2024

Physics-Informed Deep Learning With Kalman Filter Mixture for Traffic State Prediction

Niharika Deshpande
Old Dominion University

Hyoshin (John) Park
Old Dominion University

Original Publication Citation

Deshpande, N., & Park, H. (2024). Physics-informed deep learning with Kalman Filter mixture for traffic state prediction. *International Journal of Transportation Science and Technology*. Advance online publication. https://doi.org/10.1016/j.ijtst.2024.04.002

International Journal of Transportation Science and Technology xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Transportation Science and Technology

journal homepage: www.elsevier.com/locate/ijtst

Research Paper

Physics-informed deep learning with Kalman filter mixture for traffic state prediction

Niharika Deshpande, Hyoshin (John) Park *

Department of Engineering Management & Systems Engineering, Old Dominion University, 2101F Engineering Systems BLDG, Norfolk 233529, USA

ARTICLE INFO

Article history:
Received 26 November 2023
Received in revised form 1 March 2024
Accepted 3 April 2024
Available online xxxx

Keywords:
Kalman filter
Deep learning
Physics-informed
Graph neural network
Uncertainty reduction

ABSTRACT

Accurate traffic forecasting is crucial for understanding and managing congestion for efficient transportation planning. However, conventional approaches often neglect epistemic uncertainty, which arises from incomplete knowledge across different spatiotemporal scales. This study addresses this challenge by introducing a novel methodology to establish dynamic spatiotemporal correlations that captures the unobserved heterogeneity in travel time through distinct peaks in probability density functions, guided by physics-based principles. We propose an innovative approach to modifying both prediction and correction steps of the Kalman Filter (KF) algorithm by leveraging established spatiotemporal correlations. Central to our approach is the development of a novel deep learning model called the Physics Informed-Graph Convolutional Gated Recurrent Neural Network (PI-GRNN). Functioning as the state-space model within the KF, the PI-GRNN exploits established correlations to construct dynamic adjacency matrices that utilize the inherent structure and relationships within the transportation network to capture sequential patterns and dependencies over time. Furthermore, our methodology integrates insights gained from correlations into the correction step of the KF algorithm that helps in enhancing its correctional capabilities. This integrated approach proves instrumental in alleviating the inherent model drift associated with data-driven methods, as periodic corrections through update step of KF refine the predictions generated by PI-GRNN. To the best of our knowledge, this study represents a pioneering effort in integrating deep learning and KF algorithms in this unique symbiotic manner. Through extensive experimentation with real-world traffic data, we demonstrate the superior performance of our model compared to the benchmark approaches.

1. Introduction

Daily commutes are often disrupted by road closures, accidents, and adverse weather conditions, resulting in unexpectedly prolonged travel times. Traffic congestion can lead to frustrating delays and wasted time, impacting our schedules and productivity. Overall, it becomes a pressing issue that negatively impacts the quality of life and economic productivity. Fortunately, advances in data intelligence and urban computing have made it possible to collect massive amounts of traffic

* Corresponding author.
E-mail addresses: ndesh002@odu.edu (N. Deshpande), h1park@odu.edu (H. (John) Park).

https://doi.org/10.1016/j.ijtst.2024.04.002
2046-0430/

Please cite this article as: N. Deshpande and H. (John) Park, Physics-informed deep learning with Kalman filter mixture for traffic state prediction, International Journal of Transportation Science and Technology, https://doi.org/10.1016/j.ijtst.2024.04.002

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 several limitations like stationarity assumption for time series data, their incapability to capture spatiotemporal dependencies etc. These factors are crucial for accurate predictions especially in urban transportation systems. These limitations are addressed by more advanced techniques such as machine learning and DL that can effectively handle the non-stationarity, capture complex patterns, and utilize spatio-temporal dependencies present in traffic data, leading to improved forecasting performance and better transportation management strategies.

This study advances traffic prediction by extending the conventional KF with a DL approach, addressing the limitations of linear and Gaussian assumptions. We integrate historical and real-time data to enhance predictive accuracy, expand the capabilities of KF algorithm, and advance the research on uncertainty quantification. Utilizing a DL algorithm as the state estimation component within the KF offers several advantages. It enables the KF to handle the inherent non-linearity in traffic dynamics more effectively, adapt to changing conditions, and help alleviate epistemic uncertainty. Moreover, our methodology incorporates an innovative modification to the KF correction step. By considering observations from correlated links and subjecting them to a reliability test using the entropy method, we refine the correction process, that enhances the model's ability to recognize and utilize trustworthy information. This approach aims to further strengthen the KF's resilience to uncertainties, ultimately contributing to more accurate, adaptable, and robust traffic predictions. Our methodology is intricately aligned with the complexities of real-world traffic behavior, pushing the boundaries of state-of-the-art traffic prediction techniques.

The DL model that we will employ as a state space model within the KF, represents a key component of our methodology. Given the effectiveness of Graph Neural Networks (GNN) in navigating non-Euclidean topological spaces such as traffic networks, we considered it a well-suited architecture for our Deep Learning (DL) model. 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 graph based on topology alone. Hence, in our study, we use graphs evolving with time based on spatiotemporal correlations that can account for constantly changing conditions, interactions, and relationships between road segments. Despite considering dynamic graphs in GNN, it is challenging to capture long-term temporal dependencies. Thus, considering dedicated techniques in the algorithm to capture temporal correlations is crucial for the accuracy of the model. Another challenge that arises while determining the correlations is the problem of "false/coincidental correlations" which are introduced while analyzing large amounts of data. To avoid erroneous conclusions and inaccurate predictions, we have introduced the physics-informed method in this study. The main contributions of our paper are summarized as follows:

- We introduced a novel method for determining spatiotemporal correlations among non-contiguous locations in spatial and temporal dimension. We used the multimodal (multipeak) probability distributions (PDF) of traffic states to gain insights on correlations based on changing dynamics of traffic. Spurious correlations are filtered by employing physics of traffic flow theory ensuring genuineness of correlations.
- A new DL model named 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 using established spatiotemporal correlations. 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 introduction of DL with a KF mixture to minimize information uncertainty in traffic state estimation. Our approach leverages PI-GRNN as a state-space model within KF. Additionally, we refine the correction step of the KF by incorporating observations gained from established correlations, selectively utilizing reliable data to enhance the precision of predictions.

The rest of the paper is organized as follows: Section 2 reviews the literature related to multimodal-multivariate learning, KF and DL models. Section 3 discusses a data-driven multimodal (multipeak), multivariate, temporal learning approach. Section 4 presents a methodology for KF with DL technique. After that, we evaluate the performance of the proposed model against corresponding benchmarks in Section 5. Finally, we provide conclusions in Section 6.

2. Literature review

Without knowing the future traffic with confidence, the traditional choice theory considers bounded rationality of the majority of travelers taking a detour, which causes more congestion on nearby roads (Han and Timmermans, 2006). 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 (Alt et al., 2019). Recent mixture density networks in approximating multimodal (multipeak) output distribution have well-handled prediction uncertainties rather than averaging the distribution (Errica et al., 2021). While those advanced multimodal (multipeak) 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. (2022) developed a data-driven model that uses 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 (Folsom et al., 2021). However, they haven't addressed the coincidental correlations introduced due to a purely data-driven approach. This research addresses this issue by introducing a physics-informed approach utilizing multivariate traffic data to establish causality and eliminate accidental correlations.

Various studies use KF with other techniques to enhance the accuracy and robustness of traffic state predictions (Kumar, 2017). These models consider the evolution of traffic based on only neighboring segments (Mihaylova et al., 2006). Unexplored states beyond neighboring segments in space and time dimensions can negatively affect the prediction accuracy of the model (Deshpande et al., 2023). Therefore, we deal with it by establishing spatio-temporal correlations among non-contiguous locations. Traditional KF uses only the numerical value of recent observations but we customize the algorithm to use the derived PDF which helps in incorporating comprehensive information. Traditional KFs can only be used in linear systems with Gaussian white noise. To estimate the state of a nonlinear dynamic system, different variations of KFs like EKF, UKF (Julier et al., 1997) and CKF (Arasaratnam and Haykin, 2009) were proposed. Each of these variations has disadvantages associated with them. EKF relies on linearization, which means it assumes that the system dynamics and measurement models can be approximated as linear within the vicinity of the current state estimate. This assumption breaks down for highly nonlinear systems, leading to errors. While the UKF is more robust to non-linearities compared to the EKF, it may still struggle with highly nonlinear systems, and the choice of sigma point distribution can impact its performance. Like other nonlinear filters, the CKF can be sensitive to model errors and deviations from the assumed system and measurement models. If the models are significantly inaccurate, the filter's performance may degrade. To overcome the shortcomings of these non-linear filters, hybrid models that incorporate deep learning techniques can significantly improve their performance.

The studies developed algorithms that use Kalman filters and neural networks in combination, either to train the state-space model's equations or parameters with a neural network (Chen et al., 2019) or to update the neural network's parameters using a Kalman filter (Guan et al., 2013). Revach et al. (2021) replaces the Kalman Gain (KG) component in KF with RNN but it may struggle with generalizing to unseen or different situations, making it less robust compared to traditional KF's KG computation, which relies on predefined mathematical models. To overcome this shortcoming, in our study, we replace the state-space model with a novel physics-informed deep neural network technique. Since the model has physics-informed component, it positively captures domain knowledge and non-linearity in dynamic systems. This modification in KF eliminates the need for accurate knowledge and modeling of the underlying dynamics. We develop a novel KF by incorporating these improvements which outperformed the benchmarks.

In the realm of DL, correlations in data can be effectively captured using various techniques, including Recurrent Neural Networks (RNN) (Yu et al., 2017), 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. GRU and LSTM (Fu et al., 2016) techniques are specialized versions of RNN to overcome the problem of vanishing gradient posed by RNN. On the other hand, spatial dependencies in data can be modeled using CNNs, Graph Neural Networks (GNN) (Yu et al., 2017), or attention mechanisms. 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 (Rico et al., 2021). 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. To account for this critical aspect, in this study, we employ a physics-informed multimodal (multipeak), multivariate data-based approach to create dynamic graphs. A few deep learning algorithms incorporate physics-related aspects in the cost function of the model (Huang and Agarwal, 2022) but this study integrates physics-informed spatiotemporal correlations into the deep learning framework with GNN and GRU resulting in an innovative mixture algorithm which helps ensure that model training is constraint by physical principles.

3. Data-Driven Temporal Multimodal (Multipeak) multivariate learning

We start extending standard deviation-based information theory (Folsom et al., 2021) 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 PDFs. 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 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.

Fig. 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 $\mathbb{A}, \mathbb{A}\prime, \mathbb{B}, \mathbb{C}$ as an unimodal probability distribution with an expected travel time. Without knowing the future traffic with confidence, the traditional choice theory considers the bounded rationality (Han and Timmermans, 2006; Han et al., 2015; Guo, 2013; Di et al., 2016) 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-min shortcut, in the case of a scenario that turned out to be 2-min 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 min due to the extended clearance time of the incident if we know the overall travel time of the route is 4 min. 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 min, we can notify fewer vehicles to use this route, and redistribute traffic to other routes (i.e., BC) having shorter travel times.

The provided approach intertwines with the concept of bounded rationality, enriching the decision-making process within transportation systems. Firstly, the identification of "correlated cells" and the recognition that their states are probabilistic until observed aligns with bounded rationality. Bounded rationality acknowledges that decision-makers have limited cognitive abilities and access to information. By categorizing cells based on similar travel time probability distributions, the approach acknowledges the inherent uncertainty in traffic conditions and accommodates it in the decision-making. Secondly, the process of sharing information among travelers to influence route choices reflects bounded rationality's principle of adaptive decision-making under constraints. Travelers continuously update their path plans based on information about the grid's state. This adaptive behavior allows travelers to navigate the transportation network efficiently despite the constraints imposed by limited information and cognitive biases. Lastly, the confirmation mechanism, where identical cells visited by different travelers provide confirmation of correlation assumptions, aligns with bounded rationality's emphasis on learning and adjusting strategies based on experience. This confirmation process enhances the reliability of assumptions made about correlated cells, contributing to more informed decision-making in uncertain environments. In summary, the proposed approach leverages principles of bounded rationality to enhance decision-making in transportation systems by

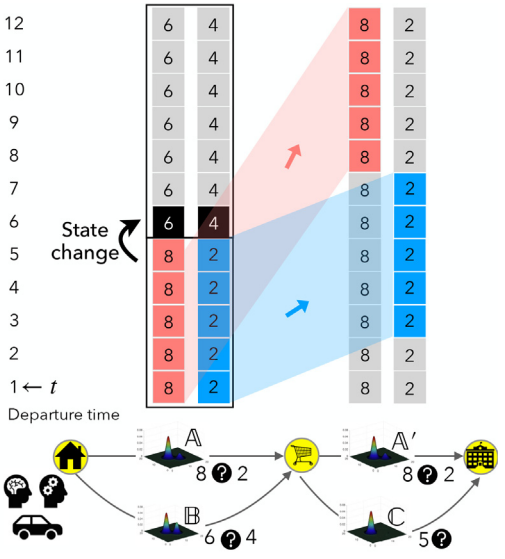


Fig. 1. Temporal multimodal Learning considering correlation between time-varying bimodal link distributions. Red represents one peak in the multimodal distribution indicating 8 min of travel time while blue represents another peak indicating 2 min. Question mark icon indicates uncertainty associated with whether the road is congested (red peak) or not congested (blue peak) at a specific departure time. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

acknowledging uncertainty, facilitating adaptive strategies, and incorporating learning from past experiences. This integration of bounded rationality principles strengthens the effectiveness and reliability of the approach in optimizing travel time predictions.

4. Kalman filtering with physics-informed deep learning state-space model (KF-PIR & MIXTURE)

We gather in Table 1 all the notations used in this section.

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 (Fig. 3). In this study, we employ a predictive model rooted in deep learning methodologies, as elaborated in the subsequent section, to anticipate the evolution of a selected state variable at the forthcoming time increment $t+1$. This chosen state variable represents a quantifiable parameter of interest, and our objective is to leverage the inherent capabilities of deep learning techniques to generate accurate forecasts for this variable's values at future time points. 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 is estimated by using Expectation Maximization. The optimal distribution of the data over K clusters is 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.

4.1. Multimodal physics-informed deep learning as a state-space model

This section outlines the creation and refinement of the prediction model employed within the context of the KF framework. 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 the benefits of GNNs. The studies like (Yu et al., 2017) use static adjacency matrix in GNN which can only capture spatial correlations based on geometry. This study introduces a novel method of calculating semantic adjacency between nodes based on the 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.

4.1.1. Framework for graph convolution gated recurrent network

A traffic prediction problem can be formulated as a time-series forecasting problem with historical data and prior knowledge. The prior knowledge used in Graph Neural Network (GNN) is a pre-defined adjacency graph $\mathscr{G}=(\mathscr{V},\mathscr{E},A)$. Here, $\mathscr{V}$ is a set of nodes that 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 L step graph signals $\mathbf{X}_{(t-L):t}$, to learn a function g which can map $\mathbf{X}_{(t-L):t}$ and $\mathscr{G}$ to next M step graph signals $\hat{\mathbf{X}}_{t:(t+M)}$, represented as follows:

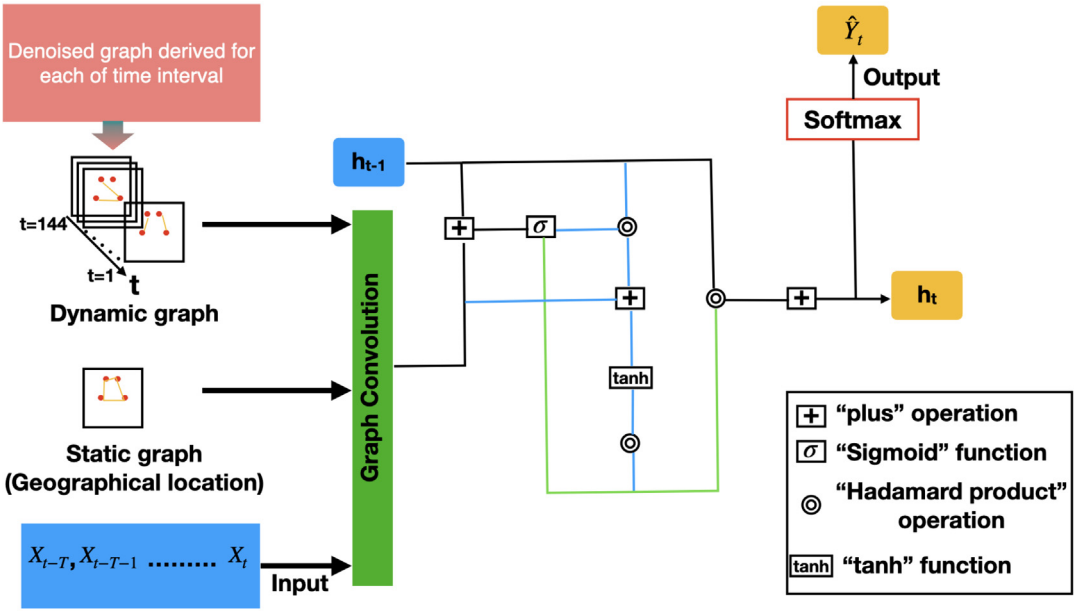
$$\left[\mathbf{X}_{(t-L):t},\mathscr{G}\right]\xrightarrow{g}\hat{\mathbf{X}}_{t:(t+M)} \tag{1}$$

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

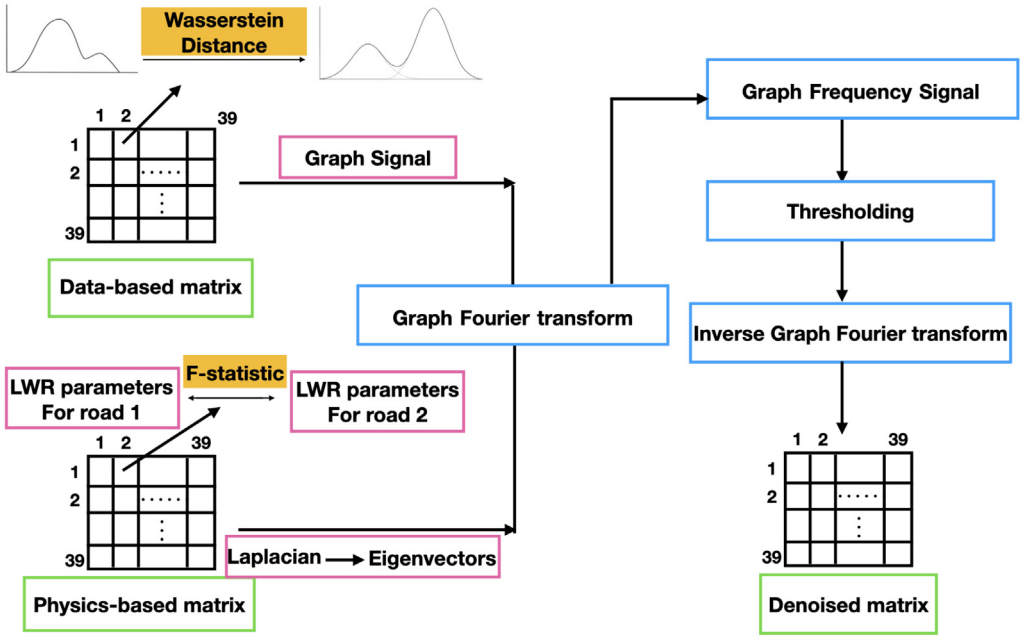
Table 1
Explanation of symbols.

Symbol	Explanation	Symbol	Explanation
z_t, r_t, c_t	Update, reset, cell state of GRU, respectively	$W_1(P,Q)$	Wasserstein distance between distributions P & Q
h_t	Hidden state of GRU	X	Graph Signal
W_u, W_r, W_c	Weight matrix associated with update, reset, cell operations, respectively	$H(X)$	Entropy of the continuous random variable X
A	Pre-defined adjacency graph	DA	Dynamic adjacency graph
SSR	Sum of squared residuals	ρ	Traffic density
t_f	traffic flow rate	y	observed speed value
$f_{updated}(x)$	Updated PDF with correlated observation	t	Time interval of 10 min
μ	mean of PDF	σ	Standard deviation of PDF
$\hat{x}_t^-$	Predicted state at time t	E_t	Error covariance matrix at t
N_Q	Process Noise	N_R	Observation Noise
K_t	Kalman gain	Z_t	Observed data for correction
$g(\cdot)$	Function learned from PI-GRNN	$g\prime(\cdot)$	Jacobian matrix of PI-GRNN

As shown in Fig. 2a, Recurrent Neural Networks (RNN) are used in the case of sequential data as they retain the previous states in memory while accepting the current state. Therefore, it becomes a suitable means to solve time-series predictions. However, RNNs are capable of capturing only short-term temporal dependencies and have the issue of vanishing gradient. These limitations of RNN can be overcome by Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014). GRU has a less complex structure than LSTM as it has less number of gates, is easy to modify, and is faster to train. Therefore, we choose GRU for extracting temporal correlations from traffic time series data. We replace the matrix multiplications in GRU with the Graph Convolution (GC) module which is described using the following equations.



(a) Overview of Graph Convolution Gated Recurrent Neural Network



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

Fig. 2. The architecture of PI-GRNN.

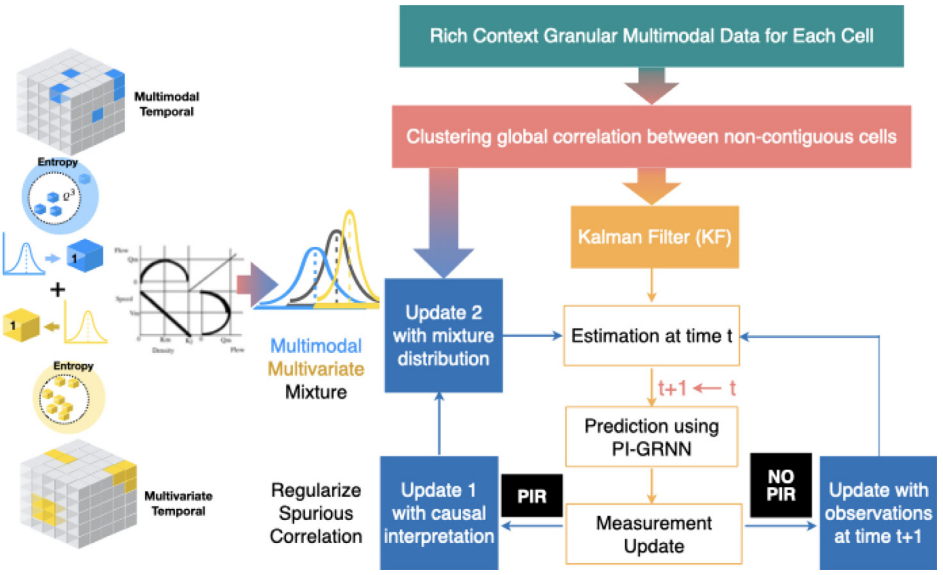


Fig. 3. Physics-informed and -regularized (PIR) KF in the hierarchical update steps with PI-GRNN as a state-space model.

$$\begin{aligned} z_t &= \sigma(W_u.[GC(DA_t, A), h_{t-1}] + b_u) \\ r_t &= \sigma(W_r.[GC(DA_t, A), h_{t-1}] + b_r) \\ c_t &= tanh(W_c.[GC(DA_t, A), (r_t * h_{t-1})] + b_c) \\ h_t &= z_t * h_{t-1} + (1 - z_t) * c_t \end{aligned} \tag{2}$$

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

4.1.2. Adjacency matrix with multimodal (multipeak) data

The historical data for speed v is collected from the TMCs for a year. The data is collected in the interval of 10 min. The data is sorted for each TMC during 144 time intervals of 10 min within 24 h. The sorted data is then categorized into 14 speed bins ranging from 2 mph to 100 mph with a difference of 7 mph for each TMC. The histogram is derived from it and then the PDF of speed for each TMC for each time interval using kernel density estimation. 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 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, the most suitable one for measuring similarity between multimodal distribution is found to be the 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 the 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 can 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 the most suitable parameter to measure the similarity between calculated speed distributions. In this study, we have established the similarity between TMCs for each 144 time intervals within 24 h. Based on 39 TMCs within the area of study, the 39 × 39 matrix is derived for each interval by filling values with Wasserstein distance using the following equation.

$$W_1(P, Q) = \min_{\gamma \in \Gamma(P,Q)} \sum_{i,j} \gamma_{ij} \cdot d(p_i, q_j) \tag{3}$$

where, $W_1(P, Q)$ represents the Wasserstein distance between distributions P & Q, $W_1(P, Q) \in [0, \inf), \Gamma$ is a transportation plan that defines the amount of mass to be moved from each point in P to Q. It satisfies the constraints of being a valid joint distribution with marginals P & Q, denoted as $\gamma \in \Gamma(P, Q)$, p_i & q_j represent individual points (samples) from distributions P &

Q, respectively, $d(p_i, q_j)$ is a distance metric (e.g., Euclidean distance or any other suitable distance measure) between p_i and q_j.

We assume that as the Wasserstein distance between the two PDFs is large, they exhibit lower degree of 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 the following formula.

$$W_{1(normalized)} = \frac{W_1(P,Q) - W_1(P,Q)_{min}}{W_1(P,Q)_{max} - W_1(P,Q)_{min}} \tag{4}$$

where, $W_1(P,Q)$ is the Wasserstein distance which is the value in matrix 39×39 to be normalized, and $W_1(P,Q)_{min}$ & $W_1(P,Q)_{max}$ is the minimum & maximum value in the matrix, respectively. The following equation is used to generate a weighted adjacency matrix.

$$W_{1(weighted)} = 1 - W_{1(normalized)} \tag{5}$$

4.1.3. Adjacency matrix with multivariate data by traffic flow theory

Adjacency matrix is also determined based on the physics of traffic flow. The LWR (Lighthill–Whitham–Richards) 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 the 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 fundamental diagram, where the flow rate is a function of traffic density and speed.

Mathematically, it can be expressed using the following equation:

$$\frac{\partial \rho}{\partial t} + \frac{\partial t_f}{\partial x} = 0 \tag{6}$$

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 and t_f is the traffic flow rate (number of vehicles per unit time). The LWR describes the evolution of traffic density and speed over time and space. If the parameters of the LWR 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 in one segment may impact the traffic conditions in the other segment comparably. Hence, in this study, we compared the parameters for all TMCs using speed and density data collected over a year using loop detectors.

This study employs a method of characteristics to obtain LWR parameters by solving differential equations backward in time from an 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 the least squares optimization method. Once the LWR parameters are estimated for each of the 39 TMCs, the sum of squared residuals (SSR) is calculated using the following equation. It represents the overall deviation of the LWR model predictions from the observed speed data.

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

where, SSR is the sum of squared residuals, y_i is the observed speed value at data point i and $\hat{y}_i$ is the predicted speed value at data point i based on the LWR model. Then F-statistic is computed using the following formula.

$$F = \frac{(SSR1 - SSR2)/(k2 - k1)}{SSR2/(n - k2)} \tag{8}$$

where, SSR1 & SSR2 are the sum of squared residuals for TMCs 1 & 2, respectively, and k1 & k2 are the degrees of freedom (DOF) for TMCs 1 & 2, respectively. In the case of LWR model, DOFs are typically 5 (number of parameters) minus the number of data points used for estimation, and n is the total number of data points used for estimation (across both TMCs). 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. For values in the adjacency matrix to fall between the range [0,1], the values are normalized using Eq. 3 described above.

The adjacency matrix weights derived across all 144 time intervals solely from empirical data undergo refinement through an assessment of the interconnection strength between two road segments by comparing their constant LWR parameters. It helps to rectify the spurious correlations that are introduced coincidentally while analyzing large amounts of data without any causal relationship between them. Hence, it is crucial to study them by considering the underlying cause using traffic flow theory.

4.1.4. Denoised adjacency matrix

We employ the 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 a denoised adjacency graph is depicted in Fig. 2b.

The Graph Fourier Transform is computed using a 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 the following formula.

$$Ln = D - A, \quad D_{ii} = \sum_{j} A_{ij} \tag{9}$$

where, D is a degree matrix, i and j represent the row and column index, and A is an adjacency matrix. Then, eigenvalues and eigenvectors of matrix L are computed using the following equation.

$$Ln \cdot U = U \cdot \Lambda \tag{10}$$

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 the adjacency matrix based on pure data.

$$S = U^T \cdot X \tag{11}$$

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 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, to transform the graph frequency signal S back into the vertex domain, the Inverse Graph Fourier Transform (IGFT) is performed.

$$X_{denoised} = U \cdot S \tag{12}$$

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 the 144 time intervals, those are used as input for the prediction algorithm.

4.2. Kalman Filter algorithm

We are using the deep learning prediction model PI-GRNN as a state-space model in KF. We are essentially altering the way the state transition matrix is computed. In traditional Kalman filters, this matrix represents the linear relationship between the state variables from one time step to the next. However, in PI-GRNN, this relationship may not be explicitly defined in a linear form but rather learned through the network's training process. We provided the modified equations of the KF to adapt them for the scenario where PI-GRNN is used as the state-space model without an explicit state transition matrix.

1. Prediction step:

$$\begin{aligned}\hat{x}_t^- &= g(\hat{x}_{t-1})\\ E_t^- &= g\prime(\hat{x}_{t-1})E_{t-1}^+ g\prime(\hat{x}_{t-1})^T + N_Q\end{aligned} \tag{13}$$

2. Update step:

$$\begin{aligned}K_t &= E_t^- B^T \left(BE_t^- B^T + N_R\right)^{-1}\\ \hat{x}_t &= \hat{x}_t^- + K_t\left(Z_t - B\hat{x}_t^-\right)\\ E_t &= (I - K_t B)P_t^-\end{aligned} \tag{14}$$

where, $\hat{x}_t$ is the predicted state at time t, E_t is the error covariance matrix at time t, N_Q is the process noise covariance matrix, K_t is the Kalman gain, N_R is the observation noise covariance matrix, $g(\cdot)$ is the function learned from PI-GRNN as shown in Eq. 1 in the manuscript, $g\prime(\cdot)$ is the Jacobian matrix of the PI-GRNN model function with respect to the state calculated using automatic differentiation in PyTorch, B is the connection matrix between the state vector and the measurement vector, and Z_t is the observed data used to correct the predicted estimate.

In the first update step of KF-PIR & MIXTURE, 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, observation from each correlated link is incorporated into the original PDF as given by following equation.

$$f_{\text{updated}}(x) = \sum_{i=1}^{peak} \frac{1}{\sqrt{2\pi}\sigma_{i_{\text{new}}}} e^{-\frac{\left(x-\mu_{i_{\text{new}}}\right)^2}{2\sigma_{i_{\text{new}}}^2}} \tag{15}$$

where, "peak" represents number of peaks in multimodal distribution. Updated mean of the new distribution is calculated using $\mu_{i_{\text{new}}} = \frac{n_i \cdot \mu_i + x_{\text{new}}}{n_i+1}$ while updated standard deviation using $\sigma_{i_{\text{new}}} = \sqrt{\frac{n_i \cdot \left(\sigma_i^2 + \left(\mu_i - \mu_{i_{\text{new}}}\right)^2\right) + \left(x_{\text{new}} - \mu_{i_{\text{new}}}\right)^2}{n_i+1}}$. Once PDF is updated with new observation, the entropy of updated PDF is calculated using following equation.

$$H(X) = -\int_{-\infty}^{\infty} f(x)\log_2(f(x))\,dx \tag{16}$$

where, H(X) represents the entropy of the continuous random variable X and f(x) is PDF. The entropy of old and updated PDF is compared. Entropy is assumed to be the measure of uncertainty. If entropy for new PDF turns out to be less than that of old one, the observation is assumed to be reliable one and used in correction step of KF. The whole algorithm of choosing reliable observation is explained in following algorithm. Algorithm 1 is used for each common observation to check its reliability. The mixture distribution is computed by incorporating all reliable observations with original distribution. Finally, this mixture distribution is used for the correction step in KF.

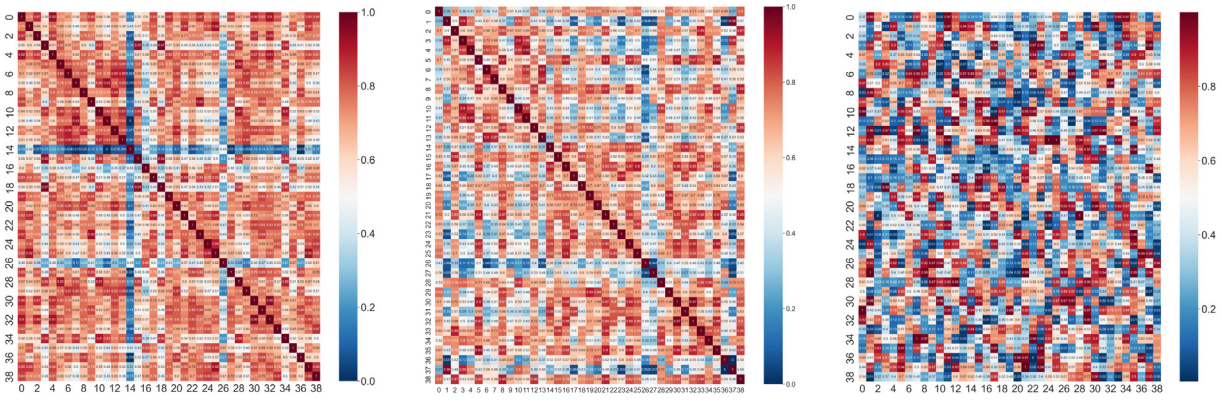
Algorithm 1. Filter reliable observations using entropy comparison and computing mixture distribution

1: Let obs be the set of all common observations from multiple peaks and multivariate clusters.
2: Compute the entropy H(Old) using the original distribution.
3: **for** each observation in obs **do**
4: Compute the entropy H(New) using the PDF incorporating new observation
5: **if** H(New) < H(Old) **then**
6: Add the new observation to the reliable observation list.
7: **else**
8: Discard the new observation.
9: **end if**
10: **end for**
11: Incorporate all reliable observations into the original distribution to obtain a mixture distribution.

5. Case study

Exploratory analyses and model development were conducted and validated on a 18.8-mile segment of the North Carolina Triangle Expressway (Fig. 4), utilizing probe vehicles and loop detectors data from 39 Traffic Message Channels (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 over each of 10 min of time interval. The 24 h in a day are divided into 10 min giving us 144-time

Fig. 4. A Case study for the network of Triangle Expressway: The interpolation method used by Park and Haghani (Park and Haghani, 2015) provides a 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-min interval, which divides a day's 24;h into 10 min.



(a) Weighted adjacency matrix across all TMCs based on multimodal PDFs at a time interval of 8-8:10 am (data-driven)

(b) Weighted adjacency matrix across all TMCs based on multimodal PDFs at a time interval of 8-8:10 pm (data-driven)

(c) Weighted adjacency matrix based on a multivariate relationship using LWR theory across all times (physics-informed)

Fig. 5. Adjacency matrices from data-driven and physics-informed approach.

intervals. The data is sorted within those intervals. This data is used in the model to determine a 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 an adjacency matrix based on interrelations between variables.

The multimodal data collected from TMC is used to calculate data-based weighted adjacency matrix of dimension 39×39 for each of 144 time interval following the method explained in Section 4.1.2. Fig. 5a and Fig. 5b shows the weighted adjacency matrix 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 across the TMCs. Therefore, when we consider adjacency entirely based on geographical proximity, we are missing out on the adjacency exhibited by the dynamic nature of traffic. Hence, the method of dynamic adjacency matrix proves to be superior in exploring a wide range of spatial correlations. Multivariate data obtained from loop detector is used to determine weighted adjacency matrix based on physics-based approach explained in Section 4.1.3. The matrix derived from the multivariate relation between speed and density based on the physics of traffic flow theory is shown in Fig. 5c. These matrices from multimodal and multivariate data are used to derive denoised adjacency graph for each time interval by employing methodology explained in Section 4.1.4.

6. Benchmark analysis

The proposed model is developed using Pytorch 1.1.0 on a virtual workstation with an NVIDIA Quadro P2200 GPU. The hyperparameters are chosen through a carefully parameter-tuning process on the validation set. We employed random search methodology where hyperparameter values were sampled randomly. This method uses random sampling guided

by prior knowledge which helps in efficiently discovering optimal configuration of hyperparameters.The model is trained using Adam optimizer. The learning rate is set to 0.001. The hidden state size is kept at 64. The batch size is set to 64 and the number of epochs is set to 100. To avoid overfitting, early stopping criteria are enforced. MAE is used as the loss function and if this metric doesn't improve for 5 number of epochs, the training is stopped. It takes 3 h to train the model.

To further improve the model's performance, we can implement adaptive sampling strategy where the probability of sampling hyperparameter values is adjusted based on the performance of previously sampled configurations. This dynamic approach can focus more on promising regions of the hyperparameter space. We can conduct sensitivity analysis to identify the most influential hyperparameters and we can focus on further exploration of them to fine-tune the model more effectively. We can experiment with the learning rate decay technique for training. It can help the model converge more effectively and achieve better performance. These potential adjustments can help the model to enhance its performance.

6.1. 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) (Hyndman, 2006). These metrics are defined as follows.

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \widehat{Y}_t\right| \\ MAPE &= \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right| \\ RMSE &= \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2} \end{aligned} \tag{17}$$

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

6.2. Benchmarks for PI-GRNN & KF models

The performance of the PI-GRNN model is compared with basic statistical models and with the latest hybrid GNN models using evaluation metrics. The prediction is determined for two horizons, 30 min (three time intervals) and 1 h (six time intervals). The baseline models are as follows & Table 2 shows evaluation results.

- **HA**: The Historical Average (HA) method predicts the future speed using an average of historical data.
- **ARIMA**: An autoregressive integrated moving average (ARIMA), is a statistical analysis model that predicts future values based on past values.
- **DCRNN** (Li et al., 2018): Diffusion Convolutional Recurrent Neural Network is a fusion model of GCN with GRU for traffic data prediction.
- **AGCRN** (Bai et al., 2020): Adaptive Graph Convolutional Recurrent Network is a model that combines GCN with GRU employing an adaptive graph structure.
- **DGCRN** (Li et al., 2021): Spatial–Temporal Graph Convolutional Gated Recurrent Network framework is designed to capture long-term dependency by mining periodic information in traffic data and models hidden spatial dependency using self-adaptive adjacency matrix.

Table 2
The evaluation metrics of the developed model and benchmarks.

Model	30 min			1 h		
	MAE	MAPE	RMSE	MAE	MAPE	RMSE
HA	4.20	7.85	13.05%	4.20	7.85	13.05%
ARIMA	5.18	10.5	12.75%	6.95	13.25	17.50%
DCRNN	3.20	6.50	8.85%	3.63	7.64	10.52%
AGCRN	3.25	6.70	9.03%	3.64	7.53	10.40%
DGCRN	2.99	6.05	8.02%	3.46	7.25	9.75%
STGCGRN	2.83	5.74	7.84%	3.43	7.22	9.74%
DGSTN	2.77	5.58	7.75%	3.4	7.2	9.69%
PI-GRNN	**2.74**	**5.50**	**7.70%**	**3.38**	**7.19**	**9.68%**
KF-PIR & MIXTURE	**2.63**	**5.26**	**7.39%**	**3.31**	**7.07**	**9.37%**

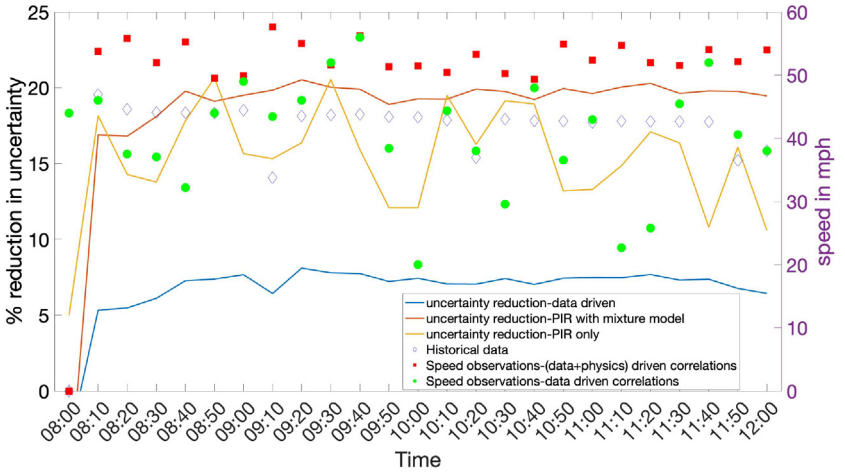


Fig. 6. Percent change in uncertainty for developed KF and benchmark models.

Table 3
Performance evaluation of the developed model and benchmarks (Part 2).

Model	Percent reduction in uncertainty
KF-PIR & Mixture	**19.3%**
KF-PIR	14.1%
KalmanNet (Revach et al., 2021)	13.5%
KF-TML (Park et al., 2022)	5%
KF-traditional	2.1%

- **STGCGRN** (Zhao et al., 2022): Dynamic Graph Convolutional Recurrent Network model employs dynamic graph in GCN for spatial correlations and then uses the GRU model to gain temporal dependencies.
- **DGSTN** (Jiang and Liu, 2023): Dynamic Graph Spatial–Temporal Neural Network model uses dynamic information maps to capture hidden node relationships and time-varying spatial correlations.

The performance of the KF-PIR & MIXTURE model is compared against the basic statistical model, traditional KF, KF-TML (Park et al., 2022) data-driven model and hybrid KF-deep learning model KalmanNet (Revach et al., 2021). We used Mean Absolute Percentage Error (MAPE) as the measure of uncertainty. We assumed that lower value of MAPE implies lower uncertainty. The percentage uncertainty reduction is calculated against the ARIMA model using the following formula.

$$\text{Percentuncertaintyreduction} = \frac{MAPE_{ARIMA} - MAPE_{after}}{MAPE_{ARIMA}} \tag{18}$$

where, $MAPE_{ARIMA}$ represents MAPE after applying the ARIMA model while $MAPE_{after}$ is MAPE following the application of the model under consideration to calculate the percent reduction in uncertainty.

The above formula is employed to calculate the percent reduction in uncertainty for each model. Fig. 6 shows the significant percent reduction in uncertainty of predictions when the PIR + Mixture model is employed. Table 3 shows the percent uncertainty reduction of all the models. The results presented in the table show that our model performs significantly better than benchmarks.

7. 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 a mixture of multiple PDFs via data-driven models, statistical transition of this knowledge across different times and space has not been investigated in the previous study. Furthermore, incorporating physics knowledge (i.e., traffic theory) can regularize the spurious correlations that may exist in the data-driven models. However, traditional ML frameworks overlook simultaneous observations of more than one variable. As a result, those high-dimensional ML-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 parent distribution of sample observations to evaluate their importance in improving the current prediction. We overcome the limitation of traditional direct (adjacent) learning by transferring online information through indirect learning of multiple modes of PDFs and multiple variables across different time stages.

The new family of statistical ML 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 the author's previous *Temporal multimodal Multivariate Learning* (Park et al., 2022). This paper introduces models capable of addressing complex tasks by grouping samples with similar distribution types in sequential data, allowing for posterior inference based on anticipated observations. It unveils research prospects in information-theoretic decision-making, particularly in nontrivial indirect learning from spatiotemporal correlations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding for this research was provided by NSF [1910397, 2106989] and NCDOT [TCE2020-01]. The authors declare that the contents of this article has not been published previously.

References

Alt, B., Šošic, A., Koeppl, H., 2019. Correlation priors for reinforcement learning.
Arasaratnam, I., Haykin, S., 2009. Cubature kalman filters. IEEE Trans. Autom. Control 54 (6), 1254–1269.
Bai, L., Yao, L., Li, C., Wang, X., Wang, C., 2020. 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.
Chen, C., Xiong, R., Yang, R., Shen, W., Sun, F., 2019. State-of-charge estimation of lithium-ion battery using an improved neural network model and extended kalman filter. Journal of Cleaner Production 234.
Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.
Deshpande, N., Park, H., Pandey, V., and Yoon, G. (2023). Advancing temporal multimodal learning with physics informed regularization. pages 1–5.
Di, X., Liu, H.X., Ban, X.J., 2016. Second best toll pricing within the framework of bounded rationality. Transport. Res. Part B: Methodol. 83, 74–90.
Errica, F., Bacciu, D., Micheli, A., 2021. Graph mixture density networks. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021), pages 3025–3035.
Folsom, L., Ono, M., Otsu, K., Park, H., 2021. Scalable information-theoretic path planning for a rover-helicopter team in uncertain environments, 18(2): 1–16. Int. J. Adv. Rob. Syst.
Fu, R., Zhang, Z., Li, L., 2016. Using lstm and gru neural network methods for traffic flow prediction. pages 324–328.
Guan, C., Luh, P.B., Michel, L.D., Chi, Z., 2013. Hybrid kalman filters for very short-term load forecasting and prediction interval estimation. IEEE Trans. Power Syst. 28 (4), 3806–3817.
Guo, X., 2013. Toll sequence operation to realize target flow pattern under bounded rationality. Transport. Res. Part B: Methodol. 56, 203–216.
Han, K., Szeto, W.Y., Friesz, T.L., 2015. Formulation, existence, and computation of boundedly rational dynamic user equilibrium with fixed or endogenous user tolerance. Transport. Res. Part B: Methodol. 79, 16–49.
Han, Q., Timmermans, H., 2006. Interactive learning in transportation networks with uncertainty, bounded rationality, and strategic choice behavior: Quantal response model. Transport. Res. Rec.: J. Transport. Res. Board 1964, 27–34.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Huang, A.J., Agarwal, S., 2022. Physics informed deep learning for traffic state estimation: Illustrations with lwr and ctm models. IEEE Open J. Intell. Transport. Syst. 3. 1–1.
Hyndman, R., 2006. Another look at forecast accuracy metrics for intermittent demand. Foresight: Int. J. Appl. Forecast. 4, 43–46.
Jiang, M., Liu, Z., 2023. Traffic flow prediction based on dynamic graph spatial-temporal neural network. Mathematics 11, 2528.
Julier, S.J., Uhlmann, J.K., 1997. New extension of the Kalman filter to nonlinear systems. In: Kadar, I. (Ed.), Signal Processing, Sensor Fusion, and Target Recognition VI, volume 3068. International Society for Optics and Photonics, SPIE, pp. 182–193.
Kumar, S., 2017. Traffic flow prediction using kalman filtering technique. Proc. Eng. 187, 582–587.
Li, F., Feng, J., Yan, H., Jin, G., Jin, D., Li, Y., 2021. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution.
Li, Y., Yu, R., Shahabi, C., Liu, Y., 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: International Conference on Learning Representations (ICLR '18).
Mihaylova, L., Boel, R., Hegyi, A., 2006. An unscented kalman filter for freeway traffic estimation.
Park, H., Darko, J., Deshpande, N., Pandey, V., Su, H., Ono, M., Barkely, D., Folsom, L., Posselt, D., Chien, S., 2022. Temporal multimodal multivariate learning. In: In SIGKDD Conference on Knowledge Discovery and Data Mining.
Park, H., Haghani, A., 2015. Optimal number and location of Bluetooth sensors considering stochastic travel time prediction. Transport. Res. Part C: Emerg. Technol. 55, 203–216.
Revach, G., Shlezinger, N., Ni, X., Escoriza, A., van Sloun, R., Eldar, Y., 2021. Kalmannet: Neural network aided kalman filtering for partially known dynamics.
Rico, J., Barateiro, J., Oliveira, A., 2021. Graph neural networks for traffic forecasting.
Yu, B., Yin, H., Zhu, Z., 2017. Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting.
Zhao, L., Chen, M., Du, Y., Yang, H., Wang, C., 2022. Spatial-temporal graph convolutional gated recurrent network for traffic forecasting.