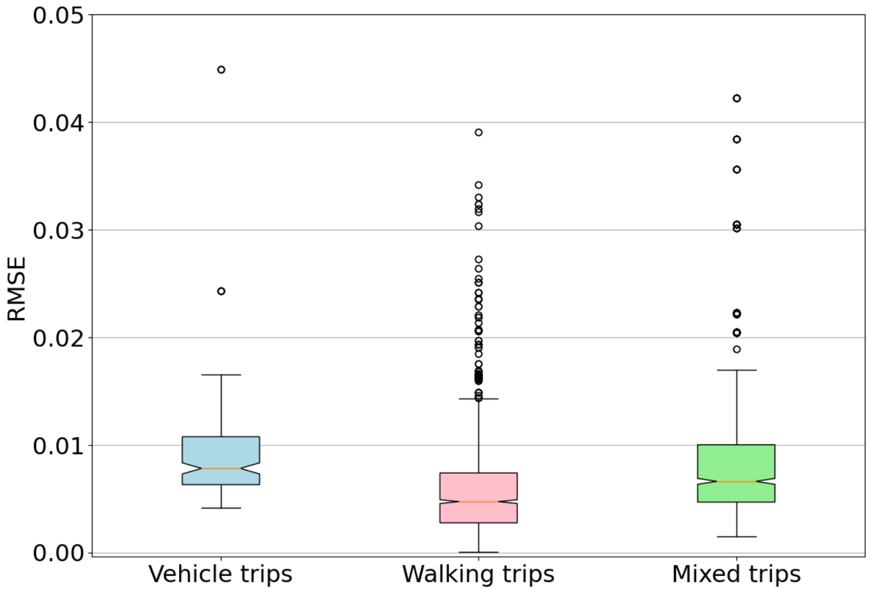
Variable	Notation	Type	Model Inputs
Unix time (normalized)	$\mathbf{t}_u$	Continuous	$[0, 1, \ldots, \tau]$
Hour Sine	$\mathbf{t}_{hs}$	Continuous	$[0, \ldots, 1]$
Hour Cosine	$\mathbf{t}_{hc}$	Continuous	$[0, \ldots, 1]$
Day of week	$\mathbf{t}_d$	Categorical	[0, 1, 2, 3, 4, 5, 6]
Week of the month	$\mathbf{t}_{wk}$	Categorical	[0, 1, 2, 3, 4]
Public holiday	$\mathbf{t}_{ph}$	Binary	[0, 1]
Weekend or not	$\mathbf{t}_{we}$	Binary	[0, 1]
AM peak	$\mathbf{t}_{am}$	Binary	[0, 1]
PM peak	$\mathbf{t}_{pm}$	Binary	[0, 1]

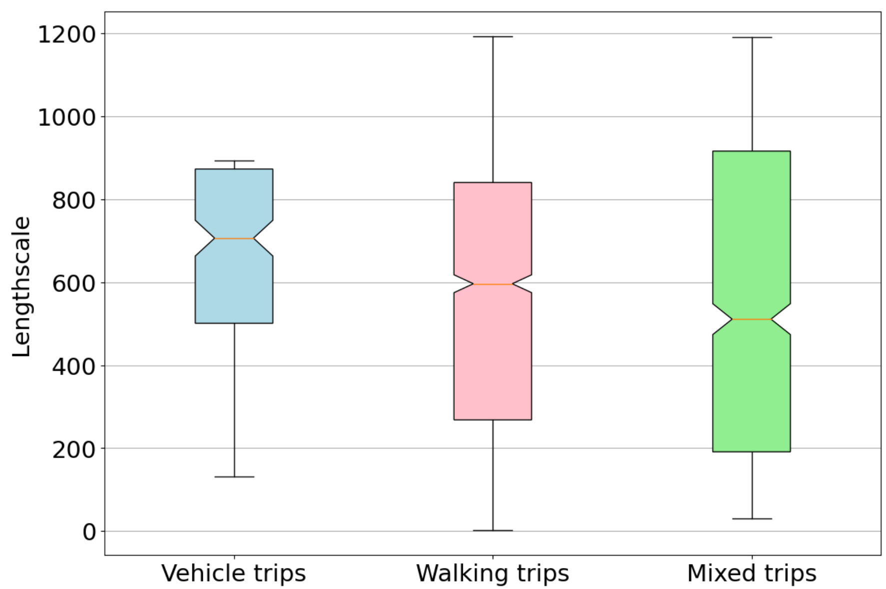
Experiments: Model Behavior for Different Types of Trips

K-means clustering by mobility metrics

Table 2: Summary of trip clusters

Cluster	Avg. Vel. [m/s]	Trip Distance [m]	Trip Duration [s]	Heading Change Rate	Velocity Change Rate	Number of Observations	Stop Rate
Slow, short trips	9.29	8,088	1,062	0.0019	0.0024	22.79	0.0007
Medium speed, medium distance	13.94	29,693	2,362	0.0007	0.0008	49.86	0.0002
Fast, distant trips	17.86	59,299	3,449	0.0005	0.0006	141.8	0.0001





Experiments: Robustness

Notation

We discretize a user's total available data time $\mathcal{T}$ into P intervals $(1, \ldots, P)$ 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

$$\mathrm{I}_p = \begin{cases} 1 & \textbf{if } p \textbf{ has at least one observation} \\ 0 & \textbf{otherwise} \end{cases}$$

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

$$q_\tau = \frac{1}{P} \sum_{p=1}^{P} \mathrm{I}_p$$

Mobility Metric Results

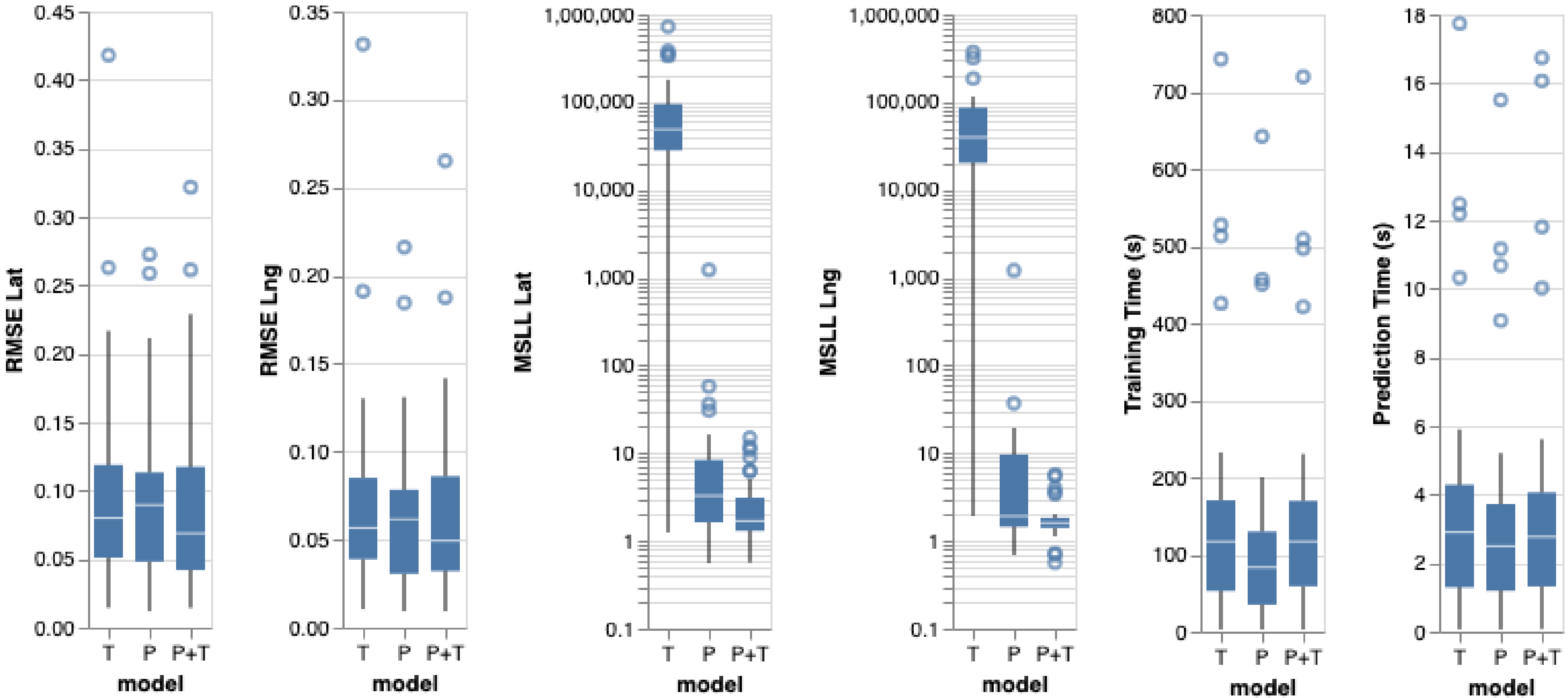
We find that the proposed method outperforms the competing algorithms in all classes of missingness gaps

Table 3: Median error of mobility metrics across varying gap lengths

Time Gap	Method	Number of Locations	Radius of Gyration	Straight Line Travel Distance	Random Entropy	Real Entropy	Uncorrelated Entropy
1 week	**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
	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
1 day	**MTGP**	**33.5**	**-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
	Holt	-1050	-0.846	-1303.06	-9.223	-9.61	-9.803
	ES	-1027	-0.718	-768.678	-4.878	-7.907	-5.225
	ARIMA	-1050	-0.834	-1303.06	-8.506	-9.609	-9.803
	SARIMAX	-1050	-0.909	-1303.06	-10.038	-9.612	-9.806
6 hours	**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
	Holt	-952	-0.645	-1171.79	-8.893	-9.493	-9.533
	ES	-929	-0.4	-768.56	-4.895	-7.869	-5.066
	ARIMA	-952	-0.645	-1178.65	-8.317	-9.493	-9.608
	SARIMAX	-956.5	-0.645	-1223.47	-9.901	-9.493	-9.751
1 hour	**MTGP**	**38.5**	**-0.074**	**389.308**	**0.053**	**0.29**	**0.157**
	RBF	-902	-0.94	-1319.47	-9.818	-9.548	-9.761
	SES	-901.5	-0.761	-1262.95	-7.989	-9.546	-9.642
	Holt	-901.5	-0.761	-1262.95	-7.041	-9.543	-9.642
	ES	-878.5	-0.627	-711.119	-4.8	-7.816	-5.174
	ARIMA	-898.5	-0.761	-1262.76	-6.801	-9.545	-9.613
	SARIMAX	-830.5	-0.644	-1161.14	-6.435	-9.455	-9.449
30 minutes	**MTGP**	**21**	**-0.435**	**123.606**	**0.048**	**0.314**	**0.166**
	RBF	-624.5	-1.36	-1043.86	-9.052	-8.954	-9.161
	SES	-614	-1.175	-1025.83	-7.405	-8.954	-9.006
	Holt	-614	-1.167	-1025.83	-6.872	-8.953	-8.952
	ES	-591	-1.145	-707.361	-4.099	-7.07	-4.445
	ARIMA	-614	-1.214	-1027.68	-6.282	-8.949	-8.952
	SARIMAX	-624.5	-1.36	-1043.86	-9.277	-8.954	-9.161
15 minutes	**MTGP**	**22**	**-0.299**	**-7.116**	**0.048**	**0.323**	**0.161**
	RBF	-670	-2.15	-1112.56	-8.925	-8.871	-9.125
	SES	-660	-1.71	-1162.11	-7.34	-8.871	-8.994
	Holt	-660	-2.099	-1162.07	-6.435	-8.871	-8.849
	ES	-637	-2.056	-513.035	-3.96	-6.771	-4.497
	ARIMA	-659	-1.931	-1168.42	-6.048	-8.871	-8.851
	SARIMAX	-670	-2.15	-1199.07	-9.39	-8.871	-9.146
5 minutes	**MTGP**	**21**	-0.824	**47.301**	**0.056**	**0.301**	**0.156**
	RBF	-896	-1.396	-1441.06	-9.791	-9.302	-9.571
	SES	-896	-1.274	-1391.03	-6.757	-9.302	-9.571
	Holt	-893	-1.012	-1391.03	-6.555	-9.302	-9.394
	ES	-872	**-0.744**	-666.535	-4.313	-6.643	-4.984
	ARIMA	-896	-1.339	-1391.03	-6.754	-9.302	-9.495
	SARIMAX	-896	-1.396	-1441.06	-9.809	-9.302	-9.571

Experiments: Physics-Regularization

Physics-regularized GP Performance



Conclusion and Future Work

- We have proposed a multi-task GP formulation to impute missing values in longitudinal mobile data
- By augmenting this model with multiple kernel learning and physics regularization, this can be a suitable generative modeling framework to generate synthetic data
- Future Work
  - Reducing computational complexity through approximation methods
  - Theoretical guarantees of convergence, accuracy bounds
  - Augmentation through Collaborative Learning

Acknowledgements

The authors are grateful to the funding support from the Center for Teaching Old Models New Tricks (TOMNET), a University Transportation Center sponsored by the US Department of Transportation through Grant No. 69A3551747116 and from the National Science Foundation for the project titled as "A whole-community effort to understand biases and uncertainties in using emerging big data for mobility analysis" (award number 2114260).

References (1/2)

> DeGiulio, A., Lee, H., Birrell, E., 2021. “Ask App Not to Track”: The Effect of Opt-In Tracking Authorization on Mobile Privacy, in: Saracino, A., Mori, P. (Eds.), Emerging Technologies for Authorization and Authentication, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 152–167. https://doi.org/10.1007/978-3-030-93747-8_11

> Gao, J., Sun, L., Cai, M., 2019. Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. Transportation Research Part C: Emerging Technologies 104, 78–94. https://doi.org/10.1016/j.trc.2019.04.022

> Rao, W., Wu, Y.-J., Xia, J., Ou, J., Kluger, R., 2018. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. Transportation Research Part C: Emerging Technologies 95, 29–46.https://doi.org/10.1016/j.trc.2018.07.002

> Sun, J., Kim, J., 2021. Joint prediction of next location and travel time from urban vehicle trajectories using long short-term memory neural networks. Transportation Research Part C: Emerging Technologies 128, 103114. https://doi.org/10.1016/j.trc.2021.103114

> Li, G., Chen, Y., Wang, Y., Nie, P., Yu, Z., He, Z., 2023. City-scale synthetic individual-level vehicle trip data. Sci Data 10, 96. https://doi.org/10.1038/s41597-023-01997

> Wang, F., J. Wang, J. Cao, C. Chen, and X. (Jeff) Ban. Extracting Trips from Multi-Sourced Data for Mobility Pattern Analysis: An App-Based Data Example. *Transportation Research Part C: Emerging Technologies*, Vol. 105, 2019, pp. 183–202. https://doi.org/10.1016/j.trc.2019.05.028.

References (2/2)

- McCool, D., P. Lugtig, and B. Schouten. Maximum Interpolable Gap Length in Missing Smartphone-Based GPS Mobility Data. *Transportation*, 2022. https://doi.org/10.1007/s11116-022-10328-2.
- Rasmussen, C. E., & Williams, C. K. I. Gaussian processes for machine learning. MIT Press, 2006.
- Gardner, J., G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. No. 31, 2018.
- Kingma, D. P., and J. Ba. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980. Accessed Jul. 9, 2022.
- Ugurel, E., Guan, X., Wang, Y., Huang, S., Wang, Q., Chen, C. Correcting Missingness in Passively-Generated Mobile Data with Multi-task Gaussian Processes. *Under Review,* 2023.
- Ugurel, E., Huang, S., Chen, C., Physics-regularized, Multi-task Gaussian Processes with Multiple Kernel Learning to Uncover Mobile Data Generation Patterns. *Under Review*, 2023.