Multivariate time series forecasting with diffusion kernels: Freeway traffic prediction

Semin Kwak
semin.kwak@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL)

Nikolas Geroliminis
nikolas.geroliminis@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL)

Pascal Frossard
pascal.frossard@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL)

ABSTRACT

Forecasting multivariate time series is challenging as the variables are intertwined in time and space. Estimating such complex spatiotemporal correlations between variables is an important factor in achieving accurate prediction results. This paper shows that this complex relationship can be effectively estimated by a composition of simple linear functions in the case of traffic signals. Particularly, we propose to regularize the prediction model by the structural information of the transportation network with the help of graph heat diffusion kernels. We confirm by extensive experiments that our proposed simple model shows comparable or better forecasting performance to Deep Neural Network-based predictors with fewer free parameters.

CCS CONCEPTS

• **Computing methodologies** → **Bayesian network models**.

KEYWORDS

multivariate time series prediction, spatiotemporal correlation, graph heat diffusion, Bayesian inference

ACM Reference Format:
Semin Kwak, Nikolas Geroliminis, and Pascal Frossard. 2021. Multivariate time series forecasting with diffusion kernels: Freeway traffic prediction. In *MileTS '21: 7th KDD Workshop on Mining and Learning from Time Series, August 14th, 2021, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Multivariate time-series prediction is an important task since many real-life problems can be modeled within this framework, such as weather forecasting, traffic prediction, etc. Since variables are complexly interrelated in time and space, proper modeling greatly influences the predictive performance. For example, in transportation sensor networks, some spatially closed sensors measure similar patterned signals, but others record significantly different data, as shown in Fig. 1(a) and (b). Therefore, sensor B's signal can be utilized to predict sensor A's one as the two signals are well correlated with each other. However, the signal of sensor C is not correlated with that of sensor A and may not contribute to the prediction.

Deep Neural Network (DNN) frameworks have been popularized to predict time series as they can approximate any complicated functions by universal approximation theorem [3], and their parameters can be effectively estimated through back-propagation. Especially, recent studies have prioritized distance-based correlations between sensors by defining signals on graphs to reduce the complexity of DNN architectures [1, 2, 7, 10–13]. Utilizing this information, the signal's spatial relationship can be correlated into a temporal one through the heat propagation kernel (or convolutional filter). By crafting the heat propagation model into DNN architectures, state-of-the-art performance can be achieved for traffic prediction.

This paper shows that even linear models can successfully predict multivariate time series with high accuracy. We integrate the sensors' structural information into existing linear traffic model e.g., [6], inheriting the proper representation of periodicity in the traffic signal. In details, we establish a heat diffusion model from sensors' structural information and assume that the traffic signals' change over time is regularized by this heat diffusion model. After that, we update the model with data and estimate the required parameters optimally through Bayesian inference [8]. The estimation process is relatively fast as most parameter estimation is performed by analytic calculations.

Predictors based on this model showed comparable performance with a much shorter learning time than state-of-the-art models based on Deep Networks. Especially, the proposed model shows great long-term prediction performance as the model captures well the periodicity of traffic signals. Since the proposed model requires minimal hyper-parameter tuning (most parameters are optimized through Bayesian inference), it might be applied to other daily periodic graph signal prediction problems easily (e.g., weather forecasting, daily energy consumption prediction).

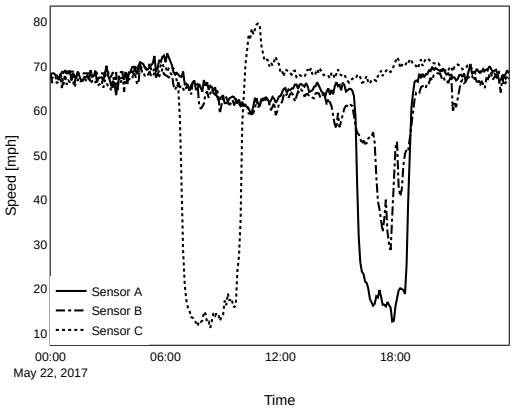
2 DATA MODEL

2.1 Graph signal

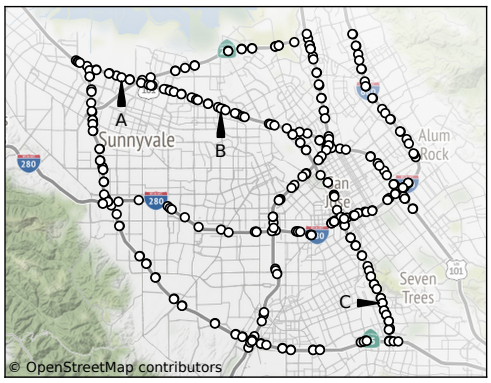
We start with modeling a transportation network using a graph. We define an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; $\mathcal{V}$ is a set of nodes where each $v \in \mathcal{V}$ denotes a node (sensor) on the network; $\mathcal{E}$ is a set of edges where each of the edges connects two nodes. We define a signal on the nodes of the graph with a traffic feature (in this paper, for instance, speed), which is expressed as a vector $\mathbf{x}_t^d \in \mathcal{R}^N$ of a day d and time t, where the constant N is the number of nodes. Therefore, the vector $\mathbf{x}_t^d$ represents a snapshot of speeds at a particular time and day. Especially, we express the day index on

MileTS '21, August 14th, 2021, Singapore

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456



(a) Signals on different sensors



(b) PEMS-BAY network

Figure 1: A transportation sensor network in California and signals of three different sensors on the network. Although the sensors B and C are close to each other in distance, as these sensors are located in different freeways (opposite direction), two traffic signals from these sensors show very different patterns.

the vector representation to exploit the periodicity of traffic signals later.

We also define a weight matrix that contains all edge weights between node v_i and v_j using a Gaussian kernel weighting function with a threshold constant κ: $[\mathbf{W}]_{i,j} = e^{-\frac{\text{dist}^2(i,j)}{\sigma^2}}$ if $\text{dist}(i,j) \leq \kappa$, 0 otherwise. The function $\text{dist}(i,j)$ denotes the shortest travel distance on $\mathcal{G}$ between the node v_i and v_j:

$$\text{dist}(i,j) = \min\{\text{dist}(v_i \to v_j), \text{dist}(v_j \to v_i)\}, \tag{1}$$

where the function $\text{dist}(v_i \to v_j)$ represents the shortest travel distance from node v_i to node v_j. As the graph $\mathcal{G}$ is undirected, the weight matrix is a symmetric matrix, i.e., $\mathbf{W}^T = \mathbf{W}$.

The constants σ and κ are the kernel width and the distance threshold. If the kernel width is large, the correlation of a pair of nodes becomes strong (close to one) even though the shortest travel distance between the two nodes is large. On the other hand, the smaller the threshold is, the sparser the weight matrix is.

2.2 Heat diffusion model on graphs

The graph heat diffusion model [5] explains how each vertex propagates its heat to its neighbors on the graph over time. As congestion evolves from a location to its neighbor over time, we can express the change of traffic features by the heat diffusion model.

The kernel on graphs that supports the heat diffusion model is introduced by [5]:

$$\mathbf{H}^{\mathcal{G}}(\tau) = e^{-\tau \mathbf{L}(\mathcal{G})}, \tag{2}$$

where the constant τ denotes the diffusion period and the matrix $\mathbf{L}(\mathcal{G})$ is the Laplacian of a graph $\mathcal{G}$ that is $\mathbf{L}(\mathcal{G}) = \text{diag}(\mathbf{W1}) - \mathbf{W}$.

Therefore, with the heat diffusion kernel, we can describe how a traffic signal is diffused through the graph $\mathcal{G}$ as follows:

$$\tilde{\mathbf{x}}^d_{t+1}(\tau) = \mathbf{H}^{\mathcal{G}}(\tau)\mathbf{x}^d_t. \tag{3}$$

We call the vector $\tilde{\mathbf{x}}^d_{t+1}(\tau)$ the internally diffused signals from $\mathbf{x}^d_t$ by the diffusion period τ on the graph $\mathcal{G}$ over one incremental time step.

We also define a convex combination of the heat diffusion kernels of K different predetermined diffusion periods with a set $\mathcal{T} = \{\tau^{(0)}, \tau^{(1)}, \cdots, \tau^{(K-1)}\}$[1] as

$$\mathbf{H}^{\mathcal{G}}(\mathcal{T}) = \sum_{\tau \in \mathcal{T}} \pi^{(\tau)} \mathbf{H}^{\mathcal{G}}(\tau), \tag{4}$$

where $\sum_{\tau \in \mathcal{T}} \pi^{(\tau)} = 1$.

2.3 Dynamic linear model (DLM) with heat diffusion kernels

In this subsection, we describe how to model the time series of traffic signals with a linear model. In particular, we describe a model that can be applied to a non-stationary time series while expressing the seasonality of a traffic signal well. In [6], authors defined a state equation of traffic in a small-scale transportation network as temporally localized linear models called Dynamic linear model (DLM) as follows:

$$\mathbf{x}^d_{t+1} = \mathbf{H}_t \mathbf{x}^d_t + \mathbf{n}^d_t, \forall t \in [0, T-1]\,. \tag{5}$$

The first time index ($t = 0$) corresponds to the beginning of a day (midnight in our work), and the last index ($t = T - 1$) refers to the end of the day. Each entry of the noise vector $\mathbf{n}^d_t \in \mathcal{R}^N$ is assumed to be an independent and identically distributed (i.i.d.) random variable, which follows a Gaussian distribution $\mathcal{N}(0, \alpha_t^{-1})$. Here the precision parameter α_t explains how precisely a data pair $(\mathbf{x}^d_t, \mathbf{x}^d_{t+1})$ fits to the model.

In this paper, we embed heat diffusion kernels into DLM to exploit topological information of the transportation network. The key idea is to express the transition matrix as a small variant from

[1] We predetermine the set $\mathcal{T}$ with two diffusion periods τ_0 and τ_∞ that correspond two extreme cases ($\tau \to 0$ and $\tau \to \infty$), respectively. In practice, we set τ_0 as the biggest one that satisfies $\left\|\mathbf{H}^{\mathcal{G}}(\tau) - \mathbf{I}\right\|_2 < \epsilon$ and τ_∞ as the smallest one that satisfies $\left\|\mathbf{H}^{\mathcal{G}}(\tau) - 1/N\mathbf{1}\mathbf{1}^T\right\|_2 < \epsilon$. After that, we define $\mathcal{T} = \text{logspace}(\tau_0, \tau_\infty, K)$, where the function returns K evenly spaced numbers on a log scale from τ_0 to τ_∞.

a mixture of diffusion kernels. We decompose the transition matrix into the time-variant internal diffusion and residual as follows:

$$\mathbf{H}_t = \mathbf{H}_t^{\mathcal{G}}(\mathcal{T}) + residual \tag{6}$$

Here, the time dependent internal transition matrix ($\mathbf{H}_t^{\mathcal{G}}(\mathcal{T})$) can be safely defined as in Eq. (4) by substituting the time-invariant parameter $\pi^{(\tau)}$ for the time-variant one $\pi_t^{(\tau)}$. The internal diffusion matrix represents how the current signal $\mathbf{x}_t^d$ diffuses through the transportation network (endogenous) whereas the residual represents how much the traffic situation is getting better or worse in the next time step based on the current signal (exogenous). With this interpretation, we model the prior distribution of the transition matrix as:

$$f(\mathbf{H}_t|\gamma_t, \Pi_t, \mathcal{G}) = \mathcal{N}\left(\mathbf{H}_t^{\mathcal{G}}(\mathcal{T}), \gamma_t^{-1}\right), \tag{7}$$

where the precision parameter γ_t represents how precisely the diffusion matrix explains the transition matrix and $\Pi_t = \{\pi_t^{(\tau)}|\tau \in \mathcal{T}\}$.

3 MODEL INFERENCE

3.1 Inference of the transition matrix

We infer the transition matrix by maximizing its posterior distribution:

$$\hat{\mathbf{H}}_t = \underset{\mathbf{H}_t}{\operatorname{argmax}}\, f(\mathbf{H}_t|\mathbf{X}_t, \mathbf{X}_{t+1}, \alpha_t, \gamma_t, \Pi_t, \mathcal{G}), \tag{8}$$

where $\mathbf{X}_t = \begin{pmatrix}\mathbf{x}_t^0 & \mathbf{x}_t^1 & \cdots & \mathbf{x}_t^{m-1}\end{pmatrix}$ and it is proportional to the product of the prior and the likelihood by Bayes' rule:

$$\text{Posterior dist.} \propto f(\mathbf{H}_t|\gamma_t, \Pi_t) f(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{H}_t, \alpha_t). \tag{9}$$

Maximizing the posterior distribution can be interpreted as balancing between the prior and likelihood of the transition matrix. For example, if there is no topological information about sensors, the transition matrix should be inferred by considering the training dataset only. In this case, we can set the prior distribution as a uniform distribution meaning that there is no strong preference for a particular value of the transition matrix and the most probable transition matrix becomes the maximum likelihood solution [6]:

$$\begin{aligned}\hat{\mathbf{H}}_t|\text{No topological info.} := \bar{\mathbf{H}}_t &= \underset{\mathbf{H}_t}{\operatorname{argmax}}\, f(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{H}_t, \alpha_t) \\ &= \mathbf{X}_{t+1}\mathbf{X}_t^T(\mathbf{X}_t\mathbf{X}_t^T)^{-1}.\end{aligned} \tag{10}$$

On the other hand, if we do not have any measurements, the most probable transition matrix should be the maximizer of the prior distribution:

$$\hat{\mathbf{H}}_t|\text{No measurements} = \underset{\mathbf{H}_t}{\operatorname{argmax}}\, f(\mathbf{H}_t|\gamma_t, \Pi_t) = \mathbf{H}_t^{\mathcal{G}}(\mathcal{T}). \tag{11}$$

The most probable transition matrix is a function of these two matrices:

$$\begin{aligned}\hat{\mathbf{H}}_t &= g(\bar{\mathbf{H}}_t, \mathbf{H}_t^{\mathcal{G}}(\mathcal{T})) \\ &= (\bar{\mathbf{H}}_t\alpha_t\mathbf{U}_t\mathbf{\Lambda}_t + \mathbf{H}_t^{\mathcal{G}}(\mathcal{T})\gamma_t\mathbf{U}_t)(\alpha_t\mathbf{\Lambda}_t + \gamma_t\mathbf{I})^{-1}\mathbf{U}_t^T,\end{aligned} \tag{12}$$

with the eigendecomposition of $\mathbf{X}_t\mathbf{X}_t^T = \mathbf{U}_t\mathbf{\Lambda}_t\mathbf{U}_t^T$.

3.2 Inference of other parameters

For the next step, we infer parameters α_t, γ_t, and Π_t. Similarly to inferring the most probable transition matrix, we infer the most probable α_t, γ_t, and Π_t by maximizing the following posterior distribution:

$$\hat{\alpha}_t, \hat{\gamma}_t, \hat{\Pi}_t = \underset{\alpha_t, \gamma_t, \Pi_t}{\operatorname{argmax}}\, f(\alpha_t, \gamma_t, \Pi_t|\mathbf{X}_{t+1}, \mathbf{X}_t). \tag{13}$$

Setting the prior distribution ($f(\alpha_t, \gamma_t, \Pi_t)$) as a uniform distribution based on the assumption that there is no preference for a certain value for these parameters before inferring, the objective changes to maximize *evidence* [8] since

$$\begin{aligned}f(\alpha_t, \gamma_t, \Pi_t|\mathbf{X}_{t+1}, \mathbf{X}_t) &\propto f(\mathbf{X}_{t+1}|\mathbf{X}_t, \alpha_t, \gamma_t, \Pi_t) f(\alpha_t, \gamma_t, \Pi_t) \\ &\propto f(\mathbf{X}_{t+1}|\mathbf{X}_t, \alpha_t, \gamma_t, \Pi_t).\end{aligned} \tag{14}$$

and the evidence is

$$\begin{aligned}f(\mathbf{X}_{t+1}|\mathbf{X}_t, \alpha_t, \gamma_t, \Pi_t) &= \int f(\mathbf{X}_{t+1}|\mathbf{X}_t, \mathbf{H}_t, \alpha_t) f(\mathbf{H}_t|\gamma_t, \Pi_t) d\mathbf{H}_t \\ &= \mathcal{N}(\mathbf{H}_t^{\mathcal{G}}(\mathcal{T})\mathbf{X}_t, \alpha_t^{-1}\mathbf{I} + \gamma_t^{-1}\mathbf{X}_t^T\mathbf{X}_t).\end{aligned} \tag{15}$$

Therefore, we infer the most probable hyper-parameters by maximizing the log-evidence with gradient-based algorithms:

$$\begin{aligned}\underset{\alpha_t, \gamma_t, \Pi_t}{\text{maximize}} \quad & \log \mathcal{N}(\mathbf{H}_t^{\mathcal{G}}(\mathcal{T})\mathbf{X}_t, \alpha_t^{-1}\mathbf{I} + \gamma_t^{-1}\mathbf{X}_t^T\mathbf{X}_t) \\ \text{subject to} \quad & 0 \le \pi_t^{(\tau)} \le 1\ \forall\tau \in \mathcal{T},\ 0 < \alpha_t,\ 0 < \gamma_t, \\ & \sum_{\tau\in\mathcal{T}} \pi_t^{(\tau)} = 1.\end{aligned} \tag{16}$$

We emphasize that inferring the parameters by maximizing the evidence leads to avoid the transition matrix to over-fit to either data measurements or prior information. In Eq. (15) we calculate the evidence by *marginalizing* the transition matrix. In other words, we set the transition matrix as a random variable instead of fixing it as a representative value, e.g., maximum likelihood estimator. Noting that these parameters determine the contributions of measurements ($\bar{\mathbf{H}}_t$) and priors ($\mathbf{H}_t^{\mathcal{G}}(\mathcal{T})$) when the transition matrix is estimated in Eq. (12), the marginalization process automatically penalizes for the transition matrix to become to be an extreme case [8].

4 PREDICTION OF TRAFFIC FEATURES

We infer the probability density function of the signal $\mathbf{x}_{t+h}^d$

$$f(\mathbf{x}_{t+h}^d|\mathbf{x}_t^d, \mathbf{x}_{t-1}^d \cdots, \mathcal{G}), \tag{17}$$

where the time index t and $t+h$ represent the current time and the future time index (h-steps ahead) respectively that we want to predict. In the expression, the probability density function is conditioned by the signals $\{\mathbf{x}_t^d, \mathbf{x}_{t-1}^d, \cdots\}$ and the graph $\mathcal{G}$ that represents a set of measurements and prior structural information, respectively.

In reality, it is common to limit the number of measurements to a fixed-sized one in a training set. In addition to the training set that contains measurements apart from the day to be predicted, it is crucial to keep measurements just before t as the temporal correlation is strong when the time difference is small. As a result,

Table 1: RMSE of different methods for PEMS-BAY dataset.

Model	Horizon		
	15 min	30 min	60 min
FC-LSTM [4]	4.19	4.55	4.96
DCRNN [7]	2.95	3.97	4.74
Graph WaveNet [10]	**2.74**	**3.70**	4.52
Proposed	2.91	3.77	**4.44**

we estimate the density function that is conditioned by a training set, the p-most recent measurements, and the graph $\mathcal{G}$:

$$f(\mathbf{x}^d_{t+h}|\mathbf{x}^d_t, \mathbf{x}^d_{t-1}, \cdots, \mathbf{x}^d_{t-(p-1)}, \mathbf{X}_{0:T-1}, \mathcal{G}), \tag{18}$$

where the training set $\mathbf{X}_{0:T-1}$ contains signals from $t = 0$ to $t = T-1$ of multiple days $d \in [0, m-1]$. The dynamic linear model further simplifies the distribution (18) as $f(\mathbf{x}^d_{t+h}|\mathbf{x}^d_t, \mathbf{X}_{t:t+h}, \mathcal{G})$ because of the temporal locality of the model.

We define a predictor $\mathbf{x}^d_{t+h|t}$ at the time step t for the horizon h as the maximizer of the probability density function

$$\mathbf{x}^d_{t+h|t} := \underset{\mathbf{x}^d_{t+h}}{\operatorname{argmax}} f(\mathbf{x}^d_{t+h}|\mathbf{x}^d_t, \mathbf{X}_{t:t+h}, \mathcal{G}). \tag{19}$$

In other words, we define the predictor $\mathbf{x}^d_{t+h|t}$ as the most probable $\mathbf{x}^d_{t+h}$ based on the current measurement vector $\mathbf{x}^d_t$, the training set $\mathbf{X}_{t:t+h}$, and the graph $\mathcal{G}$.

The posterior distribution $f(\mathbf{x}^d_{t+h}|\mathbf{x}^d_t, \mathbf{X}_{t:t+h}, \mathcal{G})$ is a Gaussian distribution that has the mean vector $\hat{\mathbf{H}}_{t+h-1} \cdots \hat{\mathbf{H}}_t \mathbf{x}^d_t$ assuming $f(\mathbf{H}_t|\mathbf{X}_t, \mathbf{X}_{t+1}, \alpha_t, \gamma_t, \Pi_t, \mathcal{G}) = \delta(\mathbf{H}_t - \hat{\mathbf{H}}_t)$, where the Dirac delta function $\delta(x) = 1$ when $x = 0$ and $\delta(x) = 0$, otherwise. The most probable transition $\hat{\mathbf{H}}_t$ is the maximizer of the posterior distribution $f(\mathbf{H}_t|\cdot)$. Since the mean value of a Gaussian distribution maximizes the distribution, the optimal predictor is

$$\mathbf{x}^d_{t+h|t} = \hat{\mathbf{H}}_{t+h-1} \cdots \hat{\mathbf{H}}_t \mathbf{x}_t := \hat{\mathbf{H}}_{t+h-1 \leftarrow t} \mathbf{x}^d_t. \tag{20}$$

Therefore, the most probable signal $\mathbf{x}^d_{t+h}$ is the successive propagation of the current measurement vector $\mathbf{x}^d_t$ through the most probable transition matrices.

5 EXPERIMENTS

5.1 Settings

The PEMS-BAY dataset is used as a benchmark to compare with other state-of-the-art models [7, 9]. This data set consists of data measured from 325 sensors (Fig. 1(b)) on the freeways of San Francisco Bay area. The training and test dataset were constructed in the same way as [7, 9] to achieve a fair comparison.

5.2 Other methods

5.2.1 FC-LSTM (Fully Connected Long Short-Term Memory). This model has been used as a representative reference for time-sequence modeling in deep learning [4]. The RMSE score for PEMS-BAY dataset is retrieved from [7].

5.2.2 DCRNN (Diffusion Convolution Recurrent Neural Network). The work in [7] constructed a successful predictor by extracting the signal's spatial features from the underlying graph structure. The model extracts spatial and temporal features of multivariate time series by diffusion convolutional and RNN layers, respectively.

5.2.3 Graph WaveNet. The authors of [10] suggested a model that extracts temporal features by dilated convolutional layers rather than an RNN structure, which shows better prediction performance than the DCRNN with a shorter learning time.

5.3 Results

Table 1 shows the RMSE of each model and our proposed method. We confirm that the performance of the proposed method reaches that of state-of-the-art methods based on a complex deep learning architecture. It even performs better for long-term prediction as our model is based on DLM that explicitly expresses the daily periodicity of traffic signals.

Our proposed method requires lower computational effort compared to the others. Also, it infers the majority of the parameters (N^2) analytically by Eq. (12). The method only requires numerical computation when it solves the optimization problem (16) to infer $K+2$ parameters where $K+2$ is noticeably smaller than N^2. On the other hand, all state-of-the-art methods require heavy numerical computations to train a large number of parameters as they are based on deep-neural-net architectures. As a result, our method successfully infers all parameters at the time scale of minutes with CPU computations, while the others compute at the time scale of hours with GPU computations.

Another advantage of our model compared to the deep-learning-based architectures is that only a small number of parameters should be decided heuristically. This can provide easy scalability to apply our model to other traffic datasets or datasets with similar properties to traffic data (daily periodicity). For example, in our model, the parameters to be determined before training are the threshold constant κ, the kernel width σ to build a proper graph, and the number of diffusion processes K to determine how many diffusion processes should be mixed. We empirically choose the constants κ and σ such that the corresponding graph $\mathcal{G}$ is a k-vertex-connected graph with a small number k. For the number of diffusion processes K, we set $K = 5$ for the PEMS-BAY dataset but the prediction performance is not sensitive to the parameter (± 0.01 minutes changes of the RMSE score from $K = 3$ to $K = 7$).

6 CONCLUSION

In this paper, we show that even linear models are enough to represent complex spatio-temporal correlations of multivariate time series when prior structural information is properly integrated into the models. We integrate topological information of the sensor network into the model by assuming that the parameters in the model are supported by the mixture of diffusion kernels with uncertainties. We exploit the Bayesian inference to optimally determine the parameters that characterize the distribution of diffusion processes and the importance of measurements against prior information. Most importantly, the proposed method reaches accurate prediction at the level of state-of-the-art methods with less computational effort. It particularly shows excellent performance in long-term

predictions by exploiting DLM's periodicity modeling. Our method can be applicable for predicting graph signals having daily patterns such as weather or energy consumption.

REFERENCES

[1] Cen Chen, Kenli Li, Sin G Teo, Xiaofeng Zou, Kang Wang, Jie Wang, and Zeng Zeng. 2019. Gated residual recurrent graph neural networks for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 485–492.

[2] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang. 2019. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems* (2019).

[3] George Cybenko. 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems* 2, 4 (1989), 303–314.

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[5] Risi Imre Kondor and John Lafferty. 2002. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning*, Vol. 2002. 315–22.

[6] Semin Kwak and Nikolas Geroliminis. 2020. Travel time prediction for congested freeways with a dynamic linear model. *IEEE Transactions on Intelligent Transportation Systems* (2020).

[7] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*.

[8] David JC MacKay. 1992. Bayesian interpolation. *Neural computation* 4, 3 (1992), 415–447.

[9] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 1907–1913.

[10] Bingbing Xu, Huawei Shen, Qi Cao, Yunqi Qiu, and Xueqi Cheng. 2018. Graph Wavelet Neural Network. In *International Conference on Learning Representations*.

[11] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 3634–3640.

[12] Chenhan Zhang, JQ James, and Yi Liu. 2019. Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting. *IEEE Access* 7 (2019), 166246–166256.

[13] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems* (2019).