A Physics-regularized, Multi-task Gaussian Process With Multiple Kernel Learning To Uncover Mobile Data Generation Processes

Ekin Ugurel, Shuai Huang, Cynthia Chen

Large-scale Data Analytics for Transportation Systems
INFORMS Community/Committee Choice Session

October 15th, 2023

UNIVERSITY *of* WASHINGTON

Motivation

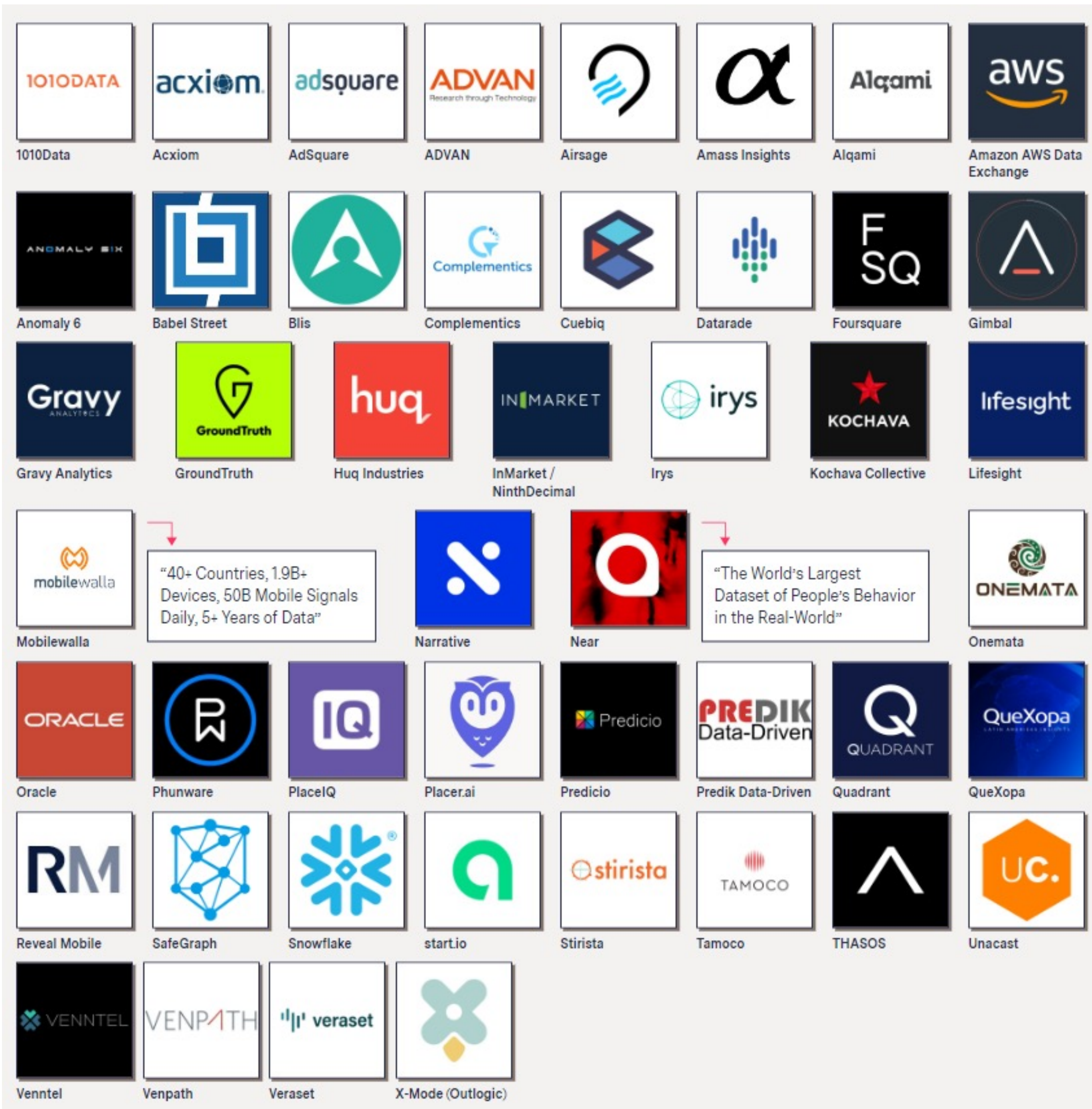
> **The past:** active solicitation (i.e., travel surveys)

- 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

The Location Data Industry: Collectors, Buyers, Sellers, and Aggregators



[Source: The Markup]

Motivation

- > Two pervasive issues:
  - – As data collection practices become more transparent and user-centric, the sparsity issue only gets worse[1]
  - – Researchers are not able to share individual mobile data used in their studies due to privacy agreements with data providers[2, 3, 4, 5]

- > The above motivate a generative modeling framework for individual mobile data to create synthetic datasets replicating real travel behavior

Challenges

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

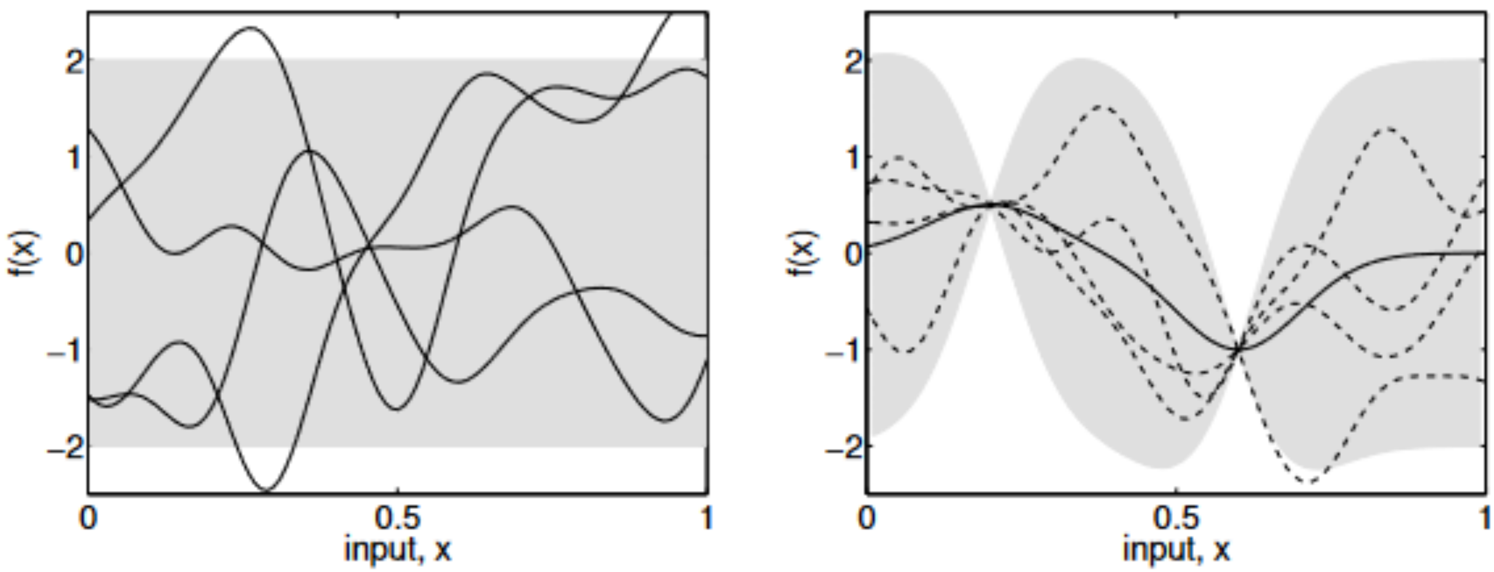
> **Heterogeneous human mobility behavior**
  - Varying tendencies to explore and exploit

Any generative method needs to be flexible enough to capture these individual-level complexities

Research Question

> **How do we generate synthetic mobile data that replicates real individuals' travel behavior?**
>
> – To what extent are kernel methods (i.e. Gaussian processes) suitable to act as generative modeling frameworks for individual trip data?

GPs consider the space of all possible models and output the most likely given your training data



(a), prior (b), posterior

Panel (a) shows four samples drawn from the prior distribution. Panel (b) shows the situation after two datapoints have been observed. The mean prediction is shown as the solid line and four samples from the posterior are shown as dashed lines. Shaded region denotes twice the standard deviation at each input value x

Multi-task Gaussian Process

The basic form of our location learning problem is

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

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

$$Y = (y_{1,\phi}, \ldots, y_{m,\phi}; y_{1,\lambda}, \ldots, y_{m,\lambda}),$$

where y_{it} 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

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

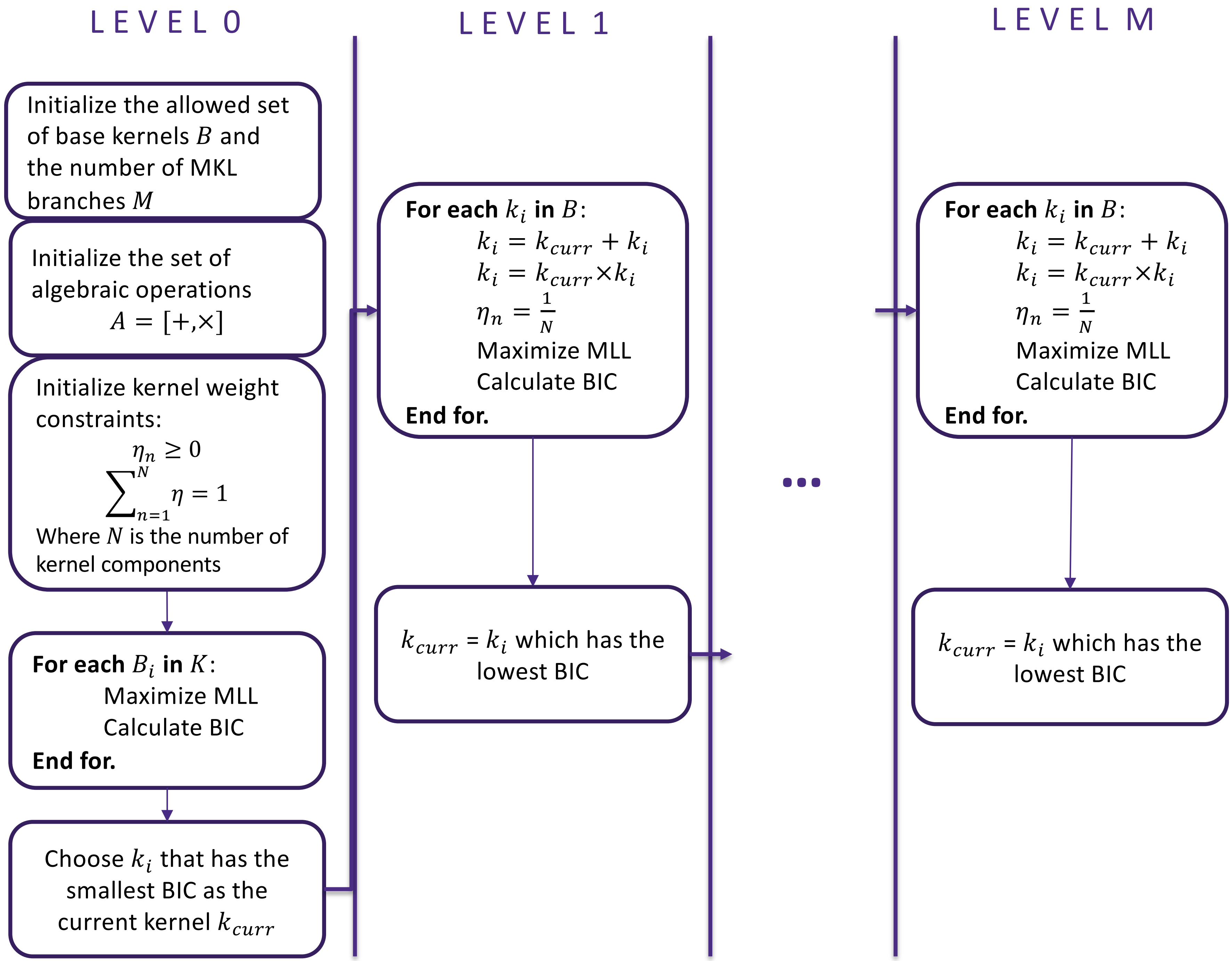
$$\boldsymbol{\mu}_* = (k_t^f \otimes k_*)(\mathbf{K}^f \otimes \mathbf{K} + D \otimes \mathbf{I})^{-1} Y$$
$$\boldsymbol{\sigma}_*^2 = (k_t^f \otimes k_{**}) - (k_t^f \otimes k_*)(\mathbf{K}^f \otimes \mathbf{K} + D \otimes \mathbf{I})^{-1}(k_t^f \otimes k_*).$$

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 $k_{**} = k(x_*, x_*)$.

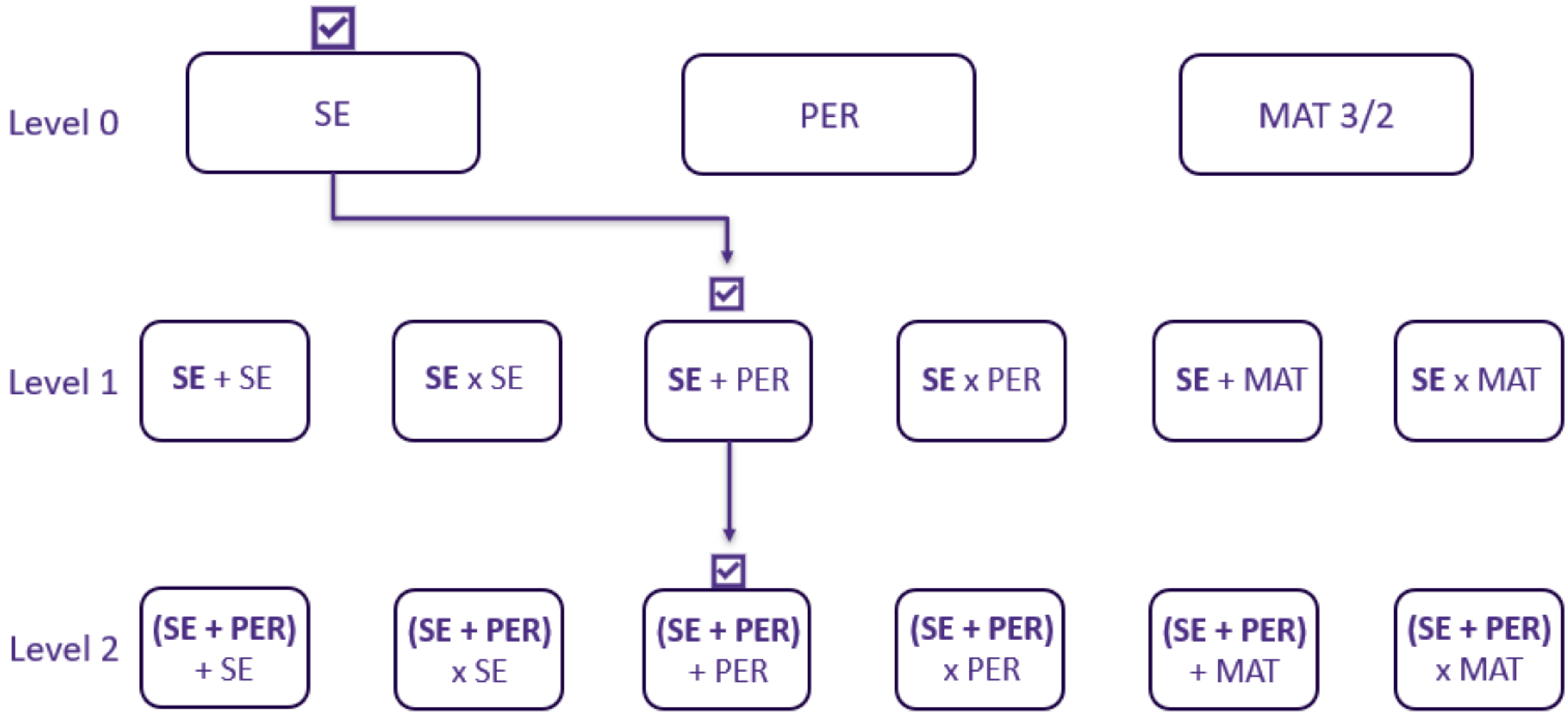
Finally, we minimize the negative marginal log-likelihood in determining the optimal model hyperparameters Θ

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

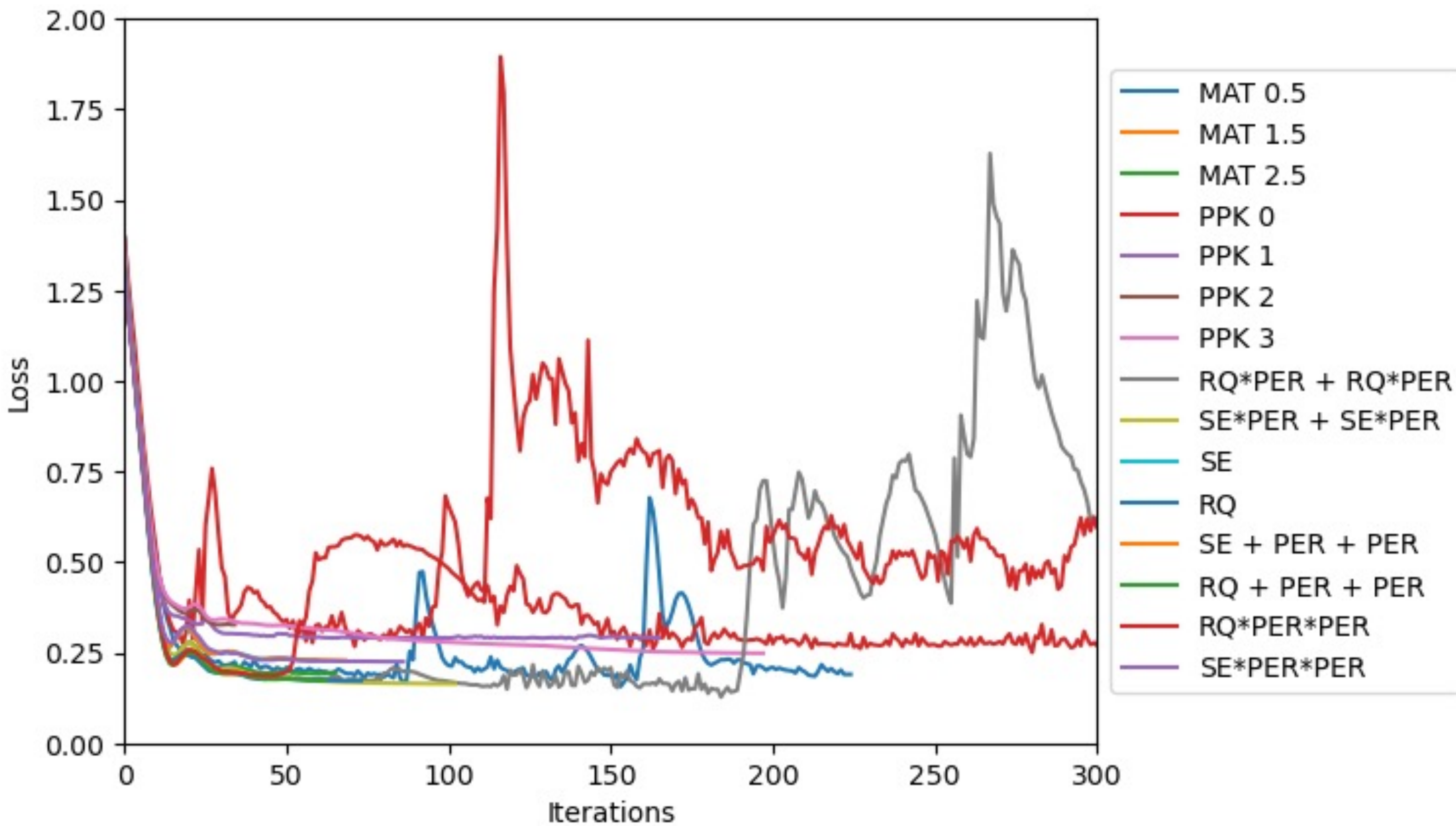
Greedy Multiple Kernel Learning



Example MKL progression

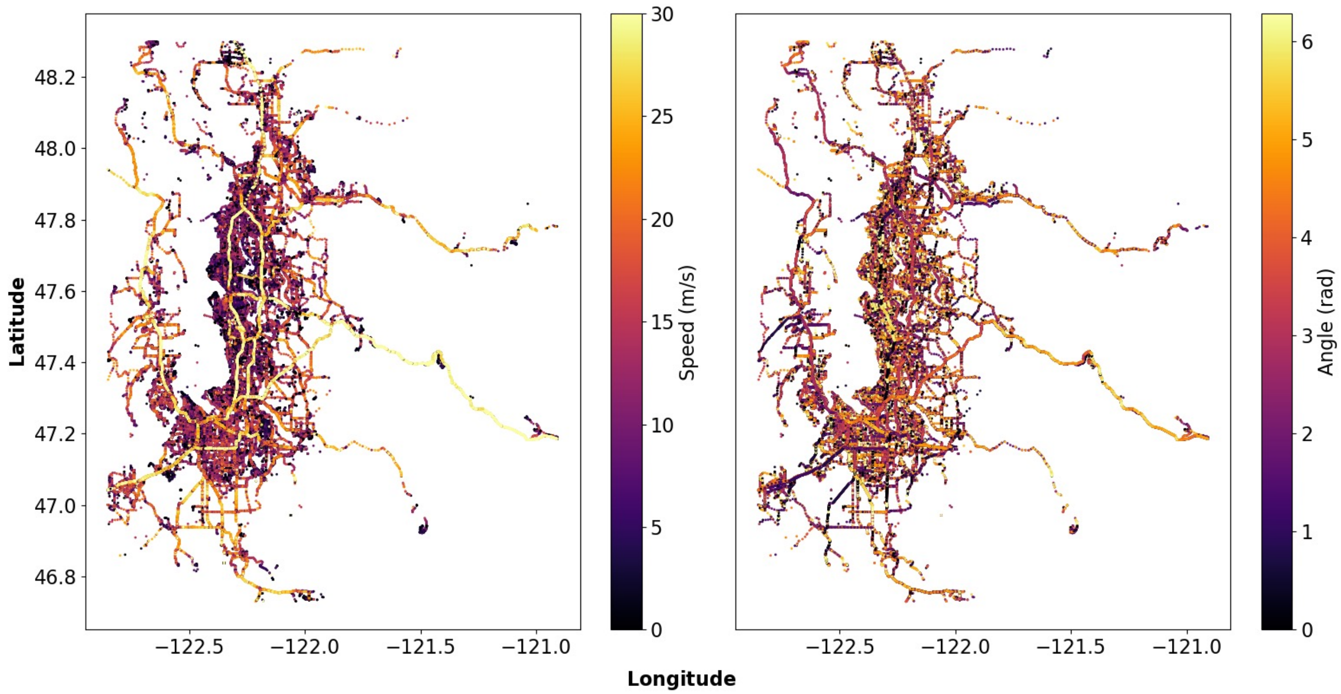


Different composite kernels showcase varying convergence behavior



Physics-informed 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

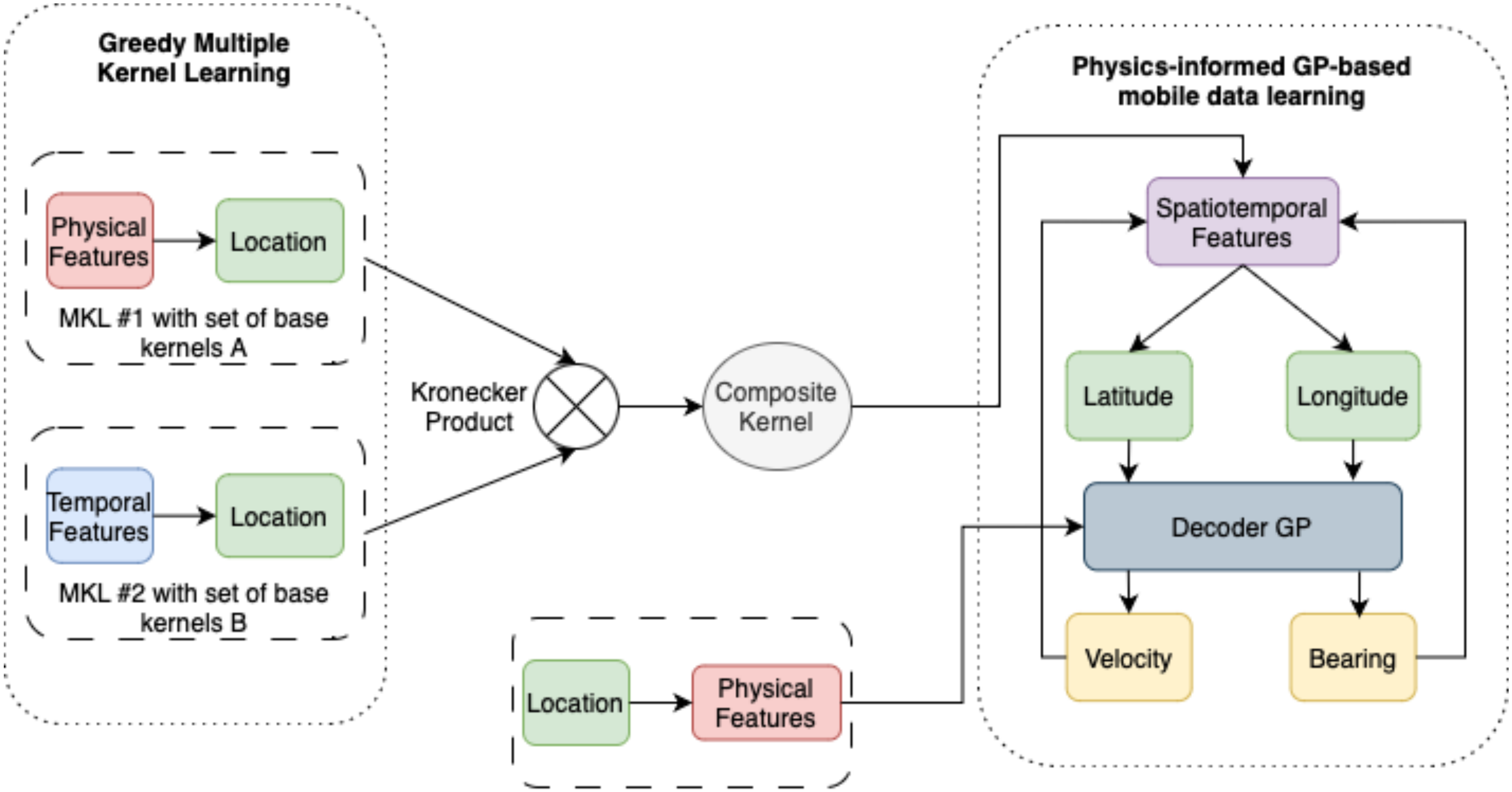
We define functional constraints that reflect the limitations of human mobility within the given spatial and temporal context

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

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

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

Model Framework

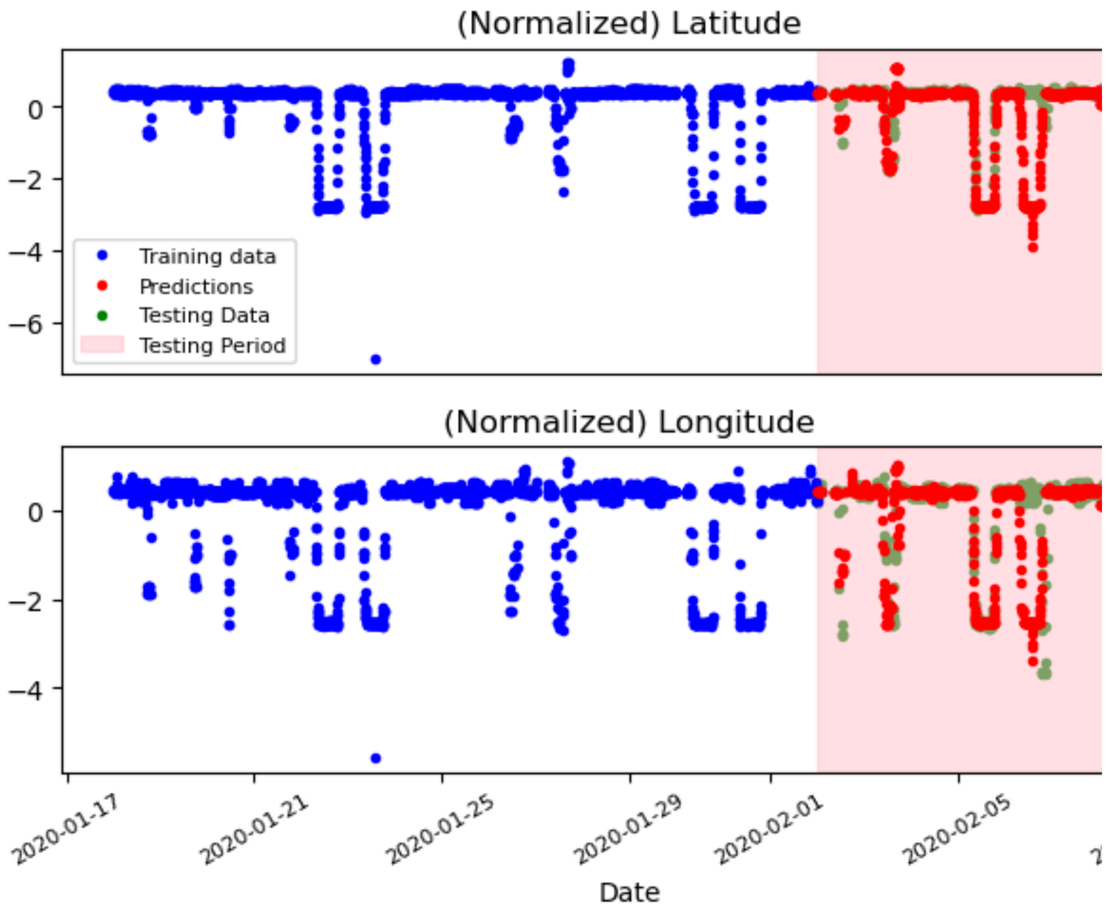


Implementation

- Jan 2020 – July 2020
- Greater Seattle Area

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]

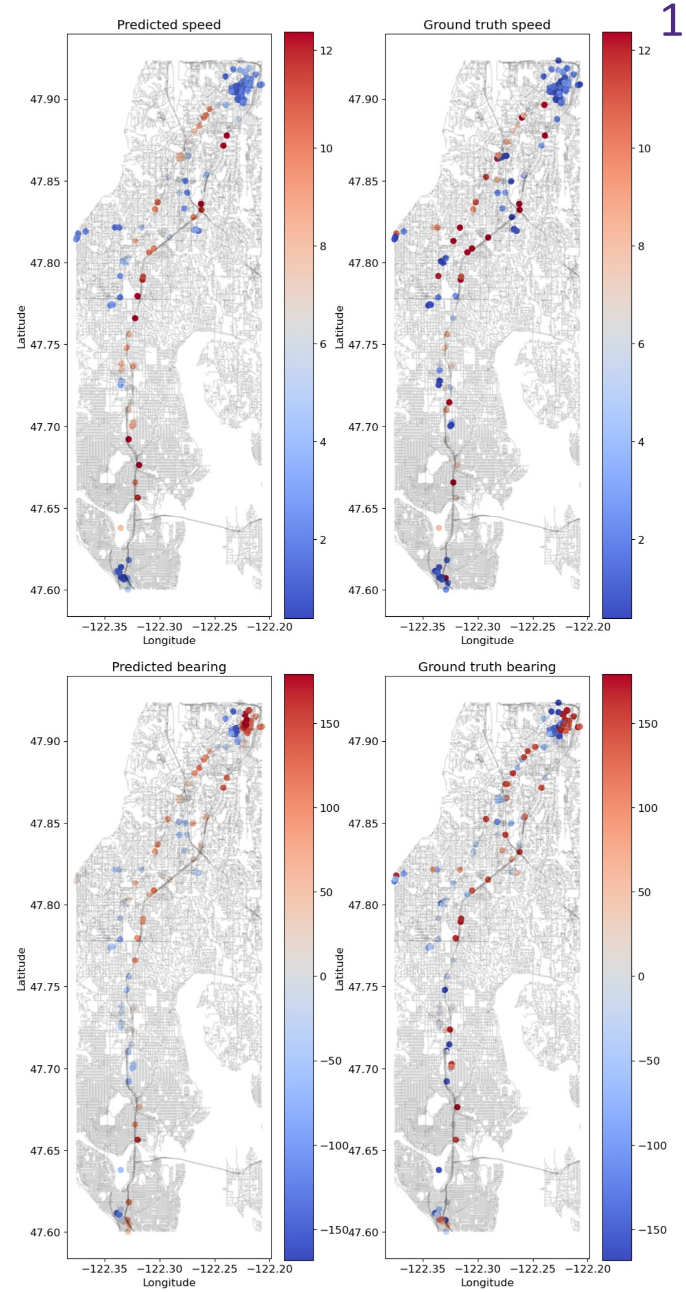
Learning Temporal Patterns



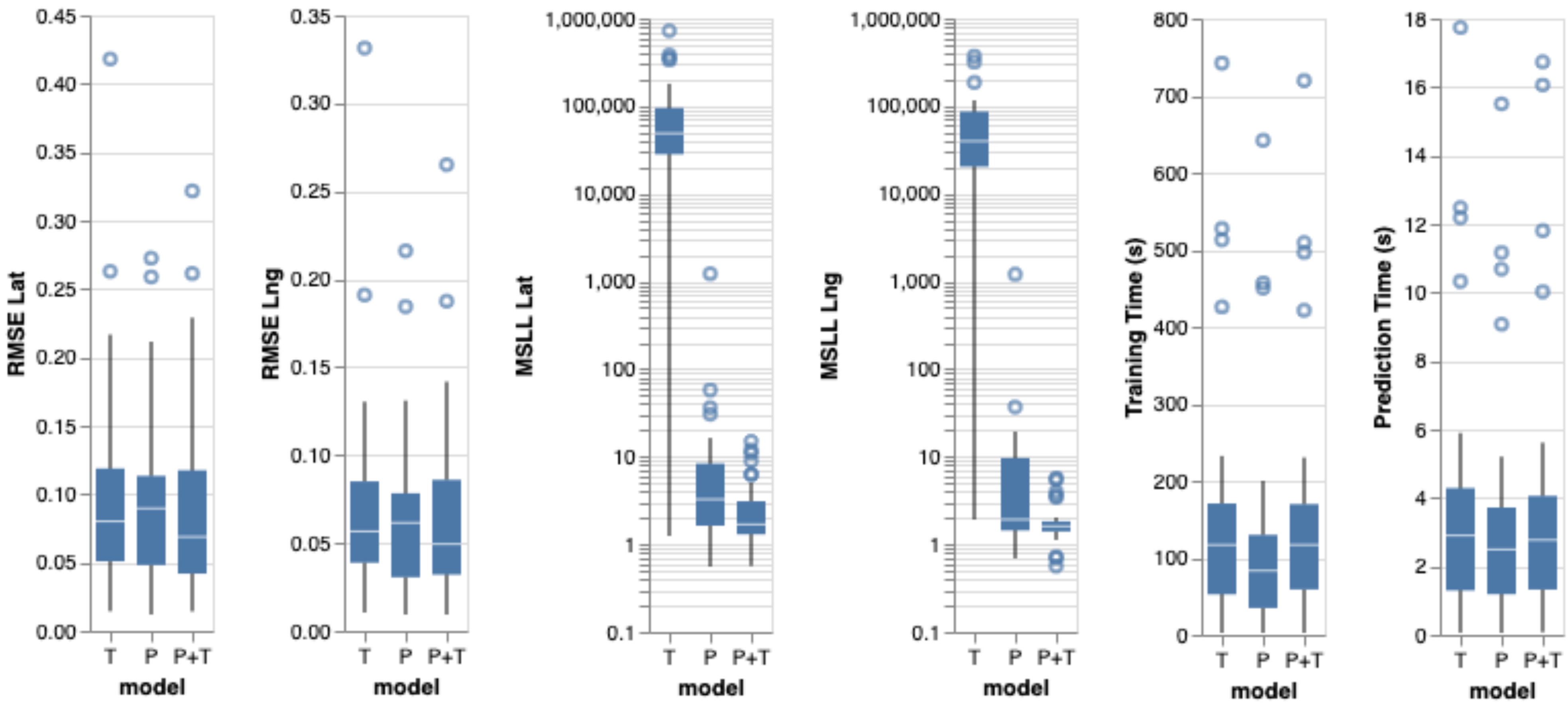
$$K = \prod_{d=1}^{n} K_{RQ,d} \times K_{PER,1} + \prod_{d=1}^{n} K_{RQ,d} \times K_{PER,1}$$

Learning Physical Constraints

$$K = K_{MAT,1} \times K_{MAT,2} + K_{RQ,1} \times K_{RQ,2}$$



Results


RMSE Lat
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45
RMSE Lng
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
MSLL Lat
0.1 1 10 100 1,000 10,000 100,000 1,000,000
MSLL Lng
0.1 1 10 100 1,000 10,000 100,000 1,000,000
Training Time (s)
0 100 200 300 400 500 600 700 800
Prediction Time (s)
0 2 4 6 8 10 12 14 16 18
T P P+T
model

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] 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

[2] 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

[3] 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

[4] 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

[5] 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/s41597-023-01997-4