Advancing Temporal Multimodal Learning with Physics Informed Regularization

Niharika Deshpande
Computational Data Science & Engineering
North Carolina A&T State University
Greensboro, USA

Hyoshin Park
Computational Data Science & Engineering
North Carolina A&T State University
Greensboro, USA
hpark1@ncat.edu

Venktesh Pandey
Civil, Architectural and Environmental Engineering
North Carolina A&T State University
Greensboro, USA

Gyugeun Yoon
Computational Data Science & Engineering
North Carolina A&T State University
Greensboro, USA

***Abstract*—Estimating multimodal distributions of travel times from real-world data is critical for understanding and managing congestion. Mixture models can estimate the overall distribution when distinct peaks exist in the probability density function, but no transfer of mixture information under epistemic uncertainty across different spatiotemporal scales has been considered for capturing unobserved heterogeneity. In this paper, a physics-informed and -regularized prediction model is developed that shares observations across similarly distributed network segments across time and space. By grouping similar mixture models, the model uses a particular sample distribution at distant non-contiguous unexplored locations and improves TT prediction. Compared to traditional prediction without those updates, the proposed model's 19% of performance show the benefit of indirect learning. Different from traditional travel time prediction tools, the developed model can be used by traffic and planning agencies in knowing how far back in history and what sample size of historic data would be useful for current prediction.**

***Index Terms*—Optimal Learning, Regularization, Spurious Correlation, Multimodal Probability Distribution, Spatiotemporal Correlation**

I. Introduction

Travel time reliability has attracted attention and "buffer time" has been used to indicate extra time to allow for traffic delays. However, unimodal assumption does not distinguish different probability density functions (PDFs). Travel time PDFs on freeways have shown *two or more modes as distinct peaks* due to the mixes of driving patterns and vehicle types [1]. The multimodal distribution exists on arterial roads, where a vehicle passing a signal at the end of the green would experience quite a different travel time than the vehicle following behind it that must make a stop for the red, although they traveled next to each other [2].

Current navigation systems (e.g., Google Maps) are not customized to users' tolerance for unexpected delays. Authors' previous work [3] could significantly reduce the traffic delays by providing en-route suggestions to informed drivers using predicted information about the time-varying route habits of uninformed drivers. However, the network is dynamic and the route suggestion users receive at the outset of their commute may not be optimal when they are on the road. While those complex patterns can be captured as unobserved heterogeneity using data-driven models, incorporating physics knowledge can regularize the spurious correlation that may exist in the data-driven models.

Temporal Multimodal Multivariate Learning (TMML) [4], [5] addressed the above challenges by indirectly learning and transferring online traffic information from multiple modes of probability distributions and multiple variables across different time stages. A location's observed traffic data could be used to forecast conditions at distant non-contiguous locations. This was achieved by aggregating the traffic data from all the grid cells and clustering cells that have similar probability distributions. When one cell of a cluster is explored, the traffic information gained from the explored cell can partially remove uncertainty about the conditions in distant non-contiguous unexplored cells of the same cluster. Those travel time mixture is clustered as white cell type numbers in Fig. 1 under lower and upper bounds with probability $P(T)$.

The local and non-contiguous spatiotemporal correlation may not be caused by the same type of unobserved heterogeneity, which we call "false causation". Under *multimodal* distribution as combination of traffic patterns, although correlation may exist between one distribution with non-recurring congestion on low volume road and another distribution with recurring congestion on high volume road, we cannot conclude that those two distributions are causally related. This paper advances the data-driven TMML by decoupling spurious correlation for two fundamental diagrams grouped in a cluster with high confidence. [6] partially filled this gap by grouping similar types of the bimodal output distribution of images classified by mixture density network. Maximum entropy seeking transfer learning was superior to partially observable Markov decision processes.

NSF 2106989, 2200590, 1910397

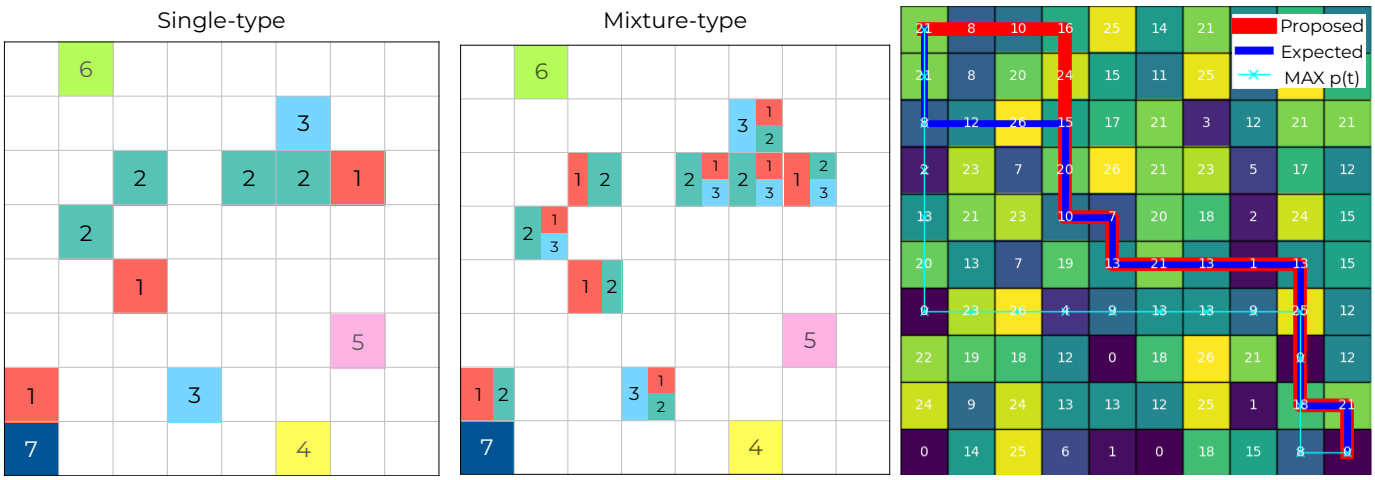

Fig. 1: Sequential information gain starting from the single-type to mixture information gain

However, a complete removal of uncertainty was made for each observation. Partial information gains from multimodal distribution and multivariate correlation have been left unaddressed, possibly due to the main focus of reinforcement learning (RL) on games and simple control problems with a lack of generalization to real-world problems. In a sequence of transfer learning, the RL does not utilize the covariance structure and ignore multimodal and multivariate gains in the reward function. Hybrid deep learning traffic studies extract spatiotemporal correlations [7]–[9], however static graph are unable to capture dynamic nature of traffic. Recent adaptive adjacency graphs [10], [11] could be alternative, however, multimodality has not been considered.

This study endeavors to create a new principled multimodal learning to predict travel time during simultaneous inference of spatiotemporal multimodal observations. We start extending [6] to ensure that locations with broad bimodal probability distributions are preferred over locations with narrow probability distributions.

II. Temporal Multimodal Learning for Mixture Model

A multi-modal urban transportation network gives travelers a variety of options for getting around. The literature treats links as a unimodal probability distribution with an expected travel time. If the assumption that the cell states are correlated is true, then visiting one cell will improve the state estimate of all cells that share similar travel time probability distribution.

The temporal multimodal learning captures mutual dependency between states under the impact of exogenous variables. Once we have additional observations within the same cluster, a new entropy method is used to estimate the mixture of multimodal and multivariate distributions. However, Shannon entropy [12] cannot distinguish distributions with multiple weights (e.g., bimodal distributions) because it only considers raw information gain, treating all information as equally valuable. Kullback-Leibler (KL) Divergence [13] introduces a bias toward only one mode (e.g., Exclusive, Reverse) or toward the mean of the modes (e.g. Inclusive, Forward) with non-symmetrical measures of information gain. Recent learning models [14] cannot address unobserved heterogeneity causing multimodal distributions since representing the information gain using KL Divergence requires comparison to an "ideal" distribution. This biases the model towards searching only for some types of solutions while ignoring more valuable solutions. When the probability distribution is heavily weighted at either extreme, the system cost either experiences positive true savings or negative true savings. The *Expectation-Maximization* algorithm cannot be guaranteed since the type of mixture probability vary across time and space and evolve as new observations become available.

If an identical cell is visited by another traveler and found to be in the same state as the original cell of that type, then all travelers have confirmation that the assumption that these cells are correlated is more likely to be true. Instead of directly using the classical entropy to combine current and historical observations, Kalman Filtering (KF) systematically split the database: starts with prediction based on historical data; and then once the new observation is available, follows hierarchical steps:

1) remove the spurious correlation
2) estimate the mixture of multimodal and multivariate sample distributions using *cross entropy* method. Here the parent distribution of observations is used to identify the cluster for the new observation.

III. Methodology

The distinguishable aspects of the physics-informed and -regularized (PIR) KF in the hierarchical update steps is the use of new data obtained from multimodal multivariate learning (Fig. 2). The global correlation between non-contiguous cells of an entire map are estimated by using Expectation Maximization.

We learn and predict traffic speed v_c^t within day by analyzing the spatiotemporal correlations between random variables v_c^t for all $(c,t) \in C \times T$. By clustering all v_c^t variables,

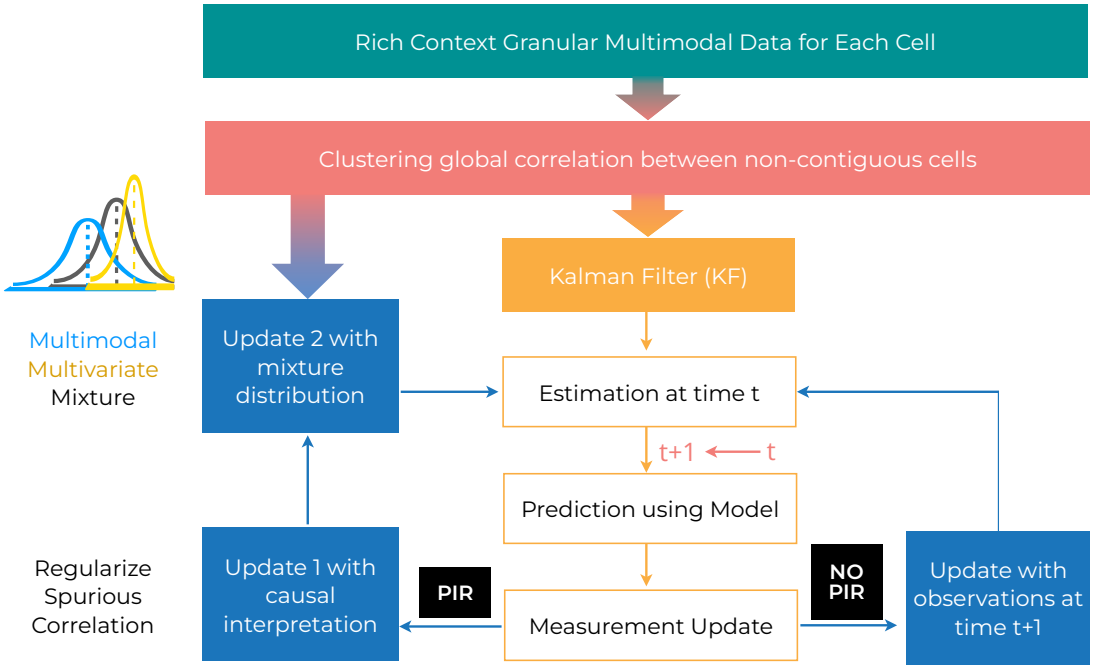

Fig. 2: Physics-informed and -regularized (PIR) KF in the hierarchical update steps

we identify spatiotemporal patterns and different combinations of traffic speed distributions. If $v_{c_1}^{t_1}$ is clustered with random variable $v_{c_2}^{t_2}$, then knowing information about the realization of $v_{c_1}^{t_1}$ will lower the uncertainty (measured using standard deviation) for $v_{c_2}^{t_2}$.

Prediction-Collection Step We project the state at time t using the prediction at previous time $t-1$ as $\hat{x}_t^- = A\hat{x}_{t-1}^+ + B\mu_t$ and error covariance of state as $P_t^- = P_{t-1}^+ A^T + Q$. We determine the Kalman Gain at time t as $K_t = P_t^- H^T (H P_t^- H^T + R)^{-1}$ where H is the connection matrix between the state vector and the measurement vector and R is the data precision matrix. In case of KF-no PIR, Z_t are the speed observations on a given day while in case of PIR-KF, Z_t are mean and variance of historical speed data. Incorporating unobserved local heterogeneity in the distributed data stream requires careful learning. Based on a cluster of similar fundamental diagrams (speed and density), density or flow are used for predicting a speed at different locations and times if the correlation is within the same cluster.

We first decouple spurious correlations and then use the entropy method to estimate the mixture of multimodal and multivariate distributions. Since the mixture could be non-Gaussian and non-linear, providing an accurately estimated distribution rather than just mean and standard deviation will increase the accuracy of updating the error covariance matrix.

First update step: When there are two conflicting observations from multimodal and multivariate clusters at the same time and location, then we investigate the original cell distribution in which observation belongs. When a spurious correlation is suspected, we should use only one of the trustable observations. Spurious correlations are filtered using the coefficient of variation parameter given by $CV = \frac{\sigma}{\mu}$, to show mean changes according to the standard deviation as a measure of relative variability. During the update step, observations available from the correlated links from previous time intervals are considered. The mean and variance of speeds of all correlated links are treated as the new observation. Instead of only one update step, we have two updates, one with mean and variance of historical data and the other with correlated speed data obtained from the clustering step. We address the question of how best can we predict $v_c^{(k+1)}$ if we know v_c^k.

Second update step: To estimate a mixture of multimodal and multivariate distributions, we use multivariate Gaussian mixture by a cross-entropy method. Stochastic likelihood maximization is based on cross entropy method. The cross-entropy metric measures the relative entropy between the true distribution f and the proposed mixture of multimodal and multivariate probability distributions g. Considering a random variable $\mathbf{X} = (X_1, \ldots, X_n)$ with support $\mathcal{X}$, the relative entropy between the two continuous probability density functions f and g will be defined as expected $\mathbb{E}_f$ value based the choice of mixture parameters θ that minimizes the cross-entropy. We will minimize the relative entropy between the true distribution f and the mixture of multimodal and multivariate distributions g parameterized by θ:

$$\boldsymbol{\theta}_g^* = \arg\min_{\boldsymbol{\theta}_g} - \int_{\mathbf{x}\in\mathcal{X}} f^*(\mathbf{x}) \log g\left(\mathbf{x} \mid \boldsymbol{\theta}_g\right) d\mathbf{x} \qquad (1)$$

The cross-entropy-method uses a multi-level algorithm to estimate $\boldsymbol{\theta}_g^*$ iteratively. Specifically, the parameter $\boldsymbol{\theta}_k$ at iteration k is used to find new parameters $\boldsymbol{\theta}_{k'}$ at the next iteration k'.

The multivariate relationship with fundamental diagram will be further considered in the regularization step by the full paper submission deadline. Clustering on TMCs as explained in section 3.1 is performed for each time interval t. We employed more efficient method of evolving adjacency matrix than [15] which is capable to capture geographical adjacency as well as dynamic traffic flow. The dynamic adjacency graph will be used in graph convolution neural network to gain information about nodes in different perspective and capture static adjacency among nodes along with its semantic adjacency.

IV. Expected Results

A. Exploratory Analysis of Multimodal Travel Time

In addition to below real-world analysis, we will expand the network size and more benchmark analysis will be presented. Fig. 3 presents the multiple modes of probability distributions of speed for four sample TMCs among a total of 39 TMCs.

B. Benchmarks

Compared against traditional prediction without those updates, the superior performance in prediction uncertainty are presented: both PIR and mixture model (19% increase), PIR only (14%), and TMML ([4] data driven (5%). It presents the temporal transition of the coefficient of variation of multimodal and multivariate for each TMC across 25 time interval.

A Gap function determines the best number of clusters for grouping the similar distributions. Across selected 36 traffic segments, PIR + mixture performs better than TMML and PIR (Fig. 4).

The higher the mean value per unit standard deviation indicates that the mean changes according to the standard deviation and we better use the multivariate data. When speed observations with PIR are close to historic observations, the reduction in uncertainty is higher. A significant reduction in uncertainty of PIR and mixture model indicates more confidence in the predictions.

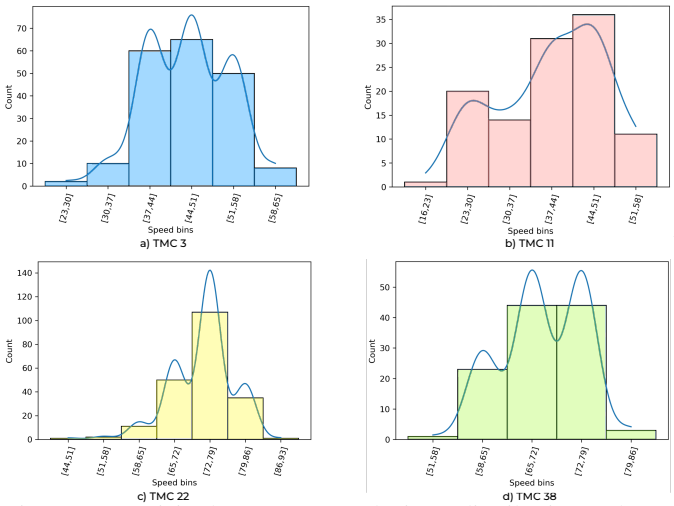

Fig. 3: Surprisingly many travel time distributions shows multimodal - 4 sample TMCs

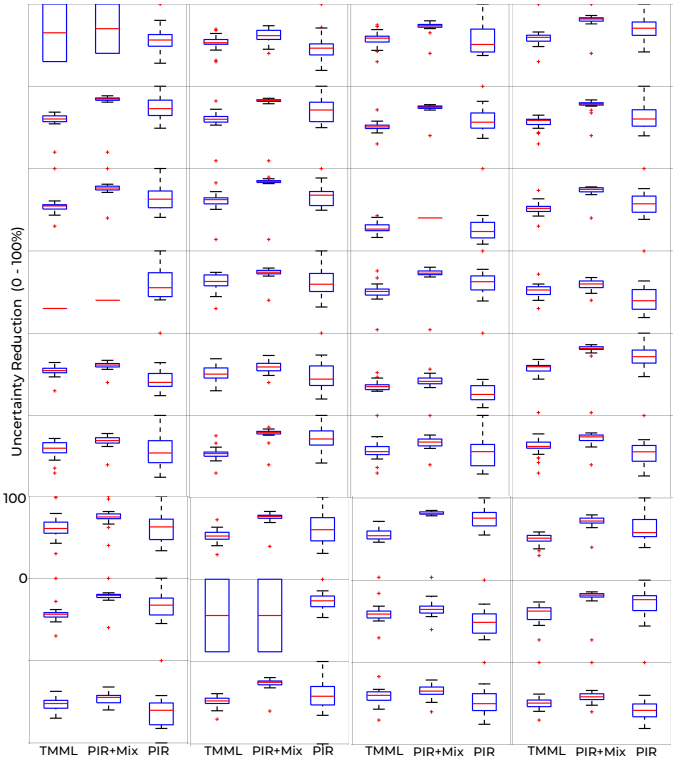

Fig. 4: Benchmarks PIR vs TMML vs PIR + Mixture across 36 TMC segments

V. Conclusion

In this study, the data space is grouped into fine grain cells featuring multimodal and multivariate clusters. Rather than handling individual data points, we analyze which parent distribution those available sample observations belong and evaluate the importance of observations to be used in improving the current prediction. We overcome the limitation of traditional direct (geographically nearby) learning by the transferring online information through indirectly learning of multiple modes of probability distributions and multiple variables across different time stages.

The new family of statistical machine learning models enhanced with traffic theory-driven regularization and cross-entropy based mixture estimation of multimodal and multivariate distribution presents superior performance in reducing travel time prediction. This paper opens appealing research opportunities in the study of information-theoretic decision making that exhibit nontrivial indirect learning from spatiotemporal correlation.

The proposed approach will be useful for traffic and planning agencies knowing how much sample observations they need to improve the traffic prediction capability and plan the future projects. Our tool simply suggests how to use those unused values in the older forecasts, balances the older and recent forecast values based on their importance, and help improving current forecast of traffic value of interest.

ACKNOWLEDGMENT

Authors would like to acknowledge the support from NSF Grant No. 2106989, 2200590, and 1910397 that partially funded the research team.

REFERENCES

[1] F. Guo, H. Rakha, and S. Park, "Multistate model for travel time reliability," *Transportation Research Record*, vol. 2188, no. 1, pp. 46–54, 2010.

[2] F. Zheng and H. V. Zuylen, "Uncertainty and predictability of urban link travel time: Delay distribution–based analysis," *Transportation Research Record*, vol. 2192, no. 1, pp. 136–146, 2010.

[3] L. Folsom, H. Park, and V. Pandey, "Dynamic routing of heterogeneous users after traffic disruptions under a mixed information framework," *Frontiers in Future Transportation*, vol. 3, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/ffutr.2022.851069

[4] H. Park, J. Darko, N. Deshpande, V. Pandey, H. Su, M. Ono, D. Barkely, L. Folsom, D. Posselt, and S. Chien, "Temporal multimodal multivariate learning," in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. [Online]. Available: https://arxiv.org/abs/1903.10304

[5] A. Neupane, V. Pandey, N. Deshpande, and H. Park, "Multimodal learning models for traffic datasets," in *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[6] L. Folsom, M. Ono, K. Otsu, and H. Park, "Scalable information-theoretic path planning for a rover-helicopter team in uncertain environments, 18(2): 1-16," *International Journal of Advanced Robotic Systems*, 2021.

[7] J. Zhang, "Tgcn: Time domain graph convolutional network for multiple objects tracking," 01 2021.

[8] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 07 2018, pp. 3634–3640.

[9] S. Du, T. Li, X. Gong, Y. Yang, and S. J. Horng, "Traffic flow forecasting based on hybrid deep learning framework," in *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2017, pp. 1–6.

[10] X. Kong, J. Zhang, X. Wei, W. Xing, and W. Lu, "Adaptive spatial-temporal graph attention networks for traffic flow forecasting," *Applied Intelligence*, vol. 52, pp. 1–17, 03 2022.

[11] N. Hu, D. Zhang, K. Xie, W. Liang, and M.-Y. Hsieh, "Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting," *Connection Science*, vol. 34, no. 1, pp. 429–448, 2022. [Online]. Available: https://doi.org/10.1080/09540091.2021.2006607

[12] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, The University of Illinois Press, 1949.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[14] M. Das Gupta, S. Srinivasa, J. Madhukara, and M. Antony, "KL divergence based agglomerative clustering for automated vitiligo grading," in *IEEE Computer Vision and Pattern Recognition*, 2015.

[15] F. Li, J. Feng, H. Yan, G. Jin, D. Jin, and Y. Li, "Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution," 2021. [Online]. Available: https://arxiv.org/abs/2104.14917