TRB Annual Meeting

A MULTIMODAL PHYSICS-INFORMED DEEP LEARNING FOR TRAFFIC STATE PREDICTION

--Manuscript Draft--

Full Title:	A MULTIMODAL PHYSICS-INFORMED DEEP LEARNING FOR TRAFFIC STATE PREDICTION
Abstract:	Accurate traffic forecasting is crucial for understanding and managing congestion for efficient transportation planning. Various studies take into consideration the spatiotemporal correlations but fail to account for epistemic uncertainty which arises from incomplete knowledge across different spatiotemporal scales. This study aims to address this issue by capturing unobserved heterogeneity in travel time by considering distinct peaks in the probability density function, which we refer to as multimodal probability distribution, while establishing causation through physics-based principles. The information obtained from this methodology is then employed in a new model called the Physics Informed-Graph Convolutional Gated Recurrent Neural Network (PI-GRNN). This deep learning model utilizes the inherent structure and relationships within the transportation network for capturing sequential patterns and dependencies in the data over time. The dynamic graph-based approach can leverage data from different locations and times to improve future travel time predictions at distant non-contiguous unobserved locations. We first designed data-driven Kalman Filtering model to demonstrate the significant benefit of removing epistemic uncertainty and extended the model to deep learning, which employs a weighted adjacency matrix to integrate non-contiguous correlations for each time interval. The clustered uncertainty profile based on these weights effectively reduces the uncertainty in prediction. To the best of our knowledge, this is the first method to employ a multimodal and multivariate data-based approach to create a dynamic graph. We compared our method with other benchmark models using a real-world freeway traffic dataset of one-year duration. Extensive experiments demonstrate that our model consistently outperforms the five baselines.
Manuscript Classifications:	Data and Data Science; Artificial Intelligence and Advanced Computing Applications AED50; Artificial Intelligence; Bayesian analysis; Deep Learning; General; Machine Learning (Artificial Intelligence); Neural Networks; Operations; Highway Traffic Monitoring ACP70; Travel Time
Manuscript Number:	TRBAM-24-05498
Article Type:	Presentation and Publication
Order of Authors:	Niharika Deshpande
	Hyoshin Park
	Venktesh Pandey
	Justice Darko
Additional Information:	
Question	**Response**
The total word count limit is 7500 words including tables. Each table equals 250 words and must be included in your count. Papers exceeding the word limit may be rejected. My word count is:	7486
Is your submission in response to a Call for Papers? (This is not required and will not affect your likelihood of publication.)	No

A MULTIMODAL PHYSICS-INFORMED DEEP LEARNING FOR TRAFFIC STATE PREDICTION

Niharika Deshpande
PhD Student
Department of Engineering Management & Systems Engineering
Old Dominion University, 2101 Engineering Systems Building, Norfolk, VA 23529
Email: ndesh002@odu.edu

Hyoshin (John) Park, Ph.D, Corresponding Author
Associate Professor
Department of Engineering Management & Systems Engineering
Old Dominion University, 2101 Engineering Systems Building, Norfolk, VA 23529
Email: h1park@odu.edu

Venktesh Pandey, Ph.D
Assistant Professor
Department of Civil, Architectural, & Environmental Engineering
North Carolina A&T State University, Greensboro, NC 27411
Email: vpandey@ncat.edu

Justice Darko, Ph.D.
Senior Data Scientist
Rockwell Automation
Email: justicedarko3@gmail.com

Word Count: 7236 words + 1 table(s) $\times$ 250 = 7486 words

Submission Date: August 1, 2023

ABSTRACT

Accurate traffic forecasting is crucial for understanding and managing congestion for efficient transportation planning. Various studies take into consideration the spatiotemporal correlations but fail to account for epistemic uncertainty which arises from incomplete knowledge across different spatiotemporal scales. This study aims to address this issue by capturing unobserved heterogeneity in travel time by considering distinct peaks in the probability density function, which we refer to as multimodal probability distribution, while establishing causation through physics-based principles. The information obtained from this methodology is then employed in a new model called the Physics Informed-Graph Convolutional Gated Recurrent Neural Network (PI-GRNN). This deep learning model utilizes the inherent structure and relationships within the transportation network for capturing sequential patterns and dependencies in the data over time. The dynamic graph-based approach can leverage data from different locations and times to improve future travel time predictions at distant non-contiguous unobserved locations. We first designed data-driven Kalman Filtering model to demonstrate the significant benefit of removing epistemic uncertainty and extended the model to deep learning, which employs a weighted adjacency matrix to integrate non-contiguous correlations for each time interval. The clustered uncertainty profile based on these weights effectively reduces the uncertainty in prediction. To the best of our knowledge, this is the first method to employ a multimodal and multivariate data-based approach to create a dynamic graph. We compared our method with other benchmark models using a real-world freeway traffic dataset of one-year duration. Extensive experiments demonstrate that our model consistently outperforms the five baselines.

INTRODUCTION

Traffic congestion in metropolitan regions, has become a pressing issue that negatively impacts the quality of life, economic productivity, and the reliability of information. The emissions from vehicles contribute to air pollution and climate change, affecting our health and the environment. With the rapid development of urbanization, transportation systems in cities are under great pressure due to growing populations and increased vehicles. Fortunately, advances in data intelligence and urban computing have made it possible to collect massive amounts of traffic data. These data serve as essential indicators reflecting the state of the transportation system and play a crucial role in predicting future traffic conditions.

Traditional models such as Historical Average (HA), Auto-Regressive Integrated Moving Average (ARIMA), Vector Auto-Regression (VAR) etc. have focused on analyzing historical data of a single traffic variable to predict its future values. While these univariate methods have proven useful in providing valuable insights, they suffer from limitations. Some of the limitations are they assume stationarity for time series data and their incapability to capture spatiotemporal dependencies, which can be crucial for accurate predictions, especially in urban transportation systems. To address these limitations and improve traffic prediction accuracy, more advanced techniques such as machine learning algorithms and deep learning models have been developed.

Kalman filtering (KF) model can be beneficial in traffic prediction because of its ability to handle a series of noisy data and adapt to dynamic traffic conditions. However, the KF primarily relies on recent measurements and does not consider long-term historical patterns. In traffic prediction, historical data can be valuable in understanding traffic behavior over time, and the KF's lack of consideration for longer-term trends can limit its effectiveness. Another challenge posed by KF is that its performance heavily relies on the quality and frequency of sensor measurements. Inaccurate or sparse data from traffic sensors can lead to poor state estimates, especially during congested or highly dynamic traffic conditions. In order to overcome both of these limitations, this study utilizes spatiotemporal correlations using multimodal multivariate learning to effectively capture the unobserved heterogeneity. These correlations would help to gain information from other correlated links in case of missing data and build robust model.

Although the new data-driven KF model can demonstrate the significant benefit of removing epistemic uncertainty, KF's linear and Gaussian assumptions may not fully capture the complexity of real-world traffic dynamics. In traffic prediction, traffic dynamics are often nonlinear, and traffic flow patterns can exhibit non-Gaussian behaviors due to various factors like congestion, incidents, or sudden changes in driving behavior. Designing an accurate linear model that can capture all these complexities can be challenging. In practice, linear models may be too simplistic to represent the true traffic dynamics, leading to suboptimal predictions. To address these limitations and improve the accuracy of traffic prediction, this study further extends the KF model to neural network approach by better handling non-linearities and data sparsity, and capturing more complex patterns in traffic dynamics.

Studies have exploited the advantage offered by Graph Neural Networks which can be effective in non-Euclidean topological space like traffic network. Despite their success, there are still some unexplored aspects of GNNs that offer exciting research opportunities. Most existing GNNs assume static graphs, where the graph structure remains unchanged during training. However, many real-world applications like traffic involve dynamic graphs, where the graph evolves over time or with changing interactions. This dynamic nature of traffic is unable to be captured by static pre-defined based on topology alone. Hence, this study uses graphs evolving with time based

on spatiotemporal correlations that can account for constantly changing conditions, interactions, and relationships between road segments. Even with considering dynamic graphs in GNN, it is challenging to model long-term temporal dependencies. Thus, considering dedicated techniques in the algorithm to capture them is crucial for accuracy of model. Another challenge that arises while determining the correlations is the problem of "false/coincidental correlations" which are introduced while analyzing large amount of data. If we do not address these correlations, it can lead to erroneous conclusions and inaccurate predictions. In order to avoid that, we have introduced the physics-regularized method. The main contributions of our paper are summarized as follows:

- We first design data-driven Kalman Filtering model to demonstrate the significant benefit of removing epistemic uncertainty and extend the model to deep learning, which employs a weighted adjacency matrix to integrate non-contiguous correlations for each time interval.
- KF model utilizes spatiotemporal correlations to estimate information gain in order to decrease the uncertainty of the unobserved road segments with sparse data and ensure accurate traffic prediction.
- New deep learning, named as Physics Informed-Graph Recurrent Neural Network (PI-GRNN), leverages the graph structure to learn meaningful representations that encode both node attributes and their interactions within the graph. PI-GRNN captures the temporal dynamics of graph data to learn the graph evolution over time. This algorithm can enhance the information aggregation process by considering both the current state of a node and its historical context enabling more effective information fusion and propagation across the graph.
- The flexibility of dynamic adjacency matrix allows us to capture both short-term and long-term dynamics of spatial correlations, making it more robust and accurate in predicting future patterns. Exploratory analysis followed by benchmark shows the effectiveness of the unique integration of the dynamic adjacency matrix to deep learning model.

We also contribute to transportation data management community by introducing new perspective in multimodal and multivariate learning as follows.

- "Multimodal" refers to a probability distribution with multiple modes, which allows us to capture unseen traffic patters. "Multivariate" refers to the inter-dependencies and interactions between various traffic variables explained by the fundamental diagram of traffic flow theory. Integrating both aspects of learning prevents the model from attributing spurious correlations to the data, where certain relationships may appear significant but are actually coincidental. This approach helps in regularizing the model and enhancing its ability to capture the true underlying patterns and relationships in the traffic data, ultimately leading to improved predictions and more effective traffic management strategies.
- We consider the probability distribution function (PDF) instead of average values to map spatiotemporal correlation. Calculating the distance between PDFs as an accurate measure of similarity between non-contiguous locations helps in keeping the information from multiple modes and hence helps in increasing the prediction accuracy of the model.

The rest of the paper is organized as follows: Section 2 reviews the literature related to multimodal-multivariate learning, Kalman Filter and deep learning models. Section 3 discusses the methodology for KF model. Section 4 presents methodology using neural network. Section 5 explore sample adjacency matrices through visualization. After that, we evaluate performance of the proposed deep learning model against corresponding benchmarks in Section 6.

LITERATURE REVIEW

Without knowing the future traffic with confidence, the traditional choice theory considers bounded rationality (*1*) of the majority of travelers taking a detour, which causes more congestion on nearby roads. Existing optimization problems with a long horizon commonly simplify the traffic states to unimodal to handle the curse of dimensionality. Commonly used Gaussian processes cannot incorporate the complex prior traffic knowledge into transition dynamics (*2*). Recent mixture density networks in approximating multimodal output distribution (*3*) have well-handled prediction uncertainties rather than averaging the distribution. While those advanced multimodal learning helped the prescriptive analytics make proactive decisions through accurate prediction of future events, sequential learning of those approximated information has depended on unimodal probability distribution. In this study, a new information theory overcomes the traditional entropy approach by actively sensing and learning information in a sequence.

Park et al. (*4*) developed a data-driven model which used location's observed data to forecast conditions at distant non-contiguous locations' unobserved data, followed by the uncertainty reduction through processing bimodal distribution and transferring information from one traveler to another traveler (*5*). But they haven't addressed the coincidental correlations introduced due to pure data-driven approach. This research addresses this issue by introducing physics-driven approach which can help to establish the causality. This approach uses multivariate traffic data to remove accidental correlations.

Various studies use KF with other techniques to enhance the accuracy and robustness of traffic state predictions (*6*). Usually real-time measurements are used in the KF models in the correction step (*7*). But they fail to account for historical data and hence cannot capture valuable insights into traffic patterns and trends. These models consider evolution of traffic based on only neighbouring segments (*8*). Unexplored states beyond neighbouring segments in space and time dimension can negatively affect the prediction accuracy of the model (*9*). Therefore, we deal with it by establishing spatio-temporal correlations among non-contiguous locations. Traditional KF uses only numerical value of recent observation but we customized the algorithm to use the derived PDF which helps in incorporating comprehensive information. We developed a novel KF model by incorporating these improvements which outperformed the benchmark KF models.

In the realm of deep learning, temporal correlations in data can be effectively captured using various techniques, including Recurrent Neural Networks (RNN) (*10*), Convolutional Neural Networks (CNN), and attention mechanisms. RNNs are well-suited for sequential data, such as time series, as they can retain information from previous time steps to capture temporal dependencies. On the other hand, spatial dependencies in data can be modeled using CNNs, Graph Neural Networks (GNN) (*10*), or attention mechanisms. CNNs are applicable for spatial data that adheres to a regular grid structure. By considering the spatial relationships encoded in the grid-based graph, CNNs can effectively capture spatial dependencies and patterns present in the data but they fail when the data is irregular. GNNs are designed to work with graph-structured data, making them suitable for modeling relationships in non-grid-like structures. They have shown great promise in various forecasting applications, especially when dealing with data structured as graphs or networks. Usually GNNs use fixed graph based on topology as an input in traffic prediction models (*11*). It does not account for dynamic changes and temporal dependencies in the data. Hence it is crucial to consider evolving graphs which can consider the changing relationship between nodes with time. In time series forecasting tasks, temporal dependency can be established using techniques like GRU and LSTM (*12*). A few deep learning algorithms incorporate physics

related aspect in the cost function of the model (*13*). This study integrates spatiotemporal correlations into the deep learning framework with GNN and GRU resulting in an innovative mixture algorithm. To the best of our knowledge, this is the first method to employ a multimodal and multivariate data-based approach to create a dynamic graph.

DATA-DRIVEN AND PHYSICS INFORMED KALMAN FILTER

Data-Driven Temporal Multimodal Multivariate Learning

We start extending standard deviation-based information theory (*5*) to ensure that locations with broad bimodal probability distributions are targeted over locations with narrow probability distributions. "Correlated cells" are defined as cells with a similar travel time probability distribution. The states of correlated cells are probabilistic until one of them is visited and the true state is observed. 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 (PDF). The observations from correlated links are used to reduce the entropy of PDF.

In the proposed entropy-based travel time prediction, information is shared with other cells, influencing their route choices. The path is planned in advance and updated as information about the grid is discovered. As travelers discover the state of the grid, that information is conveyed to the other travelers. Each traveler updates its path plan every time it moves to a new grid cell. By sharing information about the state of the grid cells, each traveler helps to define the optimal parameters to be used in the other traveler's utility functions. 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.

Figure 1 shows the benefits of the proposed data-driven learning. Assume we know that a highway connection $\mathbb{A}$ normally takes two minutes to travel without traffic, but it could take eight minutes owing to an unforeseen event (e.g., incidents). The literature treats links $\mathbb{A}, \mathbb{A}', \mathbb{B}, \mathbb{C}$ as a unimodal probability distribution with an expected travel time. Without knowing the future traffic with confidence, the traditional choice theory considers the bounded rationality (*1, 14–16*) of the majority of agents taking a detour to link $\mathbb{B}$, which causes congestion on $\mathbb{B}$ and nearby roads.

If the bimodal trip distributions for both links are similar, we can group A and A' together in the same correlated group. The literature overlooks three advantages of deploying a platoon of vehicles to A rather than B: 1) We can update the estimated travel time on this link A so other drivers can modify either their departure time or route to utilize this 2-minute shortcut, in the case of a scenario that turned out to be 2-minutes due to the quick clearance of the event. 2) We can update travel time on other links with similar probability distributions (e.g., A'). We can send extra vehicles to this route and relieve other route congestion that turned out to be 8 minutes due to the extended clearance time of the incident if we know the overall travel time of the route is 4-minutes. 3) We update travel time on other links with the same sort of probability distributions (e.g., A'). By knowing that the total travel time of a route AA' is 16-minutes, we can notify fewer vehicles to use this route, and redistribute traffic to other routes (i.e., BC) having shorter travel times.

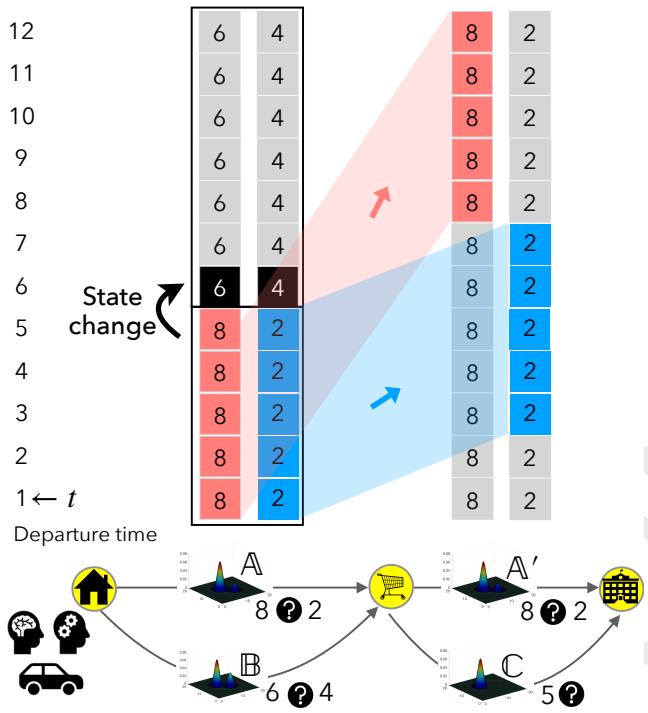


FIGURE 1: Temporal Multimodal Learning considering correlation between time-varying bimodal link distributions. While the existing literature only considers nearby links for explaining time-dependent transition of traffic, no research has been done on the realization of multimodal travel time distributions based on real-world data.

Kalman Filtering with Physics-Informed Regularization

The distinguishable aspect of the physics-informed and -regularized (PIR) model in the hierarchical update steps is the use of new information obtained from Temporal Multimodal Multivariate Learning (Figure 2). We first predict the chosen state variable at the next time interval $t+1$ using the measurement from the previous time interval t. In the update step, the predicted state is corrected using the noisy measurements at $t+1$. Clustering identifies similar travel time distributions. The global correlation between non-contiguous cells of an entire map are estimated by using Expectation Maximization. The optimal distribution of the data over K clusters are determined by maximizing the lower bound of the log of the likelihood. We decouple the spurious correlations first and then use the entropy method to estimate the mixture of multimodal and multivariate distributions. Since, the mixture is PDF with reduced entropy, providing an accurately estimated distribution rather than just mean and standard deviation, will increase the accuracy of updating the error covariance matrix.

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 $K_t = P_t^- H^T (HP_t^- H^T + R)^{-1}$ where H is the connection matrix between the state vector and the measurement vector and R & Q are Gaussian noise vectors. Z_t is the observations used to correct the predicted estimate. Observations considered for KF model with Physics Informed Regularization (PIR) is different than that for no-PIR. Z_t for KF-no PIR are the

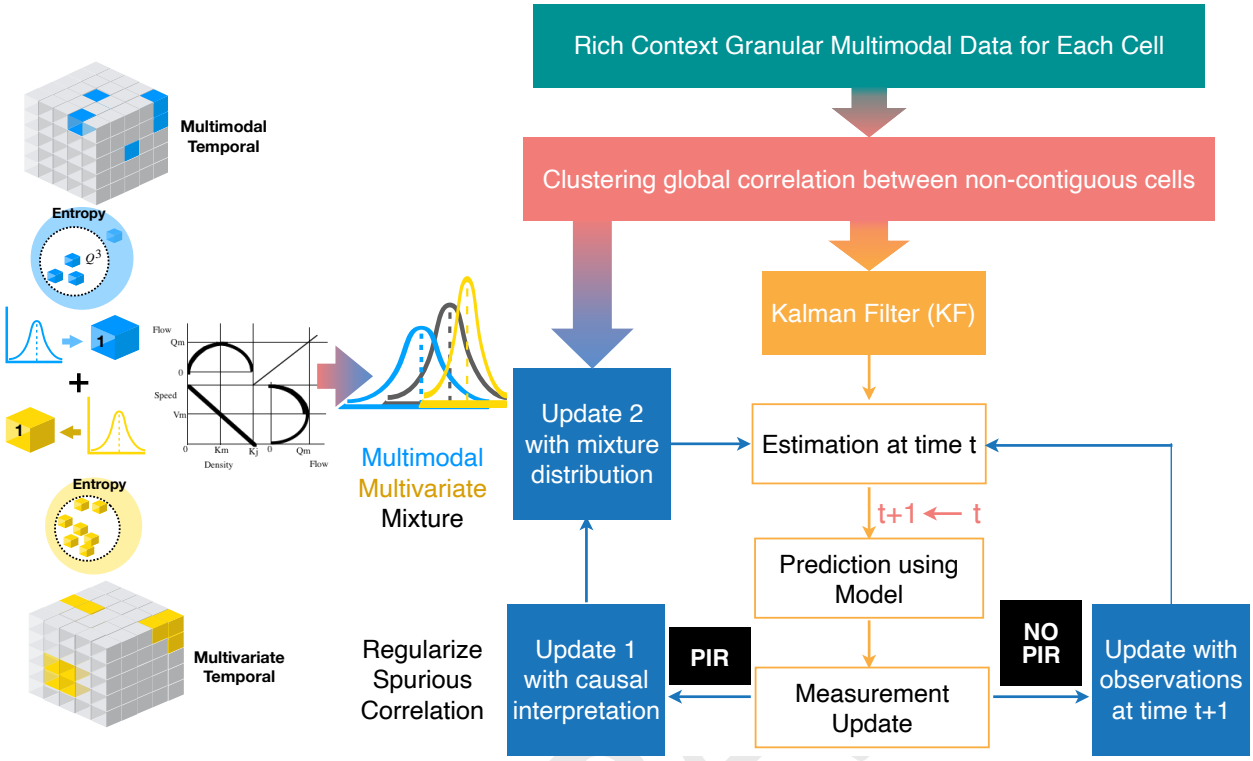


FIGURE 2: Physics-informed and -regularized (PIR) KF in the hierarchical update steps

speed observations on a given day while, in case of KF-PIR, Z_t are the observations drawn through two steps as described in Figure 2.

In the first update step of KF-PIR, common correlations are identified from multimodal and multivariate clusters at the same time and location. This will help in removing spurious correlations. Once the conflicting observations are resolved, in the second step, a mixture of two distributions obtained from two different sensors is calculated by employing cross-entropy method. In this method, entropy is assumed to be the measure of uncertainty. More entropy means less weight in the mixture distribution. Once this mixture distribution is determined, the data gained from common correlations is incorporated into the mixture PDF to lower its entropy (uncertainty) and finally it is used in the correction of prediction in KF model.

MULTIMODAL PHYSICS-INFORMED DEEP LEARNING

GNNs provide various advantages in the task of traffic state prediction due to their ability to model spatial dependencies for data from irregular topology. The sensors used for gathering data are not necessarily equally spaced and hence can have graphs with varying sizes. GNNs exhibit the ability to handle such complex graph structure making it more suitable to the real-world traffic networks compared to CNN which can primarily accept fixed-size inputs only. This study leverages these benefits of GNNs. The studies like (*10*) use static adjacency matrix in GNN which can only capture spatial correlations based on geometry. This study introduces novel method of calculating semantic adjacency between nodes based on time of the day. This helps to capture the temporal dependency along with enhanced spatial dependencies which can significantly improve the predictive capabilities of the model.

Framework for Graph Convolution Gated Recurrent Network

Traffic prediction problem can be formulated as time-series forecasting problem with historical data and prior knowledge. The prior knowledge used in Graph Neural Network (GNN) is pre-defined adjacency graph $\mathscr{G} = (\mathscr{V}, \mathscr{E}, A)$. Here, $\mathscr{V}$ is a set of nodes which represent different locations (e.g., road segments) on the road network; $\mathscr{E}$ is a set of edges and $A \in \mathbb{R}^{N \times N}$ is the adjacency matrix.

Given the graph $\mathscr{G} = (\mathscr{V}, \mathscr{E}, A)$ and its observed P step graph signals $\mathbf{X}_{(t-P):t}$, to learn a function f which is able to map $\mathbf{X}_{(t-P):t}$ and $\mathscr{G}$ to next Q step graph signals $\hat{\mathbf{X}}_{t:(t+Q)}$, represented as follows:

$$\left[\mathbf{X}_{(t-P):t}, \mathscr{G}\right] \xrightarrow{f} \hat{\mathbf{X}}_{t:(t+Q)},$$

where $\mathbf{X}_{(t-P):t} = (\mathbf{X}_{t-P}, \mathbf{X}_{t-P+1}, \ldots, \mathbf{X}_{t-1}) \in \mathbf{R}^{P \times N \times D}$, D is the number of features of each node (e.g., traffic volume, traffic speed, etc.) and $\hat{\mathbf{X}}_{t:(t+Q)} = \left(\hat{\mathbf{X}}_t, \hat{\mathbf{X}}_{t+1}, \ldots, \hat{\mathbf{X}}_{t+Q-1}\right) \in \mathbf{R}^{Q \times N \times D}$

As shown in Figure 3a, the Recurrent Neural Networks (RNN) are used in case of sequential data as it retains the previous states in memory while accepting current state. Therefore, it becomes a suitable means to solve time series predictions. However, RNNs are capable to capture only short-term temporal dependencies and has the issue of vanishing gradient. These limitations of RNN can overcome by Long Short Term Memory (LSTM) (*17*) and Gated Recurrent Unit (GRU) (*18*). GRU has less complex structure than LSTM as it has less number of gates, it is easy to modify and faster to train. Therefore, we choose GRU for extracting temporal correlations from traffic time series data. We replaced the matrix multiplications in GRU with Graph Convolution (GC) module and are described using following equations.

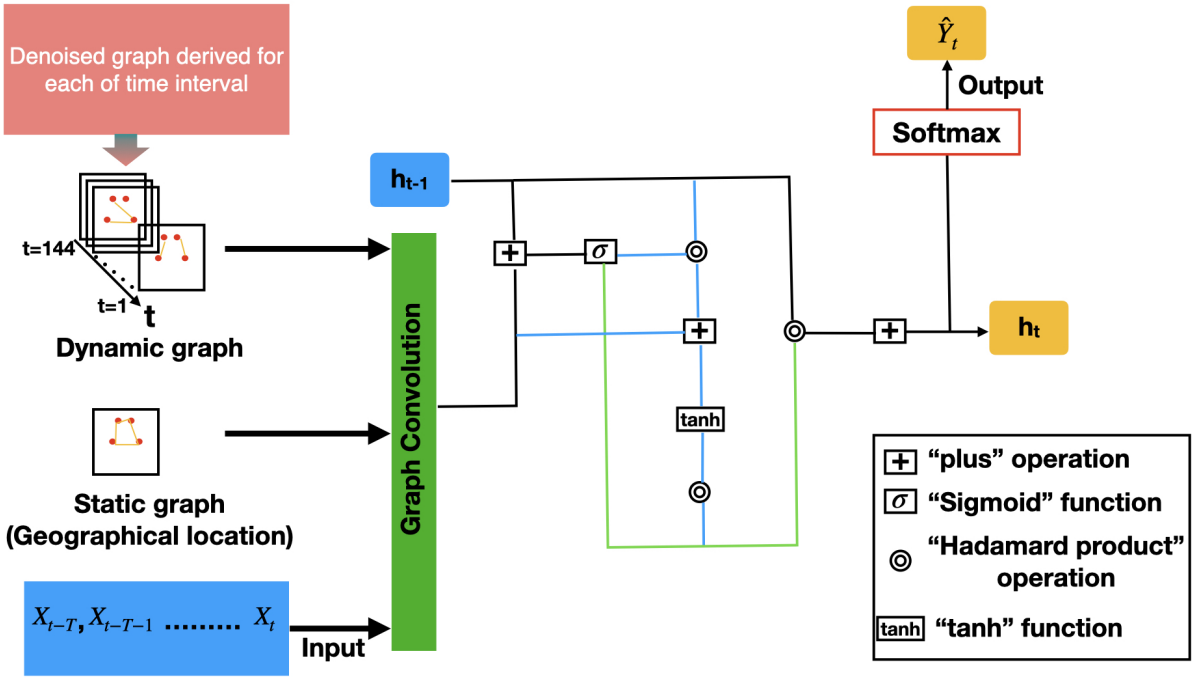
$$\begin{aligned} z_t &= \sigma(W_u.[GC(DA_t, A), h_{t-1}] + b_u) \\ r_t &= \sigma(W_r.[GC(DA_t, A), h_{t-1}] + b_r) \\ c_t &= tanh(W_c.[GC(DA_t, A), (r_t * h_{t-1})] + b_c) \\ h_t &= z_t * h_{t-1} + (1 - z_t) * c_t \end{aligned} \tag{1}$$

Where $\sigma(.)$ and $tanh(.)$ are the sigmoid functions, W and b are the weights and biases in the training, respectively. * represents the matrix multiplication. DA_t denotes dynamic adjacency graph at time interval t and A represents pre-defined adjacency graph based on geographical locations.

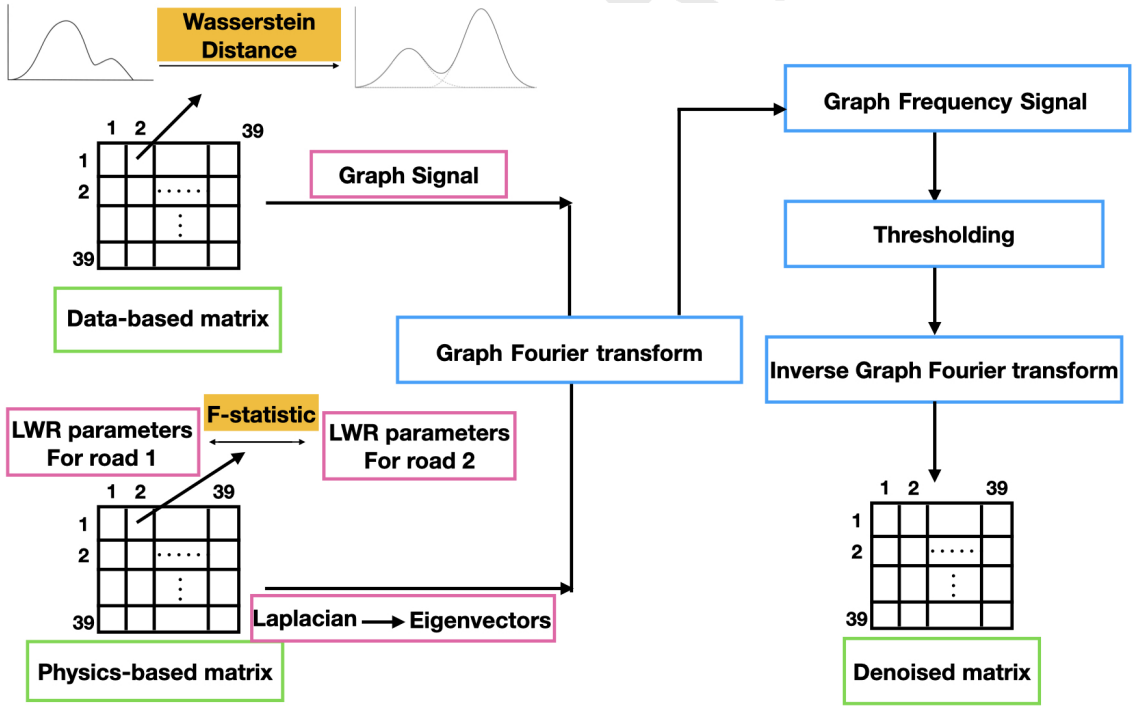
EXPLORATORY ANALYSIS OF DYNAMIC ADJACENCY MATRIX

We employed Graph Signal Denoising method to correct spurious correlations in a weighted adjacency matrix based on pure data using a weighted graph based on physics. This method involves treating the pure data weighted adjacency matrix as a graph signal, where each weight represents the correlation between two road segments. This graph signal represents the noisy or spurious correlations derived from the pure data. The overview of deriving denoised adjacency graph is depicted in Figure 3b.

For exploratory analysis and developed models were tested on the 18.8 mile stretch of North Carolina Triangle Expressway Figure 4 with probe vehicles and loop detectors on 39 TMCs collected by NPRMDS. The speed data is collected from probe vehicles from January 1, 2021 to December 31, 2021. The collected data is averaged for 10 minutes of time interval. The 24 hours in a day are divided into 10 minutes giving us 144 time intervals. The data is sorted within



(a) Overview of Graph Convolution Gated Recurrent Neural Network



(b) Flowchart of deriving dynamic adjacency matrix for each time interval

FIGURE 3: The architecture of PI-GRNN

FIGURE 4: Case study for the network of Triangle Expressway: The interpolation method used in Park and Haghani (*19*) provide fine-grain layer of speed, density, and flow data on those 39 TMC segments (cells) from January 1, 2021, to December 31, 2021. The data is averaged over a 10-minute interval, which divides a day's 24 hours into 10 minutes.

those intervals. This data is used in the model to determine data-based adjacency matrix. The multivariate data is collected from loop detectors. The loop sensors detect speed and density at the installed location. This data is used to derive adjacency matrix based on interrelations between variables.

Adjacency matrix with multimodal data

The historical data for speed v is collected from Traffic message channel (TMC) for the period of a year. The data is collected in the interval of 10 minutes. The data is sorted for each TMC during 144 time intervals of 10 minutes within 24 hours. The sorted data is then categorized into 14 speed bins ranging from 2 mph to 100 mph with the difference of 7 mph for each TMC. The histogram is derived from it and then the probability distribution function (PDF) of speed for each TMC for each time interval. The statistical test is established to identify that PDFs are multimodal meaning has more than one maxima. The similarity between PDFs needs to be established in order to understand the semantic adjacency between TMCs. The different distance parameters like KL divergence, Jensen Shannon entropy, Hellinger distance, Wasserstein distance etc. are considered. Among these parameters, most suitable one for measuring similarity between multimodal distribution was found to be Wasserstein distance. Earth mover's distance (EMD) or Wasserstein distance measures the minimum cost required to transform one distribution into another, considering the transportation of mass from one mode to another. As it accounts for spatial arrangement of the probability mass of the mode while calculating cost, it can capture differences in shape, location, and spread between modes and hence, work well with multimodal distribution. This distance parameter preserves the distributional information of the data and hence has the ability to capture complex structure of multimodal distribution. The Wasserstein distance is robust to outliers as it considers the overall mass transportation and does not heavily rely on the exact values of individual data points. It does not impose specific assumptions or constraints on the shape or type of the distributions being compared. This flexibility allows the Wasserstein distance to be used for comparing multimodal distributions that can exhibit various forms and structures. All these advantages make wasserstein distance most suitable parameter to measure similarity between calculated speed distributions. In this study, we have established the similarity between TMCs for each 144 time intervals within 24 hours. Based on 39 TMCs within the area of study, the 39×39 matrix is

established for each interval by filling values with Wasserstein distance using following equation.

$$W_1(P,Q) = \min_{\gamma \in \Gamma(P,Q)} \sum_{i,j} \gamma_{ij} \cdot d(x_i, y_j) \tag{2}$$

where:

- $W_1(P,Q)$ represents the Wasserstein distance between distributions P and Q. $W_1(P,Q) \in [0, \text{inf})$.
- Γ is a transportation plan that defines the amount of mass to be moved from each point in P to each point in Q. It satisfies the constraints of being a valid joint distribution with marginals P and Q, denoted as $\gamma \in \Gamma(P,Q)$.
- x_i and y_j represent individual points (samples) from distributions P and Q, respectively.
- $d(x_i, y_j)$ is a distance metric (e.g., Euclidean distance or any other suitable distance measure) between x_i and y_j.

We assume that as the Wasserstein distance between the two PDFs is large, they exhibit lesser correlation. A Wasserstein distance of 0 indicates that the two distributions being compared are identical. As the distributions become more dissimilar, the Wasserstein distance increases. The distance matrix is normalized between values of 0 and 1 using following formula.

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3}$$

where:

- x_i is value in matrix 39×39 to be normalize.
- x_{min} is minimum value in matrix.
- x_{max} is maximum value in matrix.

Following equation is used to generate weighted adjacency matrix.

$$(x_{weighted})_i = 1 - z_i \tag{4}$$

The above procedure is performed on all 39 TMCs and for all 144 time intervals to establish the weighted adjacency matrix of dimension 39×39 for each interval. Figure 5a and Figure 5b shows the weighted adjacency matrix for TMC-1 during time intervals 8-8:10 am and 8-8:10 pm, respectively. It can be observed from the figures that for two different time intervals, the semantic adjacency is different for the same TMC. Therefore, when we consider adjacency entirely based on geographical proximity, we are missing out on the adjacency exhibited by dynamic nature of traffic. Hence, the method of dynamic adjacency matrix proves to be superior in exploring wide-range of spatial correlations.

Adjacency matrix with multivariate data by traffic flow theory

Adjacency matrix is also determined based on the physics of traffic flow. The LWR model (Lighthill-Whitham-Richards model) is a fundamental traffic flow model that describes the evolution of traffic density along a roadway. It is a macroscopic model that represents traffic flow based on the conservation of vehicles and the fundamental relationship between traffic density, flow rate, and speed. The LWR model is based on following two assumptions.

1. Conservation of Vehicles: The total number of vehicles on the road remains constant over time. Vehicles cannot appear or disappear along the roadway.
2. Fundamental Diagram: It assumes a fundamental relationship between traffic density, flow rate, and speed. This relationship is typically represented as a triangular fundamen-

tal diagram, where the flow rate is a function of traffic density and speed.

Mathematically, the LWR model can be expressed using the following equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{5}$$

where:

- ρ represents the traffic density (number of vehicles per unit length of the road).
- t is time.
- x is the spatial coordinate along the road.
- q is the traffic flow rate (number of vehicles per unit time).

The LWR is a traffic flow model that describes the evolution of traffic density and speed over time and space. If the parameters of LWR flow model are similar for two TMCs (road segments), it signifies that the relationship between traffic density and speed is comparable for both TMCs. It suggests that drivers on these road segments experience similar congestion patterns, speed variations, flow characteristics and traffic flow capacities. Similar LWR parameters may indicate that congestion propagation between the two road segments is likely to be similar. Congestion on one segment may impact the traffic conditions on the other segment in a comparable manner. Hence, in this study, we compared the parameters for all TMCs using speed and density data collected over a period of a year using loop detectors.

This study employs method of characteristics to obtain LWR parameters by solving differential equation backwards in time from assumed set of initial conditions using the collected data of speed and density. The parameters that need to be estimated include the fundamental diagram parameters (the free-flow speed, the jam density, and the critical density), as well as the traffic demand and supply parameters. The initial conditions i.e the initial density and speed are assumed to be from the beginning of the collected data. The LWR equations are traced back in time from the final time and space measurements to the initial conditions. Then the values of parameters are determined using least squares optimization method.

Once the LWR parameters are estimated for each of 39 TMCs, the sum of squared residuals (SSR) is calculated using following equation. It represents the overall deviation of the LWR model predictions from the observed speed data.

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{6}$$

where:

- SSR is the sum of squared residuals
- y_i is the observed speed value at data point i.
- $\hat{y}_i$ is the predicted speed value at data point i based on the LWR model.

Then F-statistic is computed using the following formula.

$$F = \frac{(SSR1 - SSR2)/(k2 - k1)}{SSR2/(n - k2)} \tag{7}$$

where:

- SSR1 and SSR2 are the sum of squared residuals for road segment 1 and road segment 2, respectively.
- k1 and k2 are the degrees of freedom for road segment 1 and road segment 2, respec-

tively. In the case of LWR models, the degrees of freedom are typically 5 (number of parameters) minus the number of data points used for estimation.

- n is the total number of data points used for estimation (across both road segments).

Calculated F-values are used as weights in the adjacency matrix to reflect the strength or significance of the connection between the road segments. The weight represents the strength of the connection between the two road segments. Using F-values as weights in the adjacency matrix allows us to incorporate the significance of the parameter comparisons into the analysis of the road segment connections. It provides a quantitative measure of the relationship strength between the segments based on the comparison of their LWR parameters. In order for values in adjacency matrix to fall between the range [0,1], the values are normalized using equation 15 described above. The adjacency matrix of 39×39 derived from multivariate relation between speed and density based on physics of traffic flow called LWR theory is shown in the Figure 5c.

The weights in the adjacency matrix determined from only data are corrected using the strength of connection between two roads established using comparison of LWR parameters. It helps to rectify the spurious correlations that are introduced coincidentally while analyzing large amount data without any causal relationship between them. Hence, it is crucial to study them by considering the underlying cause using traffic flow theory.

The Graph Fourier Transform is computed using weighted adjacency graph based on physics. The Graph Fourier Transform (GFT) is a mathematical operation that transforms a graph signal from the vertex domain to the graph frequency domain. It is analogous to the Fourier Transform in signal processing, but it operates on graph signals defined on the vertices of a graph instead of continuous or discrete time signals. The GFT can be computed using the eigenvectors of the Laplacian matrix of the graph. The Laplacian can be constructed using following formula:

$$\begin{aligned} L &= D - A \\ D_{ii} &= \sum_j A_{ij} \end{aligned} \tag{8}$$

where:

- D is a degree matrix.
- i represents the row and column index.
- A is adjacency matrix.

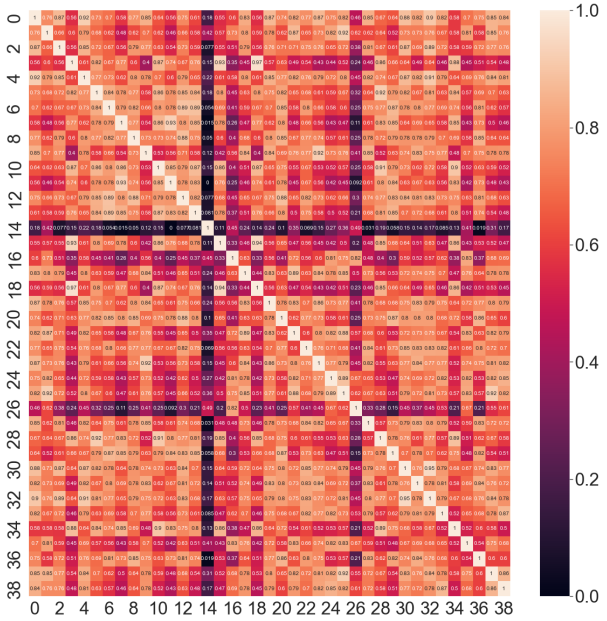
Then, eigenvalues and eigenvectors of matrix L are computed using following equation.

$$L \cdot U = U \cdot \Lambda \tag{9}$$

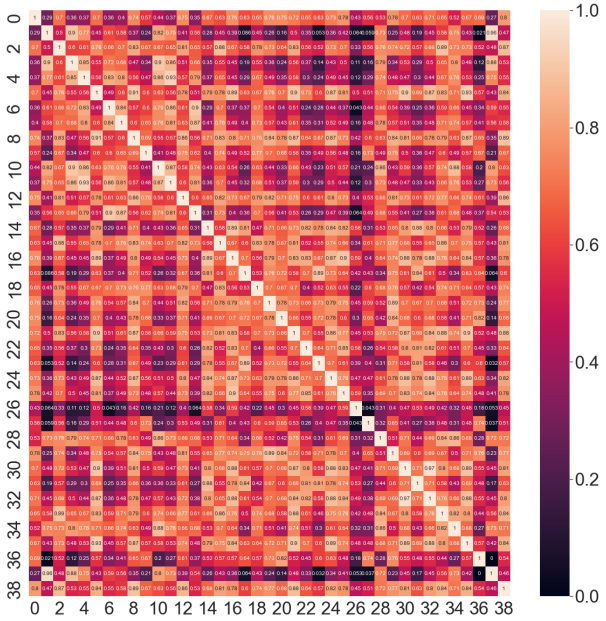
where U is a matrix whose columns are the eigenvectors, and Λ is a diagonal matrix containing the eigenvalues. Then, the graph frequency signal is computed using Laplacian and graph signal from adjacency matrix based on pure data.

$$S = U^T \cdot x \tag{10}$$

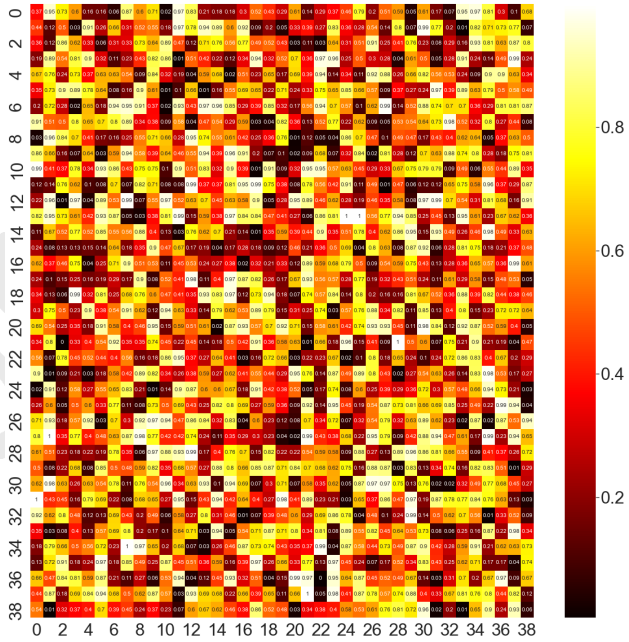
where S is the graph frequency signal, U^T is the transpose of the eigenvector matrix, and x is the original graph signal. The resulting graph frequency signal S represents the signal in the graph frequency domain. The entries of S correspond to the contributions of different graph frequencies to the original signal x. After that thresholding is used for denoising the graph signal S. It is used to determine which graph frequencies are considered significant and which ones are set to zero. The threshold value depends on the specific application and the characteristics of the graph signal. It



(a) Weighted adjacency matrix based on multimodal PDFs for TMC-1 at time interval 8-8:10 am



(b) Weighted adjacency matrix based on multimodal PDFs for TMC-1 at time interval 8-8:10 pm



(c) Weighted adjacency matrix based on multivariate relationship using LWR traffic flow theory

FIGURE 5: Adjacency matrices from data-driven and physics driven approach

determines the level of noise removal or sparsity in the denoised graph signal. A higher threshold will result in a sparser denoised signal, removing more noise but potentially discarding some valid signal components. On the other hand, a lower threshold will preserve more signal components but may also retain more noise. After experimentation and understanding of the characteristics of the graph signal and the noise present in the data, the threshold value is decided to be 0.01. After thresholding, in order to transform the graph frequency signal S back in the vertex domain, the Inverse Graph Fourier Transform (IGFT) is performed.

$$x_{denoised} = U \cdot S \tag{11}$$

where $x_{denoised}$ is the reconstructed graph signal in the vertex domain. Finally, $x_{denoised}$ is the denoised adjacency matrix, showing the effect of denoising the data-based matrix using the physics-based information and the specified threshold. Once the denoised adjacency graphs are constructed for each of 144 time intervals, those are used as an input for the prediction algorithm.

BENCHMARK ANALYSIS

The proposed model is developed using Pytorch 1.1.0 on a virtual workstation with a NVIDIA Quadro P2200 GPU. Adam optimizer is used to train the model. The learning rate is set to 0.001. The depth of layer for proposed model is set to 1 and hidden state size is kept at 64. The batch size is set to 64 and number of epochs are set to 100. In order to avoid overfitting, early stopping criteria is enforced.

Evaluation metrics of the Prediction

We evaluated the model performance based on three evaluation indicators, namely the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE) (*20*). These metrics are defined as follows.

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \\ MAPE &= \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \\ RMSE &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2} \end{aligned} \tag{12}$$

where n is the length of time series, Y_t indicates the actual measurement, $\hat{Y}_t$ represents the predicted value from the model, and $\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|$ denotes the forecast error. MAE reflects the absolute error of the prediction result. MAPE is a measure of prediction accuracy of a forecasting method in statistics. RMSE can more accurately reflect the ability of model to predict the values.

PI-GRNN Benchmarks

The performance of PI-GRNN model is compared with basic statistical models and with latest hybrid GNN models using evaluation metrics. The prediction is determined for two horizons, 30 minutes (three time intervals) and 1 hour (six time intervals). The baseline models are as follows.

HA: Historical Average (HA) method predicts the future speed using average of historical data.

ARIMA: An autoregressive integrated moving average (ARIMA), is a statistical analysis model predicts future values based on past values.

(a) The evaluation metrics of developed model and benchmarks

Model	Time	MSE	MASE	RMSE
HA	30 minutes	4.20	7.85	13.05%
ARIMA	30 minutes	5.18	10.5	12.75%
DCRNN	30 minutes	3.20	6.50	8.85%
AGCRN	30 minutes	3.25	6.70	9.03%
DGCRN	30 minutes	2.99	6.05	8.02%
PI-GRNN	**30 minutes**	**2.74**	**5.50**	**7.70%**
HA	1 hour	4.20	7.85	13.05%
ARIMA	1 hour	6.95	13.25	17.50%
DCRNN	1 hour	3.63	7.64	10.52%
AGCRN	1 hour	3.64	7.53	10.40%
DGCRN	1 hour	3.46	7.25	9.75%
PI-GRNN	**1 hour**	**3.38**	**7.19**	**9.68%**

(b) Percent uncertainty reduction of developed model and benchmarks

Model	Percent reduction in uncertainty
KF-pir and mixture model	**19.3%**
KF-pir	14.1%
KF-tml(*4*)	5%
KF-traditional	2.1%

TABLE 1: Performace evaluation of both models with respect to benchmarks

DCRNN(*21*): Diffusion Convolutional Recurrent Neural Network is fusion model of GCN with GRU for traffic data prediction.
AGCRN(*22*): Adaptive Graph Convolutional Recurrent Network is a model that combines GCN with GRU employing an adaptive graph structure.
DGCRN(*23*): Dynamic Graph Convolutional Recurrent Network model employs dynamic graph in GCN for spatial correlations and then use the GRU model to gain temporal dependencies.
Table 1a shows evaluation results.

KF Benchmark
The performance of KF-PIR and mixture model is compared against basic statistical model, traditional KF and TML ((*4*) data driven model. We used Mean Absolute Percentage Error (MAPE) as the measure of uncertainty. We assumed that, lower the value of MAPE, lower the uncertainty. The percentage uncertainty reduction is calculated against ARIMA model using following formula.

$$\text{Percent uncertainty reduction} = \frac{MAPE_{ARIMA} - MAPE_{after}}{MAPE_{ARIMA}} \tag{13}$$

where, $MAPE_{ARIMA}$ is MAPE after applying ARIMA model while $MAPE_{after}$ is MAPE after applying the model for which we want to calculate the percent reduction in uncertainty.

Above formula is employed to calculate the percent reduction in uncertainty for each model. Figure 6a shows the significant percent reduction in uncertainty of predictions when the PIR + Mixture model is employed. Table 1b shows the percent uncertainty reduction of all the models.

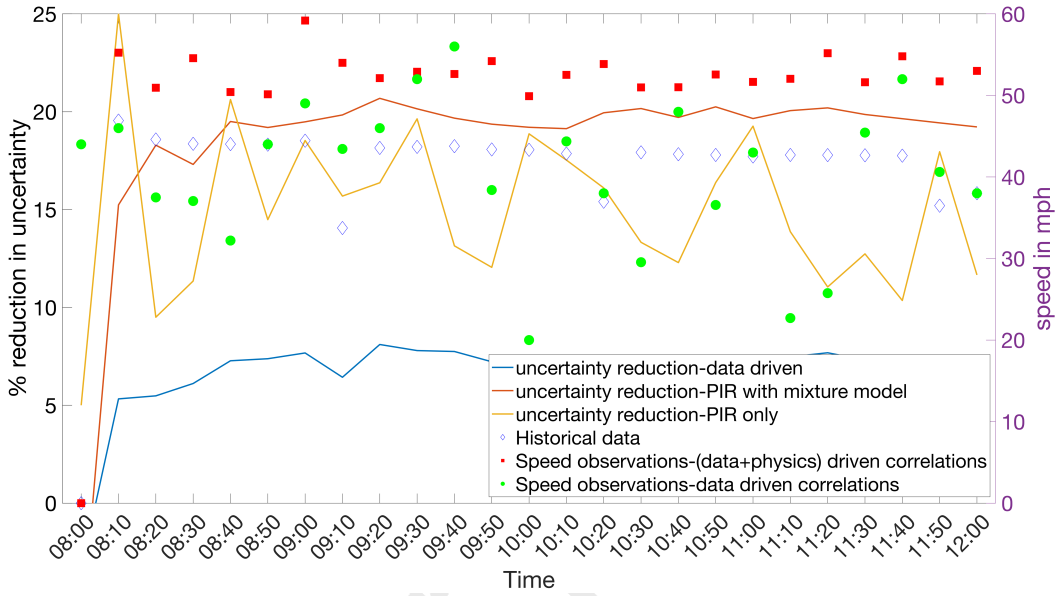
Useful tool for Traffic and Planning Agencies
Figure 6b represents that uncertainty in travel time prediction reduces as more observations become available. This tool will help traffic operators to gain insight on the amount of data required to achieve certain accuracy in predictions. Although this paper addresses the travel time prediction problem, the developed tool can be easily applied to other problems such as predicting the mixture of probability distributions of traffic flow. NCHRP Report 934 (*24*) found that one-third of all forecasts are in error by more than 30%, but there is no readily available tool that can clarify what amount of historical data would have improved it.

As shown in Figure 6b, after first 67 sample observations from multimodal multivariate clustered cells, the reduction of prediction uncertainty starts to slow down and Figure 6c shows that there are more uncertainty reduction in earlier time interval and the reduction of prediction uncertainty starts to slow down.

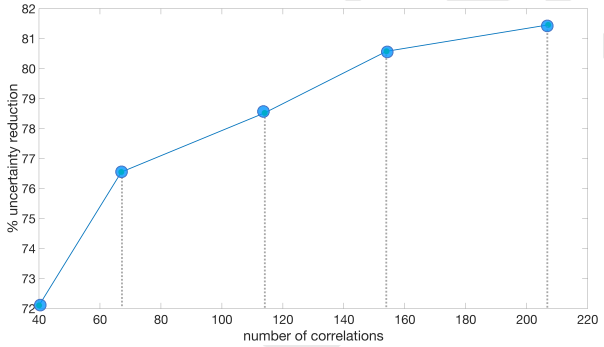
This graph for each geographical region will be specially useful for traffic and planning agencies knowing how much sample observations they need to improve the traffic prediction capability and plan the future projects. Our tool simply suggests how to use those unused values in the older forecasts,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

$$\omega(x, o_{n_{type}}) = \max\left(0, \frac{O^2_{n_{type}} - o^2_{n_{type}}}{O^2_{n_{type}} + o^2_{n_{type}}}\right) \tag{14}$$

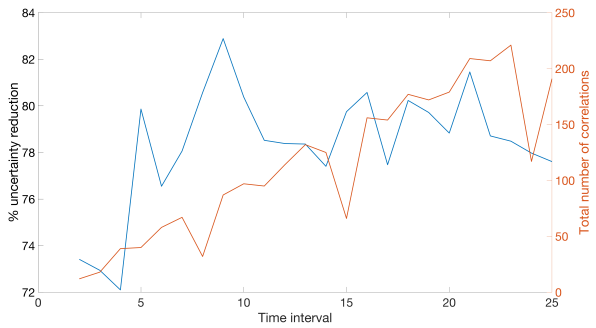
where $o_{n_{type}}$ is the number of observations of similar type cells $k \in K$ in cell j of cluster n_{type} and



(a) Percent change in uncertainty for developed KF and benchmark models



(b) Critical sample size reducing the prediction uncertainty



(c) Critical sample size reducing the prediction uncertainty across time

FIGURE 6: A new tool to guide where and how long to collect data

$O_{n_{type}}$ is the optimal number of observations (samples) needed to reduce the entropy of the cluster to zero. This weight implies a decreasing univariate entropy of the clusters as more observations of their member cells are made. In other words, as the number of observations $o_{n_{type}} \rightarrow O_{n_{type}}$, the weight $\omega(x, o_{n_{type}}) \rightarrow 0$. Also, since there is high confidence in the measurement in cells belonging to clusters with low entropy, we exploit those low entropy cluster types through a few sampling of their member cells.

In a real time operation, those updates on observations occurs sequentially. In a highly multimodal and multivariate correlated road environment, not all information gain is equally valuable. This is particularly important for traffic operators to be more agile in prioritizing historical patterns against the current observations. In general, gaining information earlier may prevent cascading effects on uncertainty in travelers behavior which may lead to inaccurate route guidance. Travelers may react similarly and collectively transfer congestion from one route to another (*25*) considering travelers' tolerance for unexpected delays (*26*). The information gained at the end of the data sampling or the trip is less likely to provide significant benefits.

CONCLUSION

The route suggestion users receive at the outset of their commute may not be optimal when they are on the road due to the uncertainty in travel time prediction. While more reliable traffic predictions can be achieved by capturing unobserved heterogeneity by analyzing mixture of multiple probability distributions via data-driven models, statistical transition of this knowledge across different time and space has not investigated in the previous study. Furthermore, incorporating physics knowledge (i.e., traffic theory) can regularize the spurious correlation that may exist in the data-driven models. However, traditional machine learning frameworks overlook simultaneous observations of more than one variable. As a result, those high-dimensional machine learning-based prediction models are intractable.

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 against author's previous *Temporal Multimodal Multivariate Learning* (*4*). These models can solve challenging tasks where the uncertainty is revealed in a sequence by grouping samples within similar distribution types and inferring the posterior based on expected observations. This paper opens appealing research opportunities in the study of information-theoretic decision making that exhibit nontrivial indirect learning from spatiotemporal correlation.

ACKNOWLEDGMENTS

Funding for this research was provided by NSF [1910397, 2106989] and NCDOT [TCE2020-01].

AUTHORS CONTRIBUTION

The authors confirm contribution to the paper as follows: study conception and design: Hyoshin Park, Venktesh Pandey, Niharika Deshpande, Justice Darko; data collection: Niharika Deshpande, Justice Darko; analysis and interpretation of results: Niharika Deshpande, Justice Darko, Hyoshin Park; draft manuscript preparation: Niharika Deshpande, Justice Darko, Hyoshin Park. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Han, Q. and H. Timmermans, Interactive Learning in Transportation Networks with Uncertainty, Bounded Rationality, and Strategic Choice Behavior: Quantal Response Model. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1964, 2006, pp. 27–34.
2. Alt, B., A. Šošic, and H. Koeppl, *Correlation Priors for Reinforcement Learning*, 2019.
3. Errica, F., D. Bacciu, and A. Micheli, Graph Mixture Density Networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, 2021, pp. 3025–3035.
4. Park, H., 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.
5. Folsom, L., 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.
6. Kumar, S., Traffic Flow Prediction using Kalman Filtering Technique. *Procedia Engineering*, Vol. 187, 2017, pp. 582–587.
7. Pueboobpaphan, R. and T. Nakatsuji, Real-Time Traffic State Estimation on Urban Road Network: The Application of Unscented Kalman Filter, 2006, pp. 542–547.
8. Mihaylova, L., R. Boel, and A. Hegyi, An unscented Kalman filter for freeway traffic estimation, 2006.
9. Deshpande, N., H. Park, V. Pandey, and G. Yoon, Advancing Temporal Multimodal Learning with Physics Informed Regularization, 2023, pp. 1–5.
10. Yu, B., H. Yin, and Z. Zhu, Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting, 2017.
11. Rico, J., J. Barateiro, and A. Oliveira, Graph Neural Networks for Traffic Forecasting, 2021.
12. Fu, R., Z. Zhang, and L. Li, Using LSTM and GRU neural network methods for traffic flow prediction, 2016, pp. 324–328.
13. Huang, A. J. and S. Agarwal, Physics Informed Deep Learning for Traffic State Estimation: Illustrations with LWR and CTM Models. *IEEE Open Journal of Intelligent Transportation Systems*, Vol. 3, 2022, pp. 1–1.
14. Han, K., W. Y. Szeto, and T. L. Friesz, Formulation, existence, and computation of boundedly rational dynamic user equilibrium with fixed or endogenous user tolerance. *Transportation Research Part B: Methodological*, Vol. 79, 2015, pp. 16–49.
15. Guo, X., Toll sequence operation to realize target flow pattern under bounded rationality. *Transportation Research Part B: Methodological*, Vol. 56, 2013, pp. 203–216.
16. Di, X., H. X. Liu, and X. J. Ban, Second best toll pricing within the framework of bounded rationality. *Transportation Research Part B: Methodological*, Vol. 83, 2016, pp. 74–90.
17. Hochreiter, S. and J. Schmidhuber, Long short-term memory. *Neural computation*, Vol. 9, No. 8, 1997, pp. 1735–1780.
18. Cho, K., B. Van Merriënboer, D. Bahdanau, and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

19. Park, H. and A. Haghani, Optimal Number and Location of Bluetooth Sensors Considering Stochastic Travel Time Prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 55, 2015, pp. 203–216.
20. Hyndman, R., Another Look at Forecast Accuracy Metrics for Intermittent Demand. *Foresight: The International Journal of Applied Forecasting*, Vol. 4, 2006, pp. 43–46.
21. Li, Y., R. Yu, C. Shahabi, and Y. Liu, Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR '18)*, 2018.
22. Bai, L., L. Yao, C. Li, X. Wang, and C. Wang, Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2020, NIPS'20.
23. Li, F., J. Feng, H. Yan, G. Jin, D. Jin, and Y. Li, *Dynamic Graph Convolutional Recurrent Network for Traffic Prediction: Benchmark and Solution*, 2021.
24. NCHRP Report 934, Traffic Forecasting Accuracy Assessment Research. In *National Academies of Sciences, Engineering, and Medicine, Washington, DC: The National Academies Press.*, 2020.
25. Ben-Akiva, M., D. McFadden, and T. e. a. Gärling, Extended Framework for Modeling Choice Behavior. *Marketing Letters*, Vol. 10, 1999, pp. 187–203.
26. Macfarlane, J., When Apps Rule the Road: Your Navigation App Is Making Traffic Unmanageable. *IEEE Spectrum*, 2019.