Probabilistic Forecasting of Long-Term Origin-Destination Demands: An Interpretable Bayesian Framework for Periodicity and Residual Learning

Zihan Wan, Zhenjie Zheng, Wei Ma *Member, IEEE*

***Abstract*—Origin-Destination (OD) demands are the backbone of traffic management and urban planning, serving as the fundamental input to numerous mobility applications. Existing studies mainly focus on modeling and predicting OD demands in the short term, while studies for long-term OD demand forecasting are limited. In particular, long-term OD demand prediction provides insights into the evolving spatiotemporal distribution of travel demands over extended periods, informing service scheduling and resource allocation that short-term forecasting cannot adequately support. One of the most distinguishing characteristics of OD demand time series is their inherent periodicity, punctuated by intermittent fluctuations. In view of this, we propose a novel interpretable Bayesian framework for long-term OD demand forecasting, which integrates both periodic patterns and transient fluctuations into a unified predictive model. By leveraging stochastic variational inference (SVI) and a modified tensor decomposition approach, the posterior distributions of the periodic and residual components in OD demands are formally derived. This enables the generation of both point-valued predictions and corresponding prediction intervals, which effectively quantify the predictive uncertainty and enhance model reliability. To validate the effectiveness of our proposed framework, we conduct experiments on real-world OD datasets. The results show that our model consistently outperforms state-of-the-art deep learning approaches under diverse scenarios. This underscores the effectiveness and robustness of our model in addressing the challenges of long-term multiple OD demand forecasting.**

***Index Terms*—OD demand forecasting, Long-term forecasting, Stochastic variational inference, Probabilistic inference, Tensor decomposition**

I. Introduction

CITIES have undergone rapid growth and development, and consequently, this expansion has led to more complicated patterns of urban movement. In response to the evolving travel demands, it is a critical challenge to provide effective human-centric mobility services for intelligent transportation systems. The cornerstone of creating an effective transportation system lies in the ability to understand and predict public travel demands, which are driven by a complex mixture of economic activities and travel motivations. In this context, the analysis of origin-destination (OD) demands plays a crucial role in strategic planning and traffic management. Therefore, a deep dive into OD demand prediction is indispensable for urban planners and traffic management authorities to build efficient and sustainable transportation systems.

Early studies on travel demand modeling typically adopt parametric spatial frameworks whose parameters are estimated by available data, and the estimated model is then applied to the prediction of population-level travel demand [1]. Consequently, their methods heavily rely on the collection of reliable data, often requiring extensive household travel surveys. With the advent of modern sensing and communication technologies, various traditional machine learning methods [2]–[4] have been developed to estimate OD demand using data from diverse sources. The OD demand data sources primarily include the Global Positioning System [5], [6], e-hailing services [7]–[9], and automated fare collection (AFC) systems in public transport networks [10]–[14]. The abundance of OD demand data captures the intricate spatial and temporal interactions inherent in human mobility patterns, allowing machine learning methods to make reliable predictions of future travel demands. However, these methods may produce suboptimal results when applied to large-scale data. Additionally, their performance often relies on manual feature engineering, which can be ineffective when handling unstructured data [3], [15], [16].

Recently, numerous studies have leveraged advanced deep learning techniques to uncover the spatial and temporal dependencies of OD demands from large-scale datasets. Compared to traditional machine learning methods, deep learning models can learn complex nonlinear relationships through automatic feature extraction and activation functions, while efficiently handling large-scale and unstructured data [17]–[19]. Given the sequential nature of time series data, a significant portion of OD demand forecasting research has adopted Recurrent Neural Network (RNN) architectures, particularly Long Short-Term Memory (LSTM) variants [20]–[23].

Although the existing deep learning models have shown promising performance in short-term (e.g., hours) prediction, long-term (e.g., days and weeks) OD demand forecasting remains relatively underexplored in the research community. Importantly, the fundamental differences between short-term and long-term prediction may cause existing deep learning methods to overlook the unique challenges inherent in long-term time series forecasting. First, short-term models are typically designed to capture immediate temporal dependencies and high-frequency variations [13]. Such prediction relies on recent observations and overlooks the inherent periodicity (e.g., daily or weekly periodicity) in long-term time series. Second, directly applying short-term models to long-term forecasting increases susceptibility to error accumulation [24]. When predictive models generate outputs over extended

Zihan Wan is with the Department of Civil and Environmental Engineering and the Research Institute for Sustainable Urban Development at the Hong Kong Polytechnic University, Hong Kong, SAR, China (E-mail: 22062327r@connect.polyu.hk).

Zhenjie Zheng is with the Department of Civil and Environmental Engineering at the Hong Kong Polytechnic University, Hong Kong, SAR, China (E-mail: zzj17.zheng@polyu.edu.hk).

Wei Ma is with the Department of Civil and Environmental Engineering and the Research Institute for Sustainable Urban Development at the Hong Kong Polytechnic University, Hong Kong, SAR, China, and also with the Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, Guangdong 518057, China (E-mail: wei.w.ma@polyu.edu.hk).

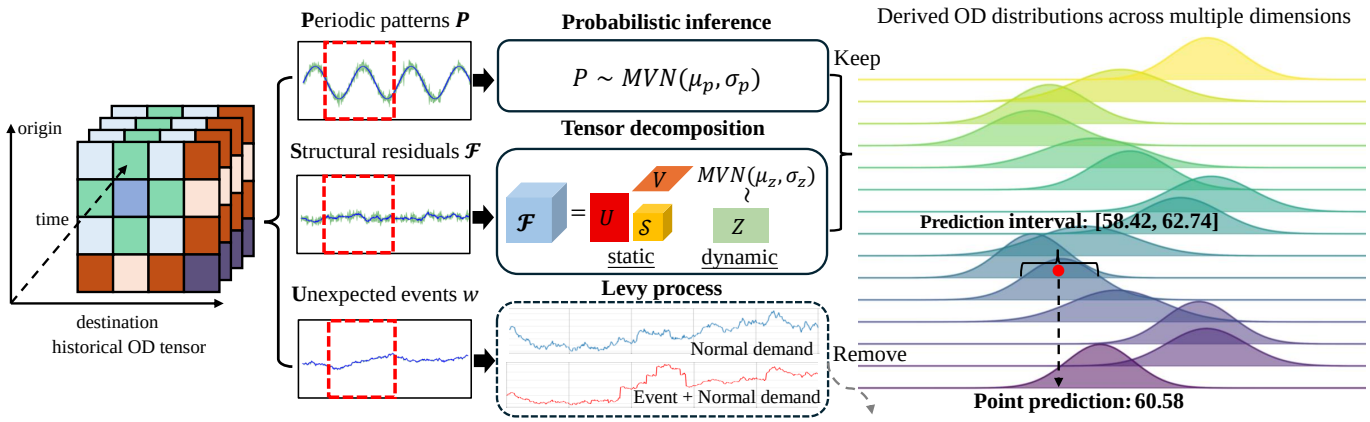


Fig. 1. An overview of our proposed framework for forecasting long-term Origin-Destination demands.

horizons, small prediction errors in earlier steps can propagate and amplify over time, leading to notable inaccuracies. Finally, short-term models mainly focus on generating point predictions and fail to account for the predictive uncertainty. This is partly due to the limited data available in short time windows, which makes it difficult to support reliable probabilistic estimation. In contrast, long-term forecasting over extended horizons naturally requires uncertainty quantification, such as prediction intervals, to account for the accumulation of errors and growing uncertainty over time. Additionally, the limited interpretability of most deep learning models for OD demand prediction has raised concerns, as their black-box nature hinders the extraction of meaningful insights that can inform management strategies and policy decisions.

To address the above challenges, we propose an interpretable Bayesian framework, named Periodicity-Residual Stochastic Variational Inference (P-R SVI), to enable accurate and reliable long-term OD forecasting (see Fig. 1). Specifically, P-R SVI is built upon stochastic variational inference (SVI) and explicitly models the periodic and residual components of OD demands, which simultaneously capture recurring regularities and stochastic fluctuations. The posterior distributions of both periodic and residual components are derived to generate point predictions and associated prediction intervals, offering a probabilistic characterization of predictive uncertainty and enhancing the robustness of the model. Importantly, the residual component is not treated as a single entity but is disentangled into several latent structural factors through a modified tensor decomposition approach. Such structured treatment enables the model to capture systematic deviations and interpretable spatiotemporal variation patterns, which reduces residual autocorrelation and mitigates error accumulation in long-term forecasting. Furthermore, the interpretable nature of the proposed framework provides practical insights for policymakers. To evaluate the effectiveness of our model, extensive experiments are conducted on two real-world OD datasets from New York City and San Francisco. Results show that the proposed P-R SVI consistently outperforms state-of-the-art deep learning approaches across diverse scenarios.

The main contributions of this research are summarized as follows:

- We propose an interpretable Bayesian framework for accurate and reliable long-term OD forecasting, which captures both periodic patterns and stochastic residuals in historical OD demand;
- By deriving the posterior distributions of the periodic and residual components, we generate both point predictions and prediction intervals, which provide a probabilistic characterization of prediction accuracy and uncertainty;
- The residual component is decomposed into latent structural factors using tensor decomposition, enabling the model to capture interpretable spatiotemporal patterns, reduce residual autocorrelation, and mitigate error accumulation in long-term forecasting; and
- We use two real-world OD datasets from New York City and San Francisco to conduct the numerical experiments. Results demonstrate the superiority of P-R SVI over other deep learning-based approaches in terms of both prediction accuracy and uncertainty quantification.

II. Related Work

In this section, we briefly review related studies on modeling and forecasting of (multivariate) time series, especially in the applications of OD demand forecasting. This section is divided into three sections: Section II-A summarizes the literature on OD demand prediction; Section II-B focuses on probabilistic forecasting methods; and Section II-C discusses the interpretable deep learning approaches.

A. OD Demand Prediction

Before the widespread adoption of deep learning models, traditional machine learning approaches dominated the field of traffic forecasting. These methods include Kalman filter [2], [25], [26], matrix/tensor factorization [3], [4], [27], [28], Bayesian network [29], and Gaussian process [30]. However, the application of traditional machine learning methods becomes infeasible due to the explosive growth of data, which requires a significantly large number of parameters. Besides, they fall short in capturing non-linear temporal correlations and spatial interactions in multivariate time series. With the success of neural networks in computer vision and natural

language processing, researchers have developed various deep learning models to handle the task of OD demand forecasting. Such forecasting has been explored across diverse transportation modes, including road networks [31], [32], rail transit systems [20], [23], [31], [33], [34], taxis [21], [22], [35], [36], and ride-hailing services [7], [9]. In addition, to enhance the predictive accuracy of OD demands in non-Euclidean spaces, researchers have incorporated similarity or adjacency structures within road and public transit networks, where origins and destinations are represented as nodes. Accordingly, various adjacency relations and corresponding graph convolution operations have been defined to capture spatial correlations and propagate information across neighboring nodes [7], [9], [32], [33], [36]–[38]. However, most of these studies focus on short-term OD demand forecasting for the future 15 or 30 minutes, except [36]. As discussed earlier, short-term prediction models are not directly applicable to long-term OD demand forecasting.

B. Probabilistic Forecasting

For the problem of long-term OD demand prediction, the uncertainty estimation is crucial for many decision-making processes in transportation systems, such as scheduling and fleet sizing. In addition to the predicted expected values as point estimates, quantifying uncertainty in the forecasts is also desired, which often requires probabilistic modeling. Typically, probabilistic modeling approaches can be broadly classified into two categories: parametric and non-parametric models. Depending on the objective function, parametric models can be further divided into Bayesian and mean-variance estimation (MVE) models. MVE obtains the optimized parameters, such as mean and variance, of a pre-specified distribution by minimizing negative log-likelihood [39]–[41]. Wang et al. [41] also examined the combination of various probability density assumptions and graph convolution operations. In Bayesian models, the goal is to obtain updated posterior distributions of latent variables given observations, and the resulting posteriors can be used in forecasting [42]. Fused with deep learning techniques, a large number of probabilistic forecasting frameworks are based on the essence of Variational Auto-Encoder (VAE) [43], which discovers independent latent factors in unsupervised learning. Various deep learning architectures, such as transformers [44] and graph convolutional networks [45], [46], are combined with VAE to achieve probabilistic forecasting by sampling from the latent representation space and feeding the latent samples into the decoder [47], [48]. Regarding the non-parametric models, they do not impose any specific distributional forms or assumptions. Instead, they quantify the uncertainty in the forecasts by directly estimating prediction intervals (PIs) [49]–[51]. For example, Rodrigues et al. [51] jointly estimate conditional expectations and quantiles by minimizing a tilted loss through a two-dimensional convolutional LSTM architecture [52].

C. Interpretable Deep Learning

Despite the satisfactory prediction performance, the black-box nature of deep learning methods is still controversial due to the lack of interpretability. Hence, the research community is devoted to the exploration of how to incorporate explainability into deep learning models, especially in the domain of transportation planning and management, since a model's explainable implications may carry more significance than prediction accuracy. In the following, we classify the deep learning-based solutions to travel demand forecasting based on the model formulation. Up to now, the majority of deep learning demand forecasting models are post hoc, in which explainability is dependent on the successful training of deep learning components to extract seemingly reasonable patterns [53]–[56]. However, the interpretation of the extracted features is susceptible to subjective bias, rendering the universal agreement on the explainability of a deep learning model an almost impossible mission. To understand the contribution of each input variable to the prediction, heat maps are generated using the Layer-wise Relevance Propagation technique [57] so that human experts could assess whether predictions by artificial neural networks are intuitively reasonable. In contrast, a priori methods explicitly combine prior information with deep learning, where domain knowledge is presented in an explainable parametric model. Based on whether the parametric models and deep learning modules are integrated or separated in the training process, the class of a priori methods could be further divided. The separated a priori methods are often composed of two consecutive stages. In the first stage, an empirical and explainable model is formulated, which accomplishes the prerequisite for the second-stage deep learning model to improve demand prediction accuracy. The explainable model appears in various forms, depending on the proposed functionality. For example, multiple linear regression acts as a prediction baseline [58] so that the LSTM framework can learn the residue between the ground truth and baseline; empirical mode decomposition is concatenated with a neural network or support vector machine to detect intrinsic mode functions of multiple frequencies underlying time series data [59], [60]. As for integrated a priori methods, the integrated training of explainable a priori models and deep learning models is scarce in the current literature.

D. Summary of the Literature

To summarize, existing studies on OD demand forecasting have primarily focused on short-term prediction using deep learning models, which are not specifically designed for long-term OD demand forecasting. These models typically rely on point-valued predictions and offer limited interpretability due to their black-box nature. In this research, we adopt a stochastic variational inference framework, where latent variables represent explainable periodic patterns, spatial and temporal factors, and how they interact. The proposed framework removes the constraint of assuming a single distribution family, allowing for greater flexibility in modeling complex OD demand patterns. We also employ a hierarchical structure that supports multi-modality and long-tail behavior, enabling the representation of diverse demand behaviors. By estimating the posterior distributions of latent variables, prediction intervals are obtained through posterior sampling rather than

direct computation of conditional quantiles, which provides a more robust characterization of uncertainty. Moreover, the interpretable modeling of periodic components offers practical insights for policymakers and serves as a reliable predictive baseline, allowing the less interpretable deep learning components to focus on learning the residual variations.

III. Preliminaries

In this section, we present the notation used throughout the paper and provide a formal description of the problem. For OD data, we denote the set of origins by $\mathcal{O}$ with size N_o. Similarly, the set of destinations is denoted as $\mathcal{D}$ with size N_d. The sets of origins and destinations are defined by traffic analysis zones (TAZs) for vehicular travel demand analysis [61] and bus/metro stations for public mobility pattern analysis [20], [34], [62]. Let T denote the number of time intervals in historical OD demand data. Based on these definitions, the observed OD data are naturally represented as a three-dimensional tensor $\boldsymbol{\mathcal{X}}^{\text{obs}} \in \mathrm{R}^{N_o \times N_d \times T}$, where each entry $\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t}$ records the number of trips or passengers from origin $i \in \mathcal{O}$ to destination $j \in \mathcal{D}$ at time interval t.

Let $\tau > 0$ denote the prediction horizon, representing the number of future time intervals to forecast, i.e., we aim to generate predictions of OD demands at time steps $T+1, \ldots, T+\tau$. Our objective is to obtain samples from the predictive distribution of $\boldsymbol{\mathcal{X}}^{\text{pred}} \in \mathrm{R}^{N_o \times N_d \times \tau}$ conditional on the observed OD demands $\boldsymbol{\mathcal{X}}^{\text{obs}}$. To achieve this, we first approximate the true posterior distribution $p_\theta(\boldsymbol{\mathcal{Z}}^{\text{latent}}|\boldsymbol{\mathcal{X}}^{\text{obs}})$ using a variational distribution $q_\phi(\boldsymbol{\mathcal{Z}}^{\text{latent}})$, where the similarity between distributions is measured by the KL divergence: $\mathrm{KL}(q_\phi(\boldsymbol{\mathcal{Z}}^{\text{latent}})||p_\theta(\boldsymbol{\mathcal{Z}}^{\text{latent}}|\boldsymbol{\mathcal{X}}^{\text{obs}}))$. The details of $\boldsymbol{\mathcal{Z}}^{\text{latent}}$ will be in the subsequent sections. Then, samples drawn from $q_\phi(\boldsymbol{\mathcal{Z}}^{\text{latent}})$ are used to calculate posterior samples of the predictive distribution $\boldsymbol{\mathcal{X}}^{\text{pred}}|\boldsymbol{\mathcal{X}}^{\text{obs}}$.

IV. Methodology

A. Overview of the proposed model

In this section, we briefly introduce the core principles and main components of our model. First, for long-term OD demand prediction, periodicity is widely recognized as one of the most prominent characteristics in historical data. For example, large OD demands consistently occur during weekday peak hours, while lower traffic demands are typically observed on weekends. Therefore, in long-term OD demand forecasting, it is natural to find the periodicity present in the historical OD demand time series. Second, although periodic patterns capture the dominant recurring structure in OD demand, consistent residual fluctuations are still observed across OD pairs. Such deviations are not purely random but are instead driven by several latent structural components, including spatial heterogeneity, temporal fluctuations, long-term global trends, and complex spatiotemporal interactions, which should be systematically disentangled and modeled. Finally, unexpected events such as traffic incidents can cause abrupt shifts in OD demands at a system-wide level. Consequently, the transient impact of such unexpected events should be decoupled to enable accurate inference of underlying spatiotemporal correlations.

To summarize, the observed OD demand tensor primarily consists of three components: periodic patterns, residual fluctuations, and unexpected events. Accordingly, we design a three-module P-R SVI framework, including a periodic module, a residual module, and an event detection module, to extract the corresponding key components from the observed OD demands. Since unexpected sudden events tend not to be persistent, only the first two components should be used for the prediction of future OD demands.

Our approach is different from conventional deep learning-based forecast models in three key ways. First, P-R SVI is a sampling-based predictive framework and learns by comparing approximate and true posterior distributions, rather than minimizing Euclidean distance between ground truth and predicted values, which yields calibrated uncertainty and stronger generalizability. Second, instead of batch normalization on mini-batches as in standard deep models, we explain the structure of variance via latent variables that capture spatial heterogeneity, temporal dynamics, and their interactions, providing an interpretable variance decomposition and improved robustness. Third, we introduce a two-stage domain knowledge-informed training scheme: we first learn periodic patterns as domain physical knowledge, then train all modules conditioned on this learned prior. This training scheme stabilizes optimization and consistently improves long-term OD demand forecasting.

In the following sections, we first introduce the formulation and underlying assumptions of each module in Sections IV-B, IV-C, and IV-D. We then present the estimation of the unknown module parameters in Section IV-E. Finally, the calibrated model is employed to predict long-term OD demand, as described in Section IV-F.

B. Formulation of the Periodic Module

The periodic module is designed to capture the latent periodic components $\boldsymbol{\mathcal{P}} \in \mathrm{R}^{N_o \times N_d \times T}$ from observed OD demands $\boldsymbol{\mathcal{X}}^{\text{obs}}$. Specifically, the periodic components $\boldsymbol{\mathcal{P}}^{i_p}$ with period length T_{i_p} are represented as a three-way tensor of size $N_o \times N_d \times T_{i_p}$, where i_p is the period index. In this work, we consider $i_p = 1$ for hourly patterns and $i_p = 2$ for daily patterns. Suppose we are interested in weekly OD demand patterns. In this case, we set $T_1 = 24$ (hours) $\times\, 7$ (days) $= 168$ for the hourly component and $T_2 = 7$ (days) for the daily component. The aggregate periodic components are the sum of periodic components of different period lengths. To enable this summation, the third dimension of the tensor representations must be aligned. In the case of $\boldsymbol{\mathcal{P}}^1$ and $\boldsymbol{\mathcal{P}}^2$, the third dimension of $\boldsymbol{\mathcal{P}}^2$ could be extended from T_2 to T_1 as follows:

$$\boldsymbol{\mathcal{P}}^2_{i,j,:} = \boldsymbol{\mathcal{P}}^2_{i,j,:} \otimes \mathbf{1}, \tag{1}$$

$$\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{P}}^1 \oplus \boldsymbol{\mathcal{P}}^2, \tag{2}$$

where $\otimes$ denotes Kronecker product, $\mathbf{1}$ is a vector of 1's of length $\frac{T_1}{T_2}$, and $\oplus$ denotes element-wise addition. Then the periodic components of OD demands up to arbitrary time T could be derived by repeating the above tensor along the time dimension until the desired time T is reached according to a cycle length of T_{1_p}.

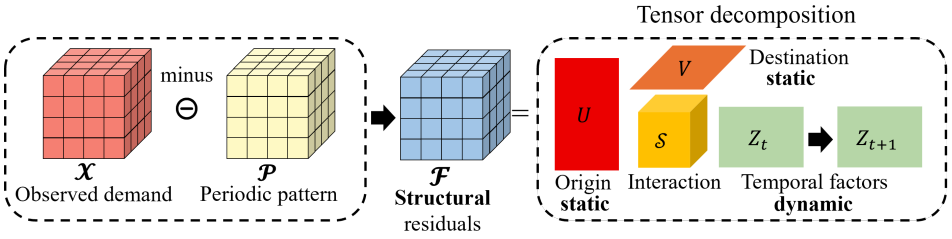


Fig. 2. The Tucker tensor decomposition of structural residuals in OD demands.

For each period i_p, the periodic components $\boldsymbol{\mathcal{P}}^{i_p}$ consist of two elements: origin patterns $\mathbf{P}^{i_p,o}$ and destination patterns $\mathbf{P}^{i_p,d}$. Both origin patterns $\mathbf{P}^{i_p,o} \in \mathrm{R}^{N_o \times T_{i_p}}$ and destination patterns $\mathbf{P}^{i_p,d} \in \mathrm{R}^{N_d \times T_{i_p}}$ are stochastic matrices. Mathematically, $\mathbf{P}^{i_p,o} = [\mathbf{P}^{i_p,o}_{n_o,t_p}]$, where $n_o \in \{1,...,N_o\}$ and $t_{i_p} \in \{1,...,T_{i_p}\}$. The prior distribution of $\mathbf{P}^{i_p,o}_{n_o,t_{i_p}}$ is assumed to independently follow a normal distribution:

$$\mathbf{P}^{i_p,o}_{n_o,t_{i_p}} \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma^{i_p,o^2}). \tag{3}$$

The prior distributions of destination patterns $\mathbf{P}^{i_p,d} \in \mathrm{R}^{N_d \times T_{i_p}}$ are specified in the same manner. Consequently, for each period i_p, the periodic components of OD trips from origin n_o to destination n_d at period t_{i_p} is calculated as follows:

$$\boldsymbol{\mathcal{P}}^{i_p}_{n_o,n_d,t_{i_p}} = \mathbf{P}^{i_p,o}_{n_o,t_{i_p}} + \mathbf{P}^{i_p,d}_{n_d,t_{i_p}}. \tag{4}$$

The posterior distribution of $\mathbf{P}^{i_p,o}_{n_o} = [\mathbf{P}^{i_p,o}_{n_o,1},...,\mathbf{P}^{i_p,o}_{n_o,T_{i_p}}] \in \mathrm{R}^{T_{i_p}}$ is assumed to take the following form:

$$\mathbf{P}^{i_p,o}_{n_o} \sim \text{MVN}(\tilde{\boldsymbol{\mu}}^{i_p,o}_{n_o}, \tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o}), \tag{5}$$

where $\tilde{\boldsymbol{\mu}}^{i_p,o}_{n_o} \in \mathrm{R}^{T_{i_p}}$, $\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o} \in \mathrm{R}^{T_{i_p}}_{+}$, and associated covariance matrix is diagonal of size $T_{i_p} \times T_{i_p}$ with diagonal $\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o}$.

To estimate the parameters $\boldsymbol{\mu}^{i_p,o}_{n_o}$ and $\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o}$, the variational distribution given by Equation (5) is used to approximate the true posterior $P(\mathbf{P}^{i_p,o}_{n_o}|\boldsymbol{\mathcal{X}}^{\text{obs}})$, which is derived from the observed OD demands.

C. Formulation of the Residual Module

In addition to the periodic components, we develop a residual module that utilizes Tucker tensor decomposition (Fig. 2) to extract residual fluctuations $\boldsymbol{\mathcal{F}} \in \mathrm{R}^{N_o \times N_d \times T}$ from observed OD demands $\boldsymbol{\mathcal{X}}^{\text{obs}}$. Such deviations include latent spatial and temporal features, as well as their spatiotemporal interactions. Specifically, $\boldsymbol{\mathcal{F}}$ is decomposed as follows:

$$\boldsymbol{\mathcal{F}} = \boldsymbol{\mathcal{S}} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{Z}, \tag{6}$$

where $\mathbf{U} \in \mathrm{R}^{N_o \times h_o}$ and $\mathbf{V} \in \mathrm{R}^{N_d \times h_d}$ are factor matrices representing the latent spatial features of origins and destinations, respectively. These matrices reduce the dimensionality of the spatial domains while preserving essential structural dependencies among locations. $\mathbf{Z} \in \mathrm{R}^{T \times h_t}$ is the latent temporal factor matrix that captures temporal dynamics and correlations across different time intervals. $\boldsymbol{\mathcal{S}} \in \mathrm{R}^{h_o \times h_d \times h_t}$ is the core tensor, which captures the spatiotemporal interactions among the latent origin, destination, and temporal components. It serves as a compact representation of how spatial and temporal patterns jointly contribute to the observed variations in OD demand. Equivalently, at the element level, the (i,j,t) entry of $\boldsymbol{\mathcal{F}}$ is defined as follows:

$$\boldsymbol{\mathcal{F}}_{i,j,t} = \sum_{k_o=1}^{h_o} \sum_{k_d=1}^{h_d} \sum_{k_t=1}^{h_t} \boldsymbol{\mathcal{S}}_{k_o,k_d,k_t} \mathbf{U}_{i,k_o} \mathbf{V}_{j,k_d} \mathbf{Z}_{t,k_t}, \tag{7}$$

where the subscript denotes the index.

Since this module is designed for OD demand time series, we place less emphasis on the static spatial factor matrices U and V, but instead focus on modeling the dynamic temporal factor matrix $\mathbf{Z}$. We assume that $\mathbf{Z}$ consists of temporally correlated latent factors as rows, i.e., $\mathbf{Z} = [\mathbf{z}_1,...,\mathbf{z}_T]^T$ with $\mathbf{z}_t \in \mathrm{R}^{h_t}$. In a Bayesian setting, we need to specify the prior distributions of the latent temporal factors $\mathbf{z}_t$. The Markovian assumption is adopted in the proposed model, which means that the current temporal factor $\mathbf{z}_t$ is only dependent on the previous one $\mathbf{z}_{t-1}$. Therefore, an initial hidden temporal factor $\mathbf{z}_0$ is introduced, and the transition in the latent temporal space is characterized by a parameterized transition function $f : \mathrm{R}^{h_t} \to \mathrm{R}^{h_t} \times \mathrm{R}^{h_t}_{+}$. The temporal factor $\mathbf{z}_t$ at time step $t \in \{1,...,T\}$ is derived as follows:

$$\mathbf{z}_0 = \mathbf{0} \tag{8}$$

$$\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t} = f(\mathbf{z}_{t-1}; \theta_f) \tag{9}$$

$$\mathbf{z}_t \sim \text{MVN}(\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t}), \tag{10}$$

where $\text{MVN}(\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t})$ denotes the multivariate Gaussian distribution whose mean vector is $\boldsymbol{\mu}_{z,t}$ and covariance matrix is diagonal with $\boldsymbol{\sigma}^2_{z,t}$ ($\boldsymbol{\mu}_{z,t} \in \mathrm{R}^{h_t}$, $\boldsymbol{\sigma}^2_{z,t} \in \mathrm{R}^{h_t}_{+}$), and θ_f denotes the set of parameters of f. Before training, $\mathbf{z}_0$ is initialized to the zero vector of dimension h_t. Any function $f : \mathrm{R}^{h_t} \to \mathrm{R}^{h_t} \times \mathrm{R}^{h_t}_{+}$ with learnable parameters θ_f is a feasible choice in the training. In our framework, f is initialized to the weighted sum of two multi-layer perceptrons (MLPs) such that $\mathbf{z}_{t-1}, \mathbf{1}_{h_t} = f(\mathbf{z}_{t-1}; \theta_f), \forall \mathbf{z}_{t-1} \in \mathrm{R}^{h_t}$. This configuration of $\mathbf{z}_0$ and f corresponds to uninformative priors of latent temporal factors, which means that the temporal factors $\mathbf{z}_t$ are assumed to be independent and identically distributed before observing any OD demand data. This represents that the fluctuations in observed OD demands come from independent Gaussian noises.

The approximate inference of the posterior distributions of latent temporal factors is based on the residual fluctuations between OD demand observations and periodic patterns. We

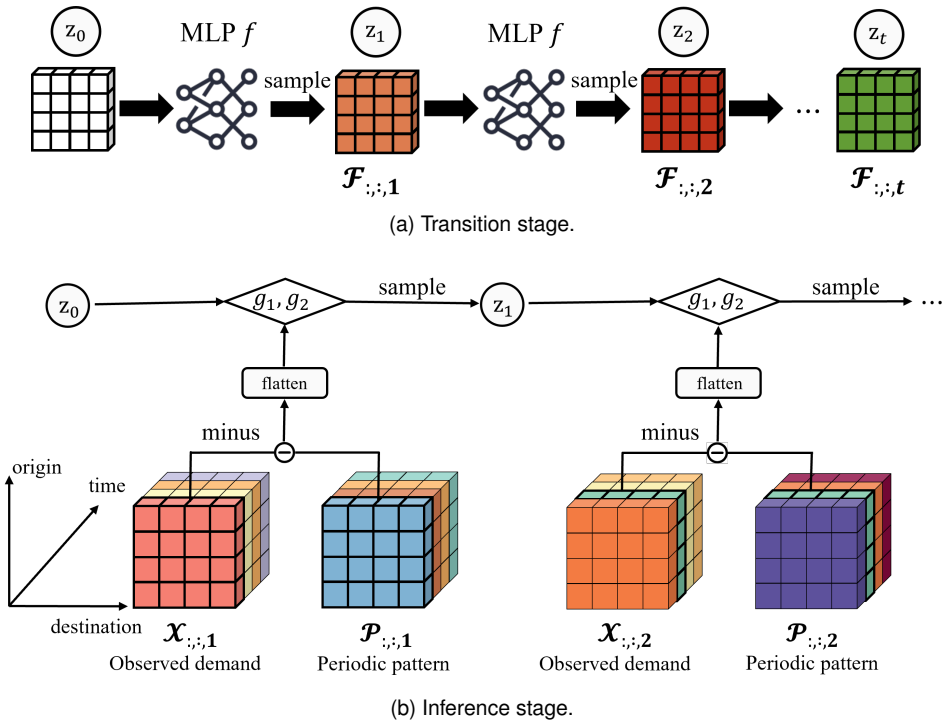


Fig. 3. Illustration of transition and inference in the residual module: in transition, $\mathbf{z}_t$ is sampled from the distribution resulting from $f(\mathbf{z}_{t-1})$ and f is initialized to output standard normal distribution corresponding to white noise in latent temporal space; in inference, the posterior distribution of $\mathbf{z}_t$ is approximated through $g_1(\mathbf{z}_{t-1}, g_2(\boldsymbol{\mathcal{X}}_{:,:,t} - \boldsymbol{\mathcal{P}}_{:,:,t}))$, where $\boldsymbol{\mathcal{X}}_{:,:,t} - \boldsymbol{\mathcal{P}}_{:,:,t}$ represents the residue of OD demand matrix at time t from the periodic patterns. In subfigure (b), yellow-colored pixels correspond to OD demand matrix $\boldsymbol{\mathcal{X}}_{:,:,t}$ and periodic patterns $\boldsymbol{\mathcal{P}}_{:,:,t}$ at time t respectively.

assume that there exists a hidden temporal factor $\tilde{\mathbf{z}}_0$ representing the initial temporal state, and the posterior distribution of the current temporal factor $\mathbf{z}_t$ can be inferred from the current observed OD demand matrix $\boldsymbol{\mathcal{X}}^{\text{obs}}_{:,:,t}$, periodic OD demands $\boldsymbol{\mathcal{P}}_{:,:,t}$, and previous temporal factor $\mathbf{z}_{t-1}$ altogether. Mathematically, there exist functions $g_1 : \mathrm{R}^{h_t} \times \mathrm{R}^{h_y} \to \mathrm{R}^{h_t} \times \mathrm{R}^{h_t}_{+}$ and $g_2 : \mathrm{R}^{N_o \times N_d} \to \mathrm{R}^{h_y}$, parametrized by ϕ_{g_1} and ϕ_{g_2}, such that:

$$\mathbf{z}_0 = \tilde{\mathbf{z}}_0; \tag{11}$$

$$\mathbf{y}_t = g_2(\boldsymbol{\mathcal{X}}_{:,:,t} - \boldsymbol{\mathcal{P}}_{:,:,t}; \phi_{g_2}); \tag{12}$$

$$\tilde{\boldsymbol{\mu}}_{z,t}, \tilde{\boldsymbol{\sigma}}^2_{z,t} = g_1(\mathbf{z}_{t-1}, \mathbf{y}_t; \phi_{g_1}); \tag{13}$$

$$\mathbf{z}_t \sim \text{MVN}(\tilde{\boldsymbol{\mu}}_{z,t}, \tilde{\boldsymbol{\sigma}}^2_{z,t}). \tag{14}$$

In the equations above, $\mathbf{y}_t$ is a latent representation of the difference between the observed OD matrix $\boldsymbol{\mathcal{X}}^{\text{obs}}_{:,:,t}$ and the matrix of periodic patterns $\boldsymbol{\mathcal{P}}_{:,:,t}$ at time t, and ϕ_{g_2} denotes the set of parameters of the encoding function g_2. The function g_1 combines the previous temporal factor $\mathbf{z}_{t-1}$ and the residual embedding $\mathbf{y}_t$ to produce the vectors $\tilde{\boldsymbol{\mu}}_{z,t}$ and $\tilde{\boldsymbol{\sigma}}^2_{z,t}$ of means and variances for the posterior distributions of $\mathbf{z}_t$. Fig. 3 provides a demonstration of the inference process for the transition between latent temporal factors $\mathbf{z}_t$.

D. Formulation of the Event Detection Module

The event detection module is designed to detect unexpected events from the observed OD demands $\boldsymbol{\mathcal{X}}^{\text{obs}}$. It employs a discrete univariate Lévy process as the prior distribution over the occurrences of unexpected events [63]. A Lévy process assumes independent and stationary incremental changes. It is essential to eliminate the impact of sudden shocks on the OD demands for accurately capturing the periodic patterns in the evolution of OD demands.

Let s_t denote the systematic increment at each time step t, representing the instantaneous deviation in OD demands introduced by unexpected events. We assume that s_t follows a normal prior. Let w_t be the cumulative sum of s_{k_s} from $k_s = 1$ to t, capturing the accumulated effect of these deviations over time. The relationship between s_t and w_t is given by:

$$s_{k_s} \overset{i.i.d.}{\sim} \text{Normal}(0, \sigma^2_{k_s}), k_s \in \{1, 2, ..., T\}, \tag{15}$$

$$w_t = \sum_{k_s=1}^{t} s_{k_s}, t \in \{1, 2, ..., T\}. \tag{16}$$

To learn periodic patterns $\boldsymbol{\mathcal{P}}$ apart from w_t, we assume the observations of OD demands follow normal distributions with learnable variances, expressed as follows:

$$\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t} \sim \text{Normal}(\boldsymbol{\mathcal{P}}_{i,j,t} + w_t | \sigma^2_{o,i} + \sigma^2_{d,j}), \tag{17}$$

where $\sigma^2_{o,i}$ and $\sigma^2_{d,j}$ indicate observational noises related to the origin i and destination j, respectively. Both vectors $\boldsymbol{\sigma}^2_o = [\sigma^2_{o,1}, ..., \sigma^2_{o,N_o}] \in \mathrm{R}^{N_o}_{+}$ and $\boldsymbol{\sigma}^2_d = [\sigma^2_{d,1}, ..., \sigma^2_{d,N_d}] \in \mathrm{R}^{N_d}_{+}$ are learnable parameters in SVI.

Assuming the historical OD demand tensor spans a duration of T time steps, the posterior distributions of the random events s_{k_s} occurring in the system are jointly modeled as independent Gaussian variables with learnable means and variances. This can be formulated as follows:

$$s_{k_s} \sim \text{Normal}(\tilde{\mu}_{s,k_s}, \tilde{\sigma}^2_{s,k_s}), \forall k_s \in \{1, ..., T\}, \tag{18}$$

where $\tilde{\mu}_{s,k_s}$ and $\tilde{\sigma}^2_{s,k_s}$ are posterior mean and variance of the increment at time k_s in $\{1, ..., T\}$, respectively.

E. Parameter Estimation

This section presents the parameter estimation procedures for the three modules described above. The estimation process is divided into two phases: the warm-up and formal phases. The warm-up phase utilizes the periodic and event detection modules to find the dominant periodic patterns by filtering out network- or system-level fluctuations. Next, leveraging the periodic components learned during the warm-up phase, the formal phase aims to: (i) uncover the underlying spatiotemporal correlations responsible for systematic OD demand deviations via the residual module; and (ii) refine the periodic representations within the module. The objective function in the SVI framework is a surrogate of $-\text{ELBO} = -\text{E}_{q_\phi}[\log p_\theta(\boldsymbol{X}, \boldsymbol{Z}^{\text{latent}}) - \log q_\phi(\boldsymbol{Z}^{\text{latent}})]$, which is formulated as follows:

$$\text{obj} = -\frac{1}{N}\sum_{i=1}^{N}[\log p_\theta(\boldsymbol{X}^{\text{obs}}, \boldsymbol{Z}_i^{\text{latent}}) - \log q_\phi(\boldsymbol{Z}_i^{\text{latent}})], \quad (19)$$

where $\boldsymbol{Z}_i^{\text{latent}}$ is a sample of latent variables from q_ϕ and $\boldsymbol{X}^{\text{obs}}$ denotes observed variables.

The reason we minimize the negative ELBO rather than the KL divergence directly is that the true posterior distribution is intractable. In contrast, the ELBO serves as a tractable surrogate objective and can be efficiently optimized. Mathematically, the relationship between the ELBO and KL divergence can be expressed as follows:

$$\log p_\theta(\boldsymbol{X}^{\text{obs}}) - \text{ELBO} = \text{KL}(q_\phi(\boldsymbol{Z}^{\text{latent}}) \| p_\theta(\boldsymbol{Z}^{\text{latent}} | \boldsymbol{X}^{\text{obs}})). \quad (20)$$

Since the KL divergence is a non-negative measure of similarity between two distributions, we have that $\log p_\theta(\boldsymbol{X}^{\text{obs}}) \geq \text{ELBO}$. Therefore, taking stochastic gradient steps in θ to minimize $-\text{ELBO}$ is supposed to increase the log evidence $\log p_\theta(\boldsymbol{X}^{\text{obs}})$. Importantly, for fixed θ, $\log p_\theta(\boldsymbol{X}^{obs})$ is also fixed, which means that the minimization of $-\text{ELBO}$ is equivalent to the minimization of $\text{KL}(q_\phi(\boldsymbol{Z}^{\text{latent}}) \| p_\theta(\boldsymbol{Z}^{\text{latent}} | \boldsymbol{X}^{\text{obs}}))$. As a result, the variational distributions q_ϕ provide a good approximation to the true posterior distributions $p_\theta(\boldsymbol{Z}^{\text{latent}} | \boldsymbol{X}^{\text{obs}})$ of latent variables $\boldsymbol{Z}^{\text{latent}}$ given observed variables $\boldsymbol{X}^{\text{obs}}$. In the following sections, we will assume that latent variables are drawn from q_ϕ only once so that cumbersome indexing summation over i to sample size N in Equation (19) could be avoided.

1) Warm-up Phase: In the warm-up phase, periodic and event detection modules are involved. The set $\boldsymbol{Z}$ of latent variables, and the sets ϕ and θ of learnable parameters are specified as follows:

$$\boldsymbol{Z}^{\text{latent}} = \{s_{k_s}, k_s = 1, ..., T, \mathbf{P}^{i_p,o}, \mathbf{P}^{i_p,d}, i_p = 1, 2\} \quad (21)$$

$$\phi = \{\tilde{\boldsymbol{\mu}}_{n_o}^{i_p,o}, \tilde{\boldsymbol{\sigma}}_{n_o}^{i_p,o^2}, n_o = 1, ..., N_o, \tilde{\boldsymbol{\mu}}_{n_d}^{i_p,d}, \tilde{\boldsymbol{\sigma}}_{n_d}^{i_p,d^2}, n_d = 1, ..., N_d, i_p = 1, 2, \tilde{\mu}_{s,k_s}, \tilde{\sigma}^2_{s,k_s}, k_s = 1, ..., T\} \quad (22)$$

$$\theta = \{\sigma^{i_p,o^2}, \sigma^{i_p,d^2}, i_p = 1, 2, \sigma^2_{o,n_o}, n_o = 1, ..., N_o, \sigma^2_{d,n_d}, n_d = 1, ..., N_d\} \quad (23)$$

The objective is formulated as follows:

$$\begin{aligned}\min_{\theta,\phi} \text{obj} &= -[\log(p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}} | \boldsymbol{Z}) p_\theta(\boldsymbol{Z})) - \log q_\phi(\boldsymbol{Z})] \\ &= \frac{1}{2}\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T}\frac{(\boldsymbol{\mathcal{X}}_{i,j,t}^{\text{obs}} - (\boldsymbol{\mathcal{P}}_{i,j,t} + \sum_{k_s=1}^{t} s_{k_s}))^2}{\sigma_{o,i}^2 + \sigma_{d,j}^2} \\ &\quad - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o}\frac{(\mathbf{P}_{n_o,t_{i_p}}^{i_p,o} - \tilde{\boldsymbol{\mu}}_{n_o,t_{i_p}}^{i_p,o})^2}{\tilde{\boldsymbol{\sigma}}_{n_o,t_{i_p}}^{i_p,o^2}} \\ &\quad - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d}\frac{(\mathbf{P}_{n_d,t_{i_p}}^{i_p,d} - \tilde{\boldsymbol{\mu}}_{n_d,t_{i_p}}^{i_p,d})^2}{\tilde{\boldsymbol{\sigma}}_{n_d,t_{i_p}}^{i_p,d^2}} \\ &\quad - \frac{1}{2}\sum_{k_s=1}^{T}\frac{(s_{k_s} - \mu_{s,k_s})^2}{\sigma^2_{s,k_s}} + \text{const.}, \end{aligned} \quad (24)$$

where we adopt a mean-field approximation to the true posteriors to simplify the expression of $\log q_\phi(\boldsymbol{Z})$.

2) Formal Phase: In the formal phase, periodic and residual modules are involved for the inference of posterior distributions of latent variables. The sets $\boldsymbol{Z}$, ϕ, and θ are defined as follows:

$$\boldsymbol{Z}^{\text{latent}} = \{\mathbf{z}_t, t = 1, ..., T, \mathbf{P}^{i_p,o}, \mathbf{P}^{i_p,d}, i_p = 1, 2\} \quad (25)$$

$$\phi = \{\tilde{\boldsymbol{\mu}}_{n_o}^{i_p,o}, \tilde{\boldsymbol{\sigma}}_{n_o}^{i_p,o^2}, n_o = 1, ..., N_o, \tilde{\boldsymbol{\mu}}_{n_d}^{i_p,d}, \tilde{\boldsymbol{\sigma}}_{n_d}^{i_p,d^2}, n_d = 1, ..., N_d, i_p = 1, 2, \tilde{\mathbf{z}}_0, \mathbf{U}, \mathbf{V}, \boldsymbol{\mathcal{S}}, \phi_{g_1}, \phi_{g_2}\} \quad (26)$$

$$\theta = \{\sigma^{i_p,o^2}, \sigma^{i_p,d^2}, i_p = 1, 2, \sigma^2_{o,n_o}, n_o = 1, ..., N_o, \sigma^2_{d,n_d}, n_d = 1, ..., N_d, \theta_f\} \quad (27)$$

The objective is formulated as follows:

$$\begin{aligned}\min_{\theta,\phi} \text{obj} &= -[\log(p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}} | \boldsymbol{Z}) p_\theta(\boldsymbol{Z})) - \log q_\phi(\boldsymbol{Z})] \\ &= \frac{1}{2}\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T}\frac{(\boldsymbol{\mathcal{X}}_{i,j,t} - (\boldsymbol{\mathcal{P}}_{i,j,t} + \boldsymbol{\mathcal{F}}_{i,j,t}))^2}{\sigma_{o,i}^2 + \sigma_{d,j}^2} \\ &\quad + \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{h_t}\log(\boldsymbol{\sigma}_{z,t,i_t}^2) \\ &\quad + \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{T}\frac{(\mathbf{z}_{t,i_t} - \boldsymbol{\mu}_{z,t,i_t})^2}{\boldsymbol{\sigma}_{z,t,i_t}^2} \\ &\quad - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o}\frac{(\mathbf{P}_{n_o,t_{i_p}}^{i_p,o} - \tilde{\boldsymbol{\mu}}_{n_o,t_{i_p}}^{i_p,o})^2}{\tilde{\boldsymbol{\sigma}}_{n_o,t_{i_p}}^{i_p,o^2}} \\ &\quad - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d}\frac{(\mathbf{P}_{n_d,t_{i_p}}^{i_p,d} - \tilde{\boldsymbol{\mu}}_{n_d,t_{i_p}}^{i_p,d})^2}{\tilde{\boldsymbol{\sigma}}_{n_d,t_{i_p}}^{i_p,d^2}} \\ &\quad - \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{T}\log\tilde{\boldsymbol{\sigma}}_{z,t,i_t}^2 \\ &\quad - \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{T}\frac{(\mathbf{z}_{t,i_t} - \tilde{\boldsymbol{\mu}}_{z,t,i_t})^2}{\tilde{\boldsymbol{\sigma}}_{z,t,i_t}^2} \\ &\quad + \text{const.} \end{aligned} \quad (28)$$

In the equations above, $\mathbf{z}_{t,i_t}$ denotes the i_t-th element of the vector $\mathbf{z}_t \in \mathrm{R}^{h_t}$; $\boldsymbol{\mu}_{z,t,i_t}$ denotes the i_t-th element of the vector $\boldsymbol{\mu}_{z,t} \in \mathrm{R}^{h_t}$ and similar for $\boldsymbol{\sigma}^2_{z,t}$, $\tilde{\boldsymbol{\mu}}^2_{z,t}$, and $\tilde{\boldsymbol{\sigma}}^2_{z,t}$. Please refer to Appendix A and B for the details in deriving Equations (24) and (28).

F. Forecasting

The long-term forecasting of OD demands relies on the periodic and residual modules by sampling from posterior distributions. The first step is to sample periodic components corresponding to origins and destinations. For the observed duration $t = 1, ..., T$, given initial temporal hidden status $\mathbf{z}_0 = \tilde{\mathbf{z}}_0$, we are able to sample $\mathbf{z}_t$ iteratively using Equations (11)–(14) with the learned functions g_1 and g_2. At $t = T$, the transition function f is used to characterize the evolution and give the approximate posteriors of $\mathbf{z}_{T+1}$ (Equations (8)–(10)) into the future. Then the predicted OD demand $\boldsymbol{\mathcal{X}}_{i,j,T+1}$ is sampled from the posterior

$$\boldsymbol{\mathcal{X}}_{i,j,T+1} \sim \text{Normal}(\boldsymbol{\mathcal{P}}_{i,j,t+1} + \boldsymbol{\mathcal{F}}_{i,j,t+1} | \sigma^2_{o,i} + \sigma^2_{d,j}), \quad (29)$$

where $\boldsymbol{\mathcal{F}}_{i,j,T+1}$ is obtained via Equation (6). This transition could be implemented repeatedly τ steps over the forecasting horizon until the predictions $\boldsymbol{\mathcal{X}}_{:,:,T+1:T+\tau}$ of future OD demands are obtained. The above process could be repeated multiple times to obtain numerous posterior samples of $\mathbf{X}_{:,:,T+1:T+\tau}$, which draws a picture of the predictive distributions of future OD demands in such a way that decision makers are given a measure of uncertainty quantification.

V. Numerical Experiments and Evaluation

In this section, the proposed model is examined with real-world OD datasets to evaluate long-term predictive performance against popular deep learning models for time series processing. The computer programs are written in Python with the PyTorch deep learning framework and an NVIDIA GeForce RTX 3070 GPU is used to run the experiments. The details of datasets, evaluation metrics, and experimental results are presented in the following sections.

A. Datasets

To evaluate the predictive performance of the proposed P-R SVI model alongside baseline methods, we conduct experiments on four real-world OD demand datasets. These datasets cover different modes of transportation, including taxi, metro, and micro-mobility systems such as bicycles and e-scooters, collected from different countries. In the following, we provide a detailed description of these datasets.

- NYC TLC Trip Record Data [64]: This dataset consists of taxi trip information including pick-up/drop-off times and locations. The pick-up and drop-off locations are aggregated into origins and destinations defined by 63 taxi zones over Manhattan, which results in a total of 63 $\times$ 63 origin-destination pairs. The hourly OD dataset is generated by counting the total number of trips on an hourly basis (according to pick-up time) for each O-D pair. The training data range from May 10, 2011 to June 9, 2011; the data between June 10, 2011 and June 16, 2011 are left for the testing set.
- Bay Area Rapid Transit (BART) Data [65]: This dataset records the number of passengers between stations. This Bay Area network system comprises 50 $\times$ 50 station-level origin-destination pairs. The training data range from February 11, 2019 to March 12, 2019; the data between March 13, 2019 and March 19, 2019 are left for the testing set.
- Helsinki Electric Scooter Data: This dataset contains the shared e-scooter trip information in the southern major district of Helsinki. The 5 origins and destinations are the sub-districts which together cover the city center. The training data range from June 17, 2021 to July 16, 2021; the data between July 17, 2021 and July 23, 2021 are left for the testing set.
- Citi Bike Trip Data: This dataset is provided by the NYC Citi Bike share system, consisting of detailed ride information. The 37 origins/destinations are obtained by dividing the Manhattan island into neighborhoods [66]. The training data range from May 1, 2023 to May 30, 2023; the data between May 31, 2023 and June 6, 2023 are left for the testing set.

Before conducting numerical experiments, we first examine whether there exists a significant distributional difference in hourly OD (origin–destination) demands between weekdays and weekends in the Bay Area Rapid Transit (BART) system. We analyze a 200-day period ending on March 12, 2019, which includes 142 weekdays and 58 weekends. For each OD pair and hour, we apply Welch's two-sample t-test to compare the 142 weekday observations with the 58 weekend observations and compute the corresponding p-values. To ensure meaningful test statistics, we exclude OD–hour combinations with an average demand below 5. Among the remaining combinations, 79.2% yield $p < 0.05$, and 73.1% yield $p < 0.01$. Fig 4 illustrates the distribution differences for hourly OD demands from MONT to DALY in BART. The t-statistics and p-values reported above the hours indicate that the differences are statistically significant across nearly all periods, especially during the peak hours (16:00–19:00), where weekday demands substantially exceed weekend levels. This pattern confirms strong weekday–weekend heterogeneity in travel behavior for this OD pair.

B. Evaluation Metrics

We use the following metrics to evaluate the performance of our model.

- Rooted Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \quad (30)$$

where $\hat{y}_i$ represents the predicted value and y_i is the ground truth.

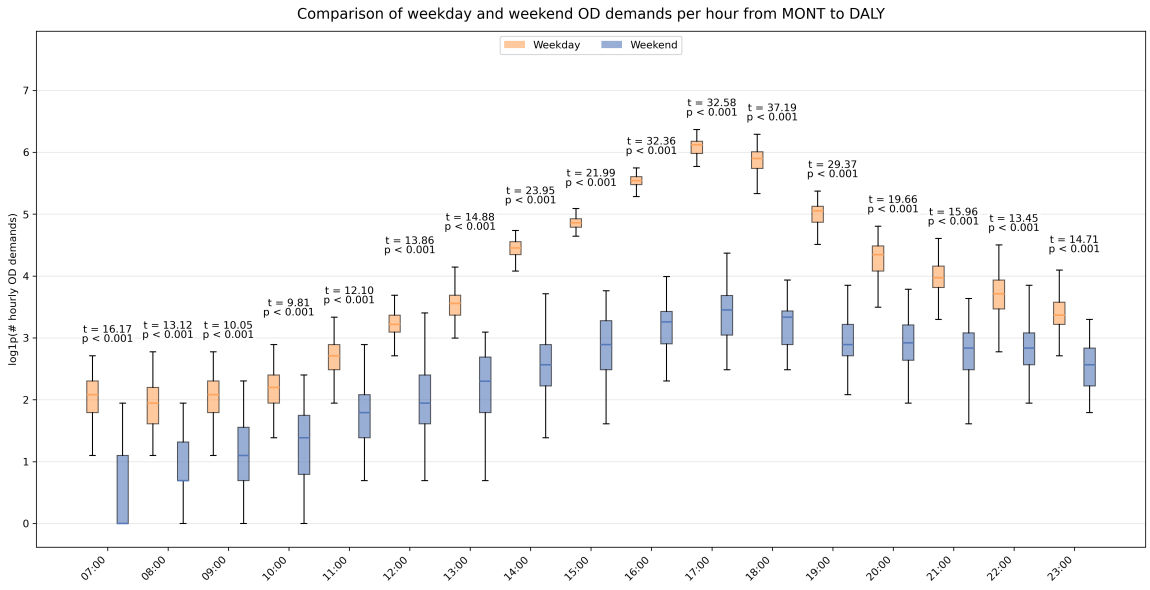


Fig. 4. Boxplots of hourly OD demands during weekdays and weekends for OD pair MONT-DALY in BART dataset.

- Weight Mean Absolute Percentage Error (WMAPE):

$$\text{WMAPE} = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{\sum_{i=1}^{N} y_i}. \tag{31}$$

WMAPE is related to the scale of the data and the use of it can reduce the disproportionate impact of large-value outliers compared to MSE. It also avoids the division by zero to reduce the calculation instability.

- Continuous Ranked Probability Score (CRPS):
CRPS measures the discrepancy between the predicted cumulative distribution function and the empirical distribution of observations. The general formula for CRPS is

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}(x \geq y)\right)^2 dx, \tag{32}$$

where F is the cumulative distribution function of the predictive distribution and y is the observation. We will use the discrete approximation of CRPS for intractable distributions. Readers could refer to [67] for a comprehensive introduction.

- Mean Interval Length (MIL):
Mean interval length is the arithmetic average length of the prediction intervals for all OD pairs and future times. In the experiments, we consider the 80% prediction interval, whose range is specified by the 90% and 10% quantiles of the prediction samples as end points. In symbols,

$$\text{MIL } 80\% = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_{i,0.9} - \hat{y}_{i,0.1}|, \tag{33}$$

where $\hat{y}_{i,q}$ denotes the q-th quantile of empirical predicted values for observation i.

- Interval Coverage Percentage (ICP):
ICP computes the fraction of the observations falling within the prediction intervals. Therefore, for 80% prediction intervals, the ICP should be approximately 80%. For example,

$$\text{ICP } 80\% = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(y_i \leq \hat{y}_{i,0.9}) \mathbf{1}(y_i \geq \hat{y}_{i,0.1}). \tag{34}$$

The RMSE metric offers a straightforward measure of the magnitude of the overall predictive error; WMAPE provides a way to account for the varying importance or scale of each item being forecast, ensuring that the forecast accuracy measurement is more reflective of the impact of the errors. The other evaluation metrics are used to assess the ability to characterize the predictive distributions.

C. Experimental Results

Table I provides a comparison of predictive performance of different models on all OD datasets, respectively, measured by various evaluation metrics. It can be seen that the proposed P-R SVI model exhibits superior overall performance compared to other forecasting baselines, which can be divided into two perspectives.

1) Predictive Accuracy: The predictive accuracy is evaluated according to RMSE and WMAPE, and our P-R SVI model shows superior performance to other models, especially in long-term forecasting. Transformer-based models are designed to capture long-term contextual information, thus theoretically well suited for long-term forecasting tasks. However, it should be noted that transformer-based models, Autoformer and Informer, display a substantial divergence from the ground truth in terms of RMSE and WMAPE. In particular, in the training of Autoformer on NYC TLC dataset and Informer on Bart dataset, the predictive results are subject to abnormally significant variance, indicated by the larger standard deviation across repeated experiments. Therefore, the predictions generated by transformer-based models are not

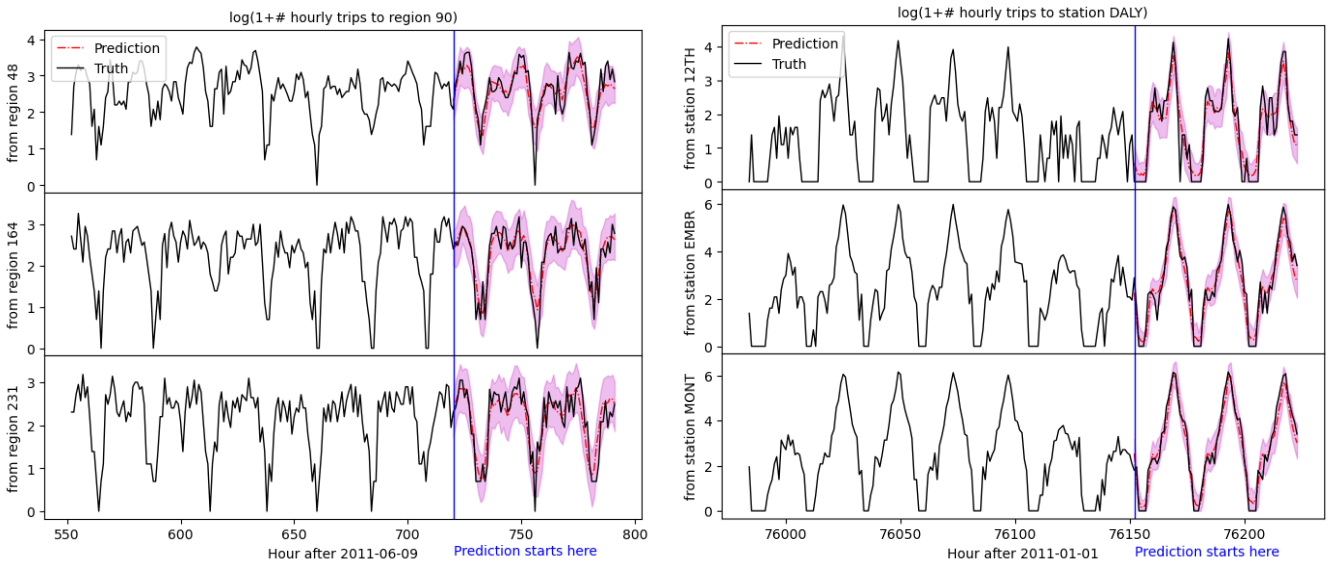
TABLE I
PERFORMANCE COMPARISON OF P-R SVI WITH OTHER LONG-TERM FORECASTING MODELS ON REAL-WORLD DATASETS (EMPIRICAL AVERAGE ± STANDARD DEVIATION OVER 5 EXPERIMENTAL REPETITIONS).

		NYC TLC Trip Record					**Bay Area Rapid Transit**				
Model	**Day(s)**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
Prophet [42]	1	3.79 ± 0.00	0.336 ± 0.000	–	6.41 ± 0.00	0.852 ± 0.000	14.15 ± 0.00	0.423 ± 0.000	–	9.72 ± 0.00	0.854 ± 0.000
	2	3.72 ± 0.00	0.341 ± 0.000	–	6.57 ± 0.00	0.852 ± 0.000	13.70 ± 0.00	0.415 ± 0.000	–	9.89 ± 0.00	0.856 ± 0.000
	3	3.89 ± 0.00	0.367 ± 0.000	–	6.60 ± 0.00	0.832 ± 0.000	12.88 ± 0.00	0.409 ± 0.000	–	9.72 ± 0.00	0.853 ± 0.000
	7	3.96 ± 0.00	0.389 ± 0.000	–	6.21 ± 0.00	0.827 ± 0.000	12.43 ± 0.00	0.465 ± 0.000	–	8.13 ± 0.00	0.831 ± 0.000
DeepAR [40]	1	4.19 ± 0.68	0.358 ± 0.038	**0.204 ± 0.001**	5.27 ± 0.78	0.788 ± 0.032	9.75 ± 0.91	0.336 ± 0.020	**0.189 ± 0.004**	9.74 ± 2.06	0.820 ± 0.026
	2	4.69 ± 0.58	0.406 ± 0.039	0.216 ± 0.007	5.51 ± 0.96	0.770 ± 0.036	12.67 ± 2.15	0.377 ± 0.032	**0.199 ± 0.002**	10.49 ± 2.28	0.814 ± 0.225
	3	4.75 ± 0.54	0.423 ± 0.038	0.224 ± 0.007	5.62 ± 1.04	0.760 ± 0.038	11.86 ± 1.62	0.379 ± 0.022	**0.211 ± 0.007**	9.87 ± 2.07	0.816 ± 0.012
	7	4.55 ± 0.64	0.435 ± 0.048	0.226 ± 0.009	5.43 ± 0.97	0.761 ± 0.040	11.40 ± 1.37	**0.413 ± 0.021**	**0.215 ± 0.006**	8.46 ± 1.62	**0.813 ± 0.013**
DeepJMQR [51]	1	7.57 ± 0.14	0.658 ± 0.007	–	9.23 ± 0.69	0.755 ± 0.005	28.28 ± 0.03	0.973 ± 0.001	–	2.02 ± 0.07	0.534 ± 0.005
	2	7.16 ± 0.12	0.646 ± 0.007	–	9.14 ± 0.69	0.759 ± 0.004	28.08 ± 0.03	0.973 ± 0.001	–	2.00 ± 0.07	0.533 ± 0.005
	3	6.98 ± 0.08	0.650 ± 0.005	–	9.12 ± 0.69	0.754 ± 0.004	26.78 ± 0.36	0.972 ± 0.002	–	2.00 ± 0.07	0.531 ± 0.005
	7	6.78 ± 0.06	0.669 ± 0.003	–	9.08 ± 0.70	0.757 ± 0.003	23.83 ± 0.03	0.972 ± 0.002	–	1.99 ± 0.07	0.567 ± 0.005
Autoformer [48]	1	17.13 ± 1.78	1.487 ± 0.092	–	–	–	29.79 ± 0.57	1.259 ± 0.057	–	–	–
	2	17.44 ± 1.74	1.509 ± 0.092	–	–	–	29.59 ± 0.46	1.248 ± 0.046	–	–	–
	3	17.58 ± 1.84	1.523 ± 0.092	–	–	–	28.45 ± 0.45	1.246 ± 0.045	–	–	–
	7	16.70 ± 1.85	1.554 ± 0.098	–	–	–	25.56 ± 0.50	1.307 ± 0.046	–	–	–
Informer [68]	1	13.07 ± 0.36	1.351 ± 0.046	0.746 ± 0.014	6.59 ± 0.95	0.431 ± 0.009	39.65 ± 6.86	1.970 ± 0.396	1.146 ± 0.108	8.39 ± 4.05	0.263 ± 0.100
	2	12.38 ± 0.30	1.254 ± 0.042	0.697 ± 0.012	6.38 ± 0.51	0.459 ± 0.009	37.60 ± 4.62	1.827 ± 0.294	1.096 ± 0.092	8.53 ± 3.93	0.295 ± 0.115
	3	11.81 ± 0.19	1.202 ± 0.031	0.684 ± 0.006	6.11 ± 0.57	0.468 ± 0.011	35.73 ± 3.71	1.775 ± 0.238	1.076 ± 0.074	8.76 ± 3.93	0.303 ± 0.114
	7	–	–	–	–	–	32.57 ± 3.59	1.858 ± 0.332	1.032 ± 0.039	8.65 ± 3.61	0.307 ± 0.098
P-R SVI	1	**3.73 ± 0.07**	**0.331 ± 0.002**	**0.204 ± 0.001**	5.67 ± 0.02	**0.810 ± 0.001**	**7.63 ± 0.13**	**0.309 ± 0.003**	0.203 ± 0.000	7.95 ± 0.05	**0.800 ± 0.000**
	2	**3.61 ± 0.05**	**0.337 ± 0.002**	**0.207 ± 0.001**	5.79 ± 0.02	**0.812 ± 0.000**	**8.62 ± 0.05**	**0.328 ± 0.002**	0.209 ± 0.000	8.02 ± 0.04	**0.796 ± 0.001**
	3	**3.61 ± 0.04**	**0.350 ± 0.001**	**0.214 ± 0.000**	5.96 ± 0.02	**0.806 ± 0.001**	**9.12 ± 0.05**	**0.346 ± 0.003**	0.215 ± 0.001	7.87 ± 0.05	**0.790 ± 0.001**
	7	**3.86 ± 0.02**	**0.380 ± 0.001**	**0.218 ± 0.000**	5.87 ± 0.03	**0.806 ± 0.001**	**10.98 ± 0.07**	0.429 ± 0.004	0.226 ± 0.001	6.16 ± 0.03	0.778 ± 0.001

		Helsinki electric scooters					**Citi Bike Trip**				
Model	**Day(s)**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
Prophet	1	28.30 ± 0.00	0.314 ± 0.000	–	63.16 ± 0.00	**0.802 ± 0.000**	3.40 ± 0.00	0.410 ± 0.000	–	3.30 ± 0.00	0.878 ± 0.000
	2	24.18 ± 0.00	0.319 ± 0.000	–	50.19 ± 0.00	**0.773 ± 0.000**	4.24 ± 0.00	0.498 ± 0.000	–	3.30 ± 0.00	0.877 ± 0.000
	3	22.63 ± 0.00	0.335 ± 0.000	–	45.86 ± 0.00	**0.783 ± 0.000**	4.01 ± 0.00	0.497 ± 0.000	–	3.36 ± 0.00	0.878 ± 0.000
	7	37.23 ± 0.00	0.548 ± 0.000	–	49.66 ± 0.00	0.707 ± 0.000	3.98 ± 0.00	0.449 ± 0.000	–	3.28 ± 0.00	0.868 ± 0.000
DeepAR	1	28.36 ± 3.42	0.293 ± 0.021	0.305 ± 0.007	21.96 ± 3.48	0.593 ± 0.018	3.30 ± 0.59	0.407 ± 0.004	**0.137 ± 0.002**	2.13 ± 0.12	0.867 ± 0.031
	2	23.94 ± 3.21	0.294 ± 0.038	0.293 ± 0.011	19.96 ± 3.79	0.643 ± 0.020	8.96 ± 2.33	0.843 ± 0.017	0.167 ± 0.003	3.33 ± 0.20	**0.835 ± 0.030**
	3	20.20 ± 4.18	0.281 ± 0.040	**0.285 ± 0.010**	18.61 ± 4.10	0.680 ± 0.020	12.11 ± 2.87	1.231 ± 0.028	0.194 ± 0.003	3.99 ± 0.28	**0.816 ± 0.033**
	7	24.78 ± 4.87	0.343 ± 0.047	0.322 ± 0.014	17.95 ± 4.77	0.638 ± 0.031	14.04 ± 3.06	1.712 ± 0.035	0.234 ± 0.005	4.68 ± 0.43	**0.787 ± 0.047**
DeepJMQR	1	28.74 ± 3.27	0.303 ± 0.005	–	18.97 ± 1.02	0.581 ± 0.007	6.97 ± 0.02	0.497 ± 0.004	–	3.87 ± 0.09	0.601 ± 0.007
	2	30.22 ± 3.46	0.312 ± 0.007	–	20.28 ± 0.98	0.577 ± 0.006	7.24 ± 0.02	0.674 ± 0.005	–	3.99 ± 1.00	0.632 ± 0.006
	3	29.90 ± 4.21	0.307 ± 0.011	–	21.33 ± 1.07	0.591 ± 0.007	8.30 ± 0.03	0.831 ± 0.007	–	4.07 ± 1.05	0.623 ± 0.006
	7	31.90 ± 4.43	**0.321 ± 0.009**	–	22.37 ± 1.01	0.601 ± 0.008	8.84 ± 0.10	0.875 ± 0.007	–	4.86 ± 1.10	0.651 ± 0.008
Autoformer	1	75.59 ± 3.92	0.974 ± 0.789	–	–	–	9.88 ± 0.72	2.392 ± 0.047	–	–	–
	2	62.63 ± 3.72	0.967 ± 0.772	–	–	–	9.32 ± 0.89	2.019 ± 0.051	–	–	–
	3	54.43 ± 4.00	0.968 ± 0.812	–	–	–	9.48 ± 0.91	2.221 ± 0.048	–	–	–
	7	50.91 ± 3.88	1.034 ± 0.801	–	–	–	9.32 ± 0.97	2.216 ± 0.057	–	–	–
Informer	1	74.56 ± 3.02	0.948 ± 0.243	1.385 ± 0.097	30.73 ± 1.11	0.293 ± 0.122	18.54 ± 3.47	4.521 ± 0.218	1.578 ± 0.131	24.95 ± 3.19	0.539 ± 0.108
	2	61.64 ± 3.18	0.962 ± 0.349	1.311 ± 0.105	42.62 ± 0.97	0.354 ± 0.101	16.73 ± 2.98	4.443 ± 0.309	1.476 ± 0.119	22.15 ± 2.89	0.536 ± 0.114
	3	54.06 ± 4.07	0.994 ± 0.240	1.281 ± 0.112	48.94 ± 1.00	0.385 ± 0.132	17.03 ± 3.99	4.447 ± 0.377	1.497 ± 0.121	21.90 ± 3.06	0.538 ± 0.117
	7	50.25 ± 3.08	1.070 ± 0.357	1.273 ± 0.103	60.75 ± 1.01	0.416 ± 0.121	16.89 ± 3.41	4.320 ± 0.331	1.379 ± 0.137	20.40 ± 3.01	0.551 ± 0.116
P-R SVI	1	**23.17 ± 0.49**	**0.263 ± 0.007**	**0.295 ± 0.001**	49.60 ± 1.02	0.833 ± 0.001	**3.02 ± 0.03**	**0.403 ± 0.003**	**0.137 ± 0.000**	3.24 ± 0.02	**0.844 ± 0.002**
	2	**20.40 ± 0.44**	**0.276 ± 0.006**	**0.286 ± 0.001**	41.65 ± 1.18	0.848 ± 0.002	**3.32 ± 0.02**	**0.433 ± 0.003**	**0.162 ± 0.001**	3.34 ± 0.03	0.846 ± 0.001
	3	**17.76 ± 0.38**	**0.279 ± 0.007**	0.290 ± 0.000	36.81 ± 1.09	0.844 ± 0.001	**3.12 ± 0.02**	**0.441 ± 0.002**	**0.165 ± 0.001**	3.39 ± 0.03	0.850 ± 0.001
	7	**24.68 ± 0.51**	0.336 ± 0.008	**0.313 ± 0.001**	34.91 ± 1.13	**0.810 ± 0.001**	**3.32 ± 0.02**	**0.447 ± 0.004**	**0.174 ± 0.001**	3.26 ± 0.04	0.840 ± 0.002

reliable for decision making. DeepJMQR yields better performance than transformers, but is still trailing by a large margin from the proposed P-R SVI model. Prophet and DeepAR achieve comparable predictive accuracy probably because of their Bayesian nature, but are not as accurate as P-R SVI.

2) Predictive Distribution Characterization: Unlike existing studies that focus primarily on point-valued estimates, our proposed model can also provide a measure of uncertainty quantification. The point estimates of the OD demand forecasts do not provide a complete foundation for future decision making. Therefore, it is essential to have a characterization of the probability distributions of future OD demands so that uncertainty quantification could be taken into consideration in the development of operational strategies. Although Prophet estimates posterior updates on each individual time series based on historical data, it overlooks the interactions among them. This drawback results in the over-estimated mean interval length (MIL), which is expected to measure the average difference between the 10% and 90% quantiles. Consequently, the wider predicted 80% prediction intervals support



(a) Number of trips to region 90

(b) Number of trips to station DALY

Fig. 5. Visualization of travel demand forecasts with predicted 80% prediction intervals: the red dashed line consists of point OD demand predictions, equal to sample means from approximate posteriors; the shaded area depicts the uncertainty quantification, bounded by 10% and 90% predictive quantiles; the dark solid line represents the ground truth; the blue vertical line marks the start of the forecasting horizon. Please note that each displayed OD demand y is transformed using $\log(1+y)$.

significantly more than 80% of actual data points. DeepAR selects a distribution family initially and then estimates the distributional parameters. The resulting CRPS from DeepAR samples is comparable to that of the proposed P-R SVI; however, P-R SVI provides a more accurate representation of the predictive distributions, capturing around 80% of observations within more accurate prediction intervals. Unlike DeepAR, DeepJMQR directly computes the corresponding prediction intervals, instead of sampling from the predictive distributions to estimate the predictive 10% and 90% quantiles. The mean 80% interval length and interval coverage percentage provided by DeepJMQR on NYC TLC dataset are nearly satisfying; but the performance on BART dataset degrades drastically. In general, P-R SVI is empirically shown to excel in depicting the predictive distributions. This result is illustrated in Fig 5, which shows OD demand predictions, 80% prediction intervals, and actual data for several OD pairs in both datasets. It is easy to observe that the OD demand forecasts align with the evolution in the ground truth, and the ground truth is consistently enveloped by the prediction intervals.

3) Explainability Exploration: This section is devoted to the discussion of the learned periodic components in the periodic module of P-R SVI, which provide explanation for daily and hourly variations in OD demands. Fig. 6 could help explain periodic patterns. The daily periodic patterns are easily discerned in the first column by observing that weekdays experience a larger volume of travel demands, which diminish in weekends. This observation is in accordance with the necessary commute trips in a five-day workweek. Different hourly variation patterns of OD pairs are illustrated in the second column of Fig. 6. Both hourly periodic patterns fluctuate around zero and reach the respective minima between midnight and dawn. However, zoom-in windows on Monday could shed a light on the different functionality of MONT and DALY from the perspective of a city. The maximum influx of trips from MONT to DALY occurs between 5:00PM and 6:00PM, which suggests that DALY is probably a common destination of commute trips, i.e. home, or a transfer station to other transit systems. In contrast, the peak hour from DALY to MONT is between 8:00AM and 9:00AM, the exact opposite situation of MONT-DALY trips, which further affirms the previous inference. By examining the hourly variations in OD demands between two stations, it is to infer that DALY is likely to be a residential area while MONT might be a commercial center in the BART system. The explicit integration of the periodic module into a deep learning framework successfully extract explainable patterns observed in real OD trips.

4) Ablation Study: Both the residual and event detection modules are aimed to explore temporal fluctuations, and it is thus necessary to examine the functionality of the modules over the long term. Therefore, in terms of the evaluation metrics mentioned above, we also assess the performance of the periodic and event detection SVI (P-E SVI) and present the results in Table II. P-E SVI is a simplified variant of P-R SVI, where the residual module is removed. Since the residual module is composed of latent random variables, its removal noticeably reduces the variance across experimental repetitions. This is supported by the negligible standard deviations in the numerical experiments of P-E SVI. Another notable feature is that P-E SVI exhibits consistently improved predictive accuracy as the duration prolongs, illustrated by the decreasing RMSE. Conversely, even though the performance of P-R SVI is better than other models, RMSE on P-R SVI declines over the extended duration. This degradation

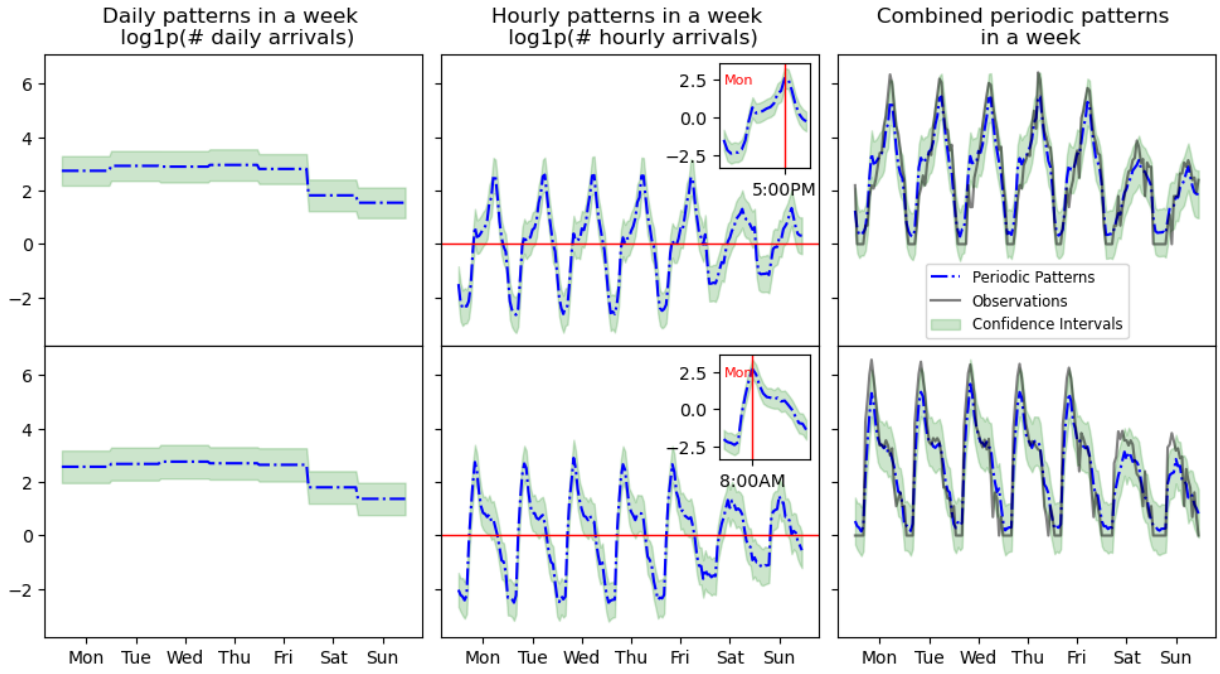


Fig. 6. The periodic patterns of OD demands between DALY and MONT stations in BART. The first row corresponds to origin MONT and destination DALY while the second row corresponds to origin DALY and destination MONT. The estimated posterior means of different periodic components are represented by blue dashed lines, enveloped by light-green shaded 80% prediction intervals. Hourly periodic patterns of OD demands are displayed in the second column, which can be used to discriminate origins and destinations of different functionality. In the zoom-in (Monday) windows, the peak hours of OD data are annotated by red vertical lines. Again, each displayed OD demand y is transformed using $\log(1+y)$.

TABLE II
PERFORMANCE COMPARISON OF P-E SVI AND P-R SVI ON REAL-WORLD DATASETS (EMPIRICAL AVERAGE ± STANDARD DEVIATION OVER 5 EXPERIMENTAL REPETITIONS).

		NYC TLC Trip Record					Bay Area Rapid Transit				
Model	**Day(s)**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
P-E SVI	1	5.20 ± 0.01	0.419 ± 0.000	0.233 ± 0.000	6.16 ± 0.00	**0.804 ± 0.001**	12.11 ± 0.05	0.414 ± 0.001	0.239 ± 0.000	10.85 ± 0.03	0.804 ± 0.001
	2	4.80 ± 0.00	0.410 ± 0.000	0.233 ± 0.000	6.26 ± 0.00	**0.806 ± 0.000**	12.03 ± 0.02	0.412 ± 0.000	0.239 ± 0.000	10.94 ± 0.02	**0.804 ± 0.000**
	3	4.50 ± 0.00	0.404 ± 0.000	0.234 ± 0.000	6.29 ± 0.00	**0.805 ± 0.000**	11.59 ± 0.02	0.414 ± 0.000	0.241 ± 0.000	10.71 ± 0.02	**0.802 ± 0.000**
	7	4.31 ± 0.00	0.410 ± 0.000	0.233 ± 0.000	6.01 ± 0.00	**0.806 ± 0.000**	11.03 ± 0.01	0.447 ± 0.000	0.236 ± 0.000	8.49 ± 0.01	**0.806 ± 0.001**
P-R SVI	1	**3.73 ± 0.07**	**0.331 ± 0.002**	**0.204 ± 0.001**	5.67 ± 0.02	0.810 ± 0.001	**7.63 ± 0.13**	**0.309 ± 0.003**	**0.203 ± 0.000**	7.95 ± 0.05	**0.800 ± 0.000**
	2	**3.61 ± 0.05**	**0.337 ± 0.002**	**0.207 ± 0.001**	5.79 ± 0.02	0.812 ± 0.000	**8.62 ± 0.05**	**0.328 ± 0.002**	**0.209 ± 0.000**	8.02 ± 0.04	**0.796 ± 0.001**
	3	**3.61 ± 0.04**	**0.350 ± 0.001**	**0.214 ± 0.000**	5.96 ± 0.02	0.806 ± 0.001	**9.12 ± 0.05**	**0.346 ± 0.003**	**0.215 ± 0.001**	7.87 ± 0.05	0.790 ± 0.001
	7	**3.86 ± 0.02**	**0.380 ± 0.001**	**0.218 ± 0.000**	5.87 ± 0.03	**0.806 ± 0.001**	**10.98 ± 0.07**	**0.429 ± 0.004**	**0.226 ± 0.001**	6.16 ± 0.03	0.778 ± 0.001

is inevitable since a recurrent neural network (RNN) is a component of the residual module in P-R SVI. The RNN helps capture transient fluctuations accurately in the near future, but would result in error accumulation in the long run. As a consequence, P-E SVI is better at depicting the predictive distributions, supported by the more accurate interval coverage percentage (ICP).

We also investigate the sensitivity of P-R SVI to different hyperparameter settings, and the results are present in Table III. Using the NYC TLC trip records, we evaluate performance under different settings of the residual module of P-R SVI, with $h_o = h_d \in 24, 36$ and $h_t \in 48, 72, 96$. Across these configurations, the evaluation metrics remain closely similar. While larger hyperparameter values slightly increase variations across experimental runs, they do not diminish P-R SVI's predictive accuracy or its uncertainty quantification performance.

D. *Practical challenges in real-time applications*

Real-time deployment of the proposed model could present several practical challenges. Another practical limitation encountered in real-world applications is data availability. In long-term forecasting, even when accumulated errors are mitigated, predictive performance inevitably degrades over time. This requires frequent update of the model parameters based on newly observed OD demands. For certain transportation modes, such as e-hailing, trip origins and destinations are obtained immediately. In contrast, within metro systems, the origin station can be captured when a passenger taps their smart card upon entry, but the destination remains unknown until the card is tapped again upon exit. Consequently, there exists an inherent delay in OD data acquisition, which constrains the real-time adaptability of the model.

The computational efficiency and scalability of the proposed framework are critical for practical deployment. In our model,

TABLE III
PERFORMANCE COMPARISON OF P-R SVI WITH DIFFERENT HYPERPARAMETER SETTINGS ON NYC TLC TRIP RECORD DATASET (EMPIRICAL AVERAGE ± STANDARD DEVIATION OVER 5 EXPERIMENTAL REPETITIONS).

	$h_o = h_d = 24, h_t = 48$					$h_o = h_d = 36, h_t = 48$				
Day(s)	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
1	3.81 ± 0.02	0.337 ± 0.000	0.211 ± 0.001	5.87 ± 0.00	0.813 ± 0.001	3.89 ± 0.06	0.340 ± 0.001	0.217 ± 0.001	5.50 ± 0.03	0.790 ± 0.001
2	3.79 ± 0.01	0.347 ± 0.002	0.213 ± 0.001	5.91 ± 0.01	0.814 ± 0.002	3.91 ± 0.04	0.343 ± 0.001	0.214 ± 0.001	5.51 ± 0.02	0.782 ± 0.001
3	3.87 ± 0.02	0.361 ± 0.001	0.220 ± 0.001	5.98 ± 0.01	0.810 ± 0.002	3.97 ± 0.03	0.365 ± 0.003	0.221 ± 0.000	5.68 ± 0.02	0.783 ± 0.000
7	3.95 ± 0.01	0.391 ± 0.001	0.230 ± 0.000	6.06 ± 0.00	0.810 ± 0.002	4.03 ± 0.04	0.398 ± 0.002	0.225 ± 0.000	5.80 ± 0.01	0.791 ± 0.001
	$h_o = h_d = 24, h_t = 72$					$h_o = h_d = 36, h_t = 72$				
Day(s)	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
1	3.77 ± 0.07	0.337 ± 0.002	0.209 ± 0.001	5.87 ± 0.03	0.809 ± 0.001	3.75 ± 0.08	0.330 ± 0.002	0.202 ± 0.000	5.72 ± 0.03	0.805 ± 0.000
2	3.72 ± 0.04	0.343 ± 0.002	0.217 ± 0.001	5.89 ± 0.02	0.813 ± 0.000	3.73 ± 0.05	0.339 ± 0.002	0.207 ± 0.000	5.70 ± 0.04	0.797 ± 0.001
3	3.74 ± 0.03	0.362 ± 0.001	0.218 ± 0.001	5.86 ± 0.03	0.814 ± 0.001	3.78 ± 0.04	0.357 ± 0.002	0.218 ± 0.001	5.87 ± 0.02	0.807 ± 0.001
7	3.90 ± 0.04	0.389 ± 0.002	0.227 ± 0.000	5.97 ± 0.03	0.813 ± 0.001	3.98 ± 0.02	0.386 ± 0.001	0.229 ± 0.001	6.03 ± 0.03	0.810 ± 0.001
	$h_o = h_d = 24, h_t = 96$					$h_o = h_d = 36, h_t = 96$				
Day(s)	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**	**RMSE**	**WMAPE**	**CRPS**	**MIL 80%**	**ICP 80%**
1	3.79 ± 0.05	0.329 ± 0.003	0.204 ± 0.001	5.71 ± 0.01	0.807 ± 0.000	3.73 ± 0.07	0.331 ± 0.002	0.204 ± 0.001	5.67 ± 0.02	0.810 ± 0.001
2	3.60 ± 0.04	0.338 ± 0.002	0.206 ± 0.001	5.77 ± 0.02	0.807 ± 0.001	3.61 ± 0.05	0.337 ± 0.002	0.207 ± 0.001	5.79 ± 0.02	0.812 ± 0.000
3	3.79 ± 0.04	0.361 ± 0.002	0.213 ± 0.001	6.01 ± 0.03	0.810 ± 0.001	3.61 ± 0.04	0.350 ± 0.001	0.214 ± 0.000	5.96 ± 0.02	0.806 ± 0.001
7	3.88 ± 0.04	0.380 ± 0.001	0.218 ± 0.001	5.97 ± 0.02	0.811 ± 0.000	3.86 ± 0.02	0.380 ± 0.001	0.218 ± 0.000	5.87 ± 0.03	0.806 ± 0.001

the periodic module exhibits a computational complexity of $\mathcal{O}((N_o + N_d) \cdot T_{i_p})$ per training epoch, which scales linearly with the number of origins/destinations and the number of time intervals within a period. The calculation of residuals $\boldsymbol{\mathcal{F}}$, as formulated in Equations (6) and (7), also scales linearly with $\mathcal{O}(N_o \cdot h_o \cdot h_d \cdot h_t + N_o \cdot N_d \cdot h_d \cdot h_t + N_o \cdot N_d \cdot T \cdot h_t)$. For the transition and inference of latent variables $\mathbf{z}_t \in \mathrm{R}^{h_t}$, the use of multi-layer perceptrons (MLPs) yields a complexity of $\mathcal{O}(N_o \cdot N_d \cdot h + (k-1) \cdot h^2) + h \cdot h_t$, where h is the number of neurons per layer and k denotes the number of layers. Following the universal approximation theorem for MLPs, h may increase with $N_o \cdot N_d$ to ensure sufficient representation capability. While the model remains computationally tractable for moderate-sized networks, its runtime naturally grows with the number of OD pairs, comparable to transformer-based models [44], in large-scale systems.

VI. CONCLUSION

In this paper, we investigate the task of long-term origin-destination demand forecasting in a Bayesian framework. The proposed P-R SVI model integrates stochastic variational inference and tensor decomposition to capture spatiotemporal interactions in a stochastic manner. Being cast in a stochastic variational inference (SVI) framework, the resulting predictive distributions of P-R SVI do not belong to a single distribution family. The hierarchical nature of latent variables enhances the expressive power of P-R SVI so that it is capable of capturing complex distributional traits such as multi-mode and long-tail. Extensive experiments are performed on real OD demand datasets to demonstrate the superiority of P-R SVI over other popular baselines. The strengths of the P-R SVI model are reflected in the predictive accuracy and the accurate depiction of the predictive distributions. Therefore, P-R SVI has the potential to assist decision making processes in transportation by providing reliable OD demand forecasts. Overall, this paper is one of the first works to exploit the fusion of explainable periodic patterns and deep learning techniques for stochastic long-term OD demand forecasting.

As for the future research directions, first, it is worth attention to explore the integration of more complicated deep learning models into the SVI framework. Deep learning models demonstrate outstanding learning ability in various artificial intelligence tasks; however, how to incorporate uncertainty quantification into deep learning still awaits further exploration and is instrumental for solving many operational problems. Secondly, since P-R SVI is fully built upon latent variables and training data consist of historical OD demand time series solely, one could further examine how the inclusion of exogenous covariates could improve the predictive performance. Currently, such datasets including social media, socio-economics, and point of interest (POI) information are easy to collect.

ACKNOWLEDGMENTS

The work described in this paper is supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/15206322 and PolyU/15227424). The contents of this article reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein.

APPENDIX

In the appendix, we present the derivation of Equations (24) and (28), which is omitted in the paper to save space.

A. Derivation of Equation (24)

$$\min_{\theta,\phi} \text{obj} = -[\log(p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}|\boldsymbol{Z})p_\theta(\boldsymbol{Z})) - \log q_\phi(\boldsymbol{Z})] \tag{35}$$

$$= -\log \prod_{i=1}^{N_o}\prod_{j=1}^{N_d}\prod_{t=1}^{T} p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t}|\boldsymbol{\mathcal{P}}_{i,j,t}, \sum_{k_s=1}^{t} s_{k_s}) + \log q_\phi(\boldsymbol{Z}) + \text{const.} \tag{36}$$

$$= -\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T} \log p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t}|\boldsymbol{\mathcal{P}}_{i,j,t}, \sum_{k_s=1}^{t} s_{k_s}) + \sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o} \log q_\phi(\mathbf{P}^{i_p,o}_{n_o,t_{i_p}}) + \sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d} \log q_\phi(\mathbf{P}^{i_p,d}_{n_d,t_{i_p}}) + \sum_{k_s=1}^{T} \log q_\phi(s_{k_s}) + \text{const.} \tag{37}$$

$$= \frac{1}{2}\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T} \frac{(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t} - (\boldsymbol{\mathcal{P}}_{i,j,t} + \sum_{k_s=1}^{t} s_{k_s}))^2}{\sigma^2_{o,i} + \sigma^2_{d,j}} - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o} \frac{(\mathbf{P}^{i_p,o}_{n_o,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,o}_{n_o,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o,t_{i_p}}} - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d} \frac{(\mathbf{P}^{i_p,d}_{n_d,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,d}_{n_d,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,d^2}_{n_d,t_{i_p}}} - \frac{1}{2}\sum_{k_s=1}^{T} \frac{(s_{k_s} - \mu_{s,k_s})^2}{\sigma^2_{s,k_s}} + \text{const.} \tag{38}$$

B. Derivation of Equation (28)

$$\min_{\theta,\phi} \text{obj} = -[\log(p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}|\boldsymbol{Z})p_\theta(\boldsymbol{Z})) - \log q_\phi(\boldsymbol{Z})] \tag{39}$$

$$= -\log \prod_{i=1}^{N_o}\prod_{j=1}^{N_d}\prod_{t=1}^{T} p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t}|\boldsymbol{\mathcal{P}}_{i,j,t}, \boldsymbol{\mathcal{F}}_{i,j,t}) - \log \prod_{t=1}^{T} p_\theta(\mathbf{z}_t|\mathbf{z}_{t-1}) + \log q_\phi(\boldsymbol{Z}) + \text{const.} \tag{40}$$

$$= -\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T} \log p_\theta(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t}|\boldsymbol{\mathcal{P}}_{i,j,t}, \boldsymbol{\mathcal{F}}_{i,j,t}) - \sum_{t=1}^{T} \log p_\theta(\mathbf{z}_t|f(\mathbf{z}_{t-1})) + \sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o} \log q_\phi(\mathbf{P}^{i_p,o}_{n_o,t_{i_p}}) + \sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d} \log q_\phi(\mathbf{P}^{i_p,d}_{n_d,t_{i_p}}) + \sum_{t=1}^{T} \log q_\phi(\mathbf{z}_t|g_1(\mathbf{z}_{t-1}, g_2(\boldsymbol{\mathcal{X}}^{\text{obs}}_{:,:,t} - \boldsymbol{\mathcal{P}}_{:,:,t}))) + \text{const.} \tag{41}$$

$$= \frac{1}{2}\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T} \frac{(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t} - (\boldsymbol{\mathcal{P}}_{i,j,t} + \boldsymbol{\mathcal{F}}_{i,j,t}))^2}{\sigma^2_{o,i} + \sigma^2_{d,j}} + \frac{1}{2}\sum_{t=1}^{T} \log\det(\text{diag}(\boldsymbol{\sigma}^2_{z,t})) + \frac{1}{2}\sum_{t=1}^{T} (\mathbf{z}_t - \boldsymbol{\mu}_{z,t})^T (\text{diag}(\boldsymbol{\sigma}^2_{z,t}))^{-1} (\mathbf{z}_t - \boldsymbol{\mu}_{z,t}) - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o} \frac{(\mathbf{P}^{i_p,o}_{n_o,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,o}_{n_o,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o,t_{i_p}}} - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d} \frac{(\mathbf{P}^{i_p,d}_{n_d,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,d}_{n_d,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,d^2}_{n_d,t_{i_p}}} - \frac{1}{2}\sum_{t=1}^{T} \log\det(\text{diag}(\tilde{\boldsymbol{\sigma}}^2_{z,t})) - \frac{1}{2}\sum_{t=1}^{T} (\mathbf{z}_t - \tilde{\boldsymbol{\mu}}_{z,t})^T (\text{diag}(\tilde{\boldsymbol{\sigma}}^2_{z,t}))^{-1} (\mathbf{z}_t - \tilde{\boldsymbol{\mu}}_{z,t}) + \text{const.} \tag{42}$$

$$= \frac{1}{2}\sum_{i=1}^{N_o}\sum_{j=1}^{N_d}\sum_{t=1}^{T} \frac{(\boldsymbol{\mathcal{X}}^{\text{obs}}_{i,j,t} - (\boldsymbol{\mathcal{P}}_{i,j,t} + \boldsymbol{\mathcal{F}}_{i,j,t}))^2}{\sigma^2_{o,i} + \sigma^2_{d,j}} + \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{h_t} \log(\boldsymbol{\sigma}^2_{z,t,i_t}) + \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{h_t} \frac{(\mathbf{z}_{t,i_t} - \boldsymbol{\mu}_{z,t,i_t})^2}{\boldsymbol{\sigma}^2_{z,t,i_t}} - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_o=1}^{N_o} \frac{(\mathbf{P}^{i_p,o}_{n_o,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,o}_{n_o,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,o^2}_{n_o,t_{i_p}}} - \frac{1}{2}\sum_{i_p=1}^{2}\sum_{t_{i_p}=1}^{T_{i_p}}\sum_{n_d=1}^{N_d} \frac{(\mathbf{P}^{i_p,d}_{n_d,t_{i_p}} - \tilde{\boldsymbol{\mu}}^{i_p,d}_{n_d,t_{i_p}})^2}{\tilde{\boldsymbol{\sigma}}^{i_p,d^2}_{n_d,t_{i_p}}} - \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{h_t} \log \tilde{\boldsymbol{\sigma}}^2_{z,t,i_t} - \frac{1}{2}\sum_{t=1}^{T}\sum_{i_t=1}^{h_t} \frac{(\mathbf{z}_{t,i_t} - \tilde{\boldsymbol{\mu}}_{z,t,i_t})^2}{\tilde{\boldsymbol{\sigma}}^2_{z,t,i_t}} + \text{const.} \tag{43}$$

REFERENCES

[1] K. W. Axhausen and T. Gärling, "Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems," *Transport reviews*, vol. 12, no. 4, pp. 323–341, 1992.

[2] G. J. Grindey, S. M. Amin, E. Y. Rodin, and A. Garcia-Ortiz, "Kalman filter approach to traffic modeling and prediction," in *Intelligent Transportation Systems*, vol. 3207. SPIE, 1998, pp. 234–240.

[3] T. Djukic, G. Flötteröd, H. Van Lint, and S. Hoogendoorn, "Efficient real time od matrix estimation based on principal component analysis," in *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2012, pp. 115–121.

[4] X. Chen and L. Sun, "Bayesian temporal factorization for multidimensional time series prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4659–4673, 2021.

[5] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE transactions on intelligent transportation systems*, vol. 12, no. 1, pp. 141–151, 2010.

[6] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 186–194.

[7] Y. Wang, H. Yin, H. Chen, T. Wo, J. Xu, and K. Zheng, "Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1227–1235.

[8] K. Zhang, Z. Liu, and L. Zheng, "Short-term prediction of passenger demand in multi-zone level: Temporal convolutional neural network with multi-task learning," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 4, pp. 1480–1490, 2019.

[9] J. Ke, X. Qin, H. Yang, Z. Zheng, Z. Zhu, and J. Ye, "Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network," *Transportation Research Part C: Emerging Technologies*, vol. 122, p. 102858, 2021.

[10] J. Kwon and P. Varaiya, "Real-time estimation of origin–destination matrices with partial trajectories from electronic toll collection tag data," *Transportation Research Record*, vol. 1923, no. 1, pp. 119–126, 2005.

[11] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system," in *Proceedings of the ACM SIGKDD international workshop on urban computing*, 2012, pp. 142–148.

[12] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model," *Transportation Research Part B: Methodological*, vol. 83, pp. 121–135, 2016.

[13] Y. Li and C. Shahabi, "A brief overview of machine learning methods for short-term traffic forecasting and future directions," *Sigspatial Special*, vol. 10, no. 1, pp. 3–9, 2018.

[14] D. Zhuang, S. Hao, D.-H. Lee, and J. G. Jin, "From compound word to metropolitan station: Semantic similarity analysis using smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 114, pp. 322–337, 2020.

[15] Z. Cheng, S. Jian, T. H. Rashidi, M. Maghrebi, and S. T. Waller, "Integrating household travel survey and social media data to improve the quality of od matrix: A comparative case study," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 6, pp. 2628–2636, 2020.

[16] Z. Xu, Z. Lv, J. Li, H. Sun, and Z. Sheng, "A novel perspective on travel demand prediction considering natural environmental and socioeconomic factors," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 1, pp. 136–159, 2022.

[17] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.

[18] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.

[19] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "Lstm network: a deep learning approach for short-term traffic forecast," *IET intelligent transport systems*, vol. 11, no. 2, pp. 68–75, 2017.

[20] F. Toqué, E. Côme, M. K. El Mahrsi, and L. Oukhellou, "Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks," in *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*. IEEE, 2016, pp. 1071–1076.

[21] K.-F. Chu, A. Y. Lam, and V. O. Li, "Deep multi-scale convolutional lstm network for travel demand and origin-destination predictions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3219–3232, 2019.

[22] Z. Duan, K. Zhang, Z. Chen, Z. Liu, L. Tang, Y. Yang, and Y. Ni, "Prediction of city-scale dynamic taxi origin-destination flows using a hybrid deep neural network combined with travel time," *IEEE Access*, vol. 7, pp. 127 816–127 832, 2019.

[23] D. Li, J. Cao, R. Li, and L. Wu, "A spatio-temporal structured lstm model for short-term prediction of origin-destination matrix in rail transit with multisource data," *IEEE Access*, vol. 8, pp. 84 000–84 019, 2020.

[24] K. F. Chu, A. Y. Lam, and V. O. Li, "Travel demand prediction using deep multi-scale convolutional lstm network," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 1402–1407.

[25] Y. Wang and M. Papageorgiou, "Real-time freeway traffic state estimation based on extended kalman filter: a general approach," *Transportation Research Part B: Methodological*, vol. 39, no. 2, pp. 141–167, 2005.

[26] J. Guo, W. Huang, and B. M. Williams, "Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[27] J. Ren and Q. Xie, "Efficient od trip matrix prediction based on tensor decomposition," in *2017 18th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2017, pp. 180–185.

[28] Z. Cheng, M. Trepanier, and L. Sun, "Real-time forecasting of metro origin-destination matrices with high-order weighted dynamic mode decomposition," *Transportation science*, vol. 56, no. 4, pp. 904–918, 2022.

[29] Y. Ma, R. Kuik, and H. J. van Zuylen, "Day-to-day origin–destination tuple estimation and prediction with hierarchical bayesian networks using multiple data sources," *Transportation research record*, vol. 2343, no. 1, pp. 51–61, 2013.

[30] D. Gammelli, I. Peled, F. Rodrigues, D. Pacino, H. A. Kurtaran, and F. C. Pereira, "Estimating latent demand of shared mobility through censored gaussian processes," *Transportation Research Part C: Emerging Technologies*, vol. 120, p. 102775, 2020.

[31] X. Fu, H. Yang, C. Liu, J. Wang, and Y. Wang, "A hybrid neural network for large-scale expressway network od prediction based on toll data," *PloS one*, vol. 14, no. 5, p. e0217241, 2019.

[32] D. Koca, J. D. Schmöcker, and K. Fukuda, "Origin-destination matrix estimation by deep learning using maps with new york case study," in *2021 7th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 2021, pp. 1–6.

[33] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, "Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3377–3391, 2020.

[34] J. Zhang, H. Che, F. Chen, W. Ma, and Z. He, "Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method," *Transportation Research Part C: Emerging Technologies*, vol. 124, p. 102928, 2021.

[35] L. Liu, Z. Qiu, G. Li, Q. Wang, W. Ouyang, and L. Lin, "Contextualized spatial–temporal network for taxi origin-destination demand prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3875–3887, 2019.

[36] X. Zou, S. Zhang, C. Zhang, J. James, and E. Chung, "Long-term origin-destination demand prediction with graph deep learning," *IEEE Transactions on Big Data*, vol. 8, no. 6, pp. 1481–1495, 2021.

[37] H. Shi, Q. Yao, Q. Guo, Y. Li, L. Zhang, J. Ye, Y. Li, and Y. Liu, "Predicting origin-destination flow via multi-perspective graph convolutional network," in *2020 IEEE 36th International conference on data engineering (ICDE)*. IEEE, 2020, pp. 1818–1821.

[38] X. Xiong, K. Ozbay, L. Jin, and C. Feng, "Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter," *Transportation Research Record*, vol. 2674, no. 8, pp. 491–503, 2020.

[39] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, vol. 1. IEEE, 1994, pp. 55–60.

[40] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International journal of forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[41] Q. Wang, S. Wang, D. Zhuang, H. Koutsopoulos, and J. Zhao, "Uncertainty quantification of spatiotemporal travel demand with probabilistic graph neural networks," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[42] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[45] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[47] G. Jin, Y. Cui, L. Zeng, H. Tang, Y. Feng, and J. Huang, "Urban ride-hailing demand prediction with multiple spatio-temporal information fusion network," *Transportation Research Part C: Emerging Technologies*, vol. 117, p. 102665, 2020.

[48] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[49] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower upper bound estimation method for construction of neural network-based prediction intervals," *IEEE transactions on neural networks*, vol. 22, no. 3, pp. 337–346, 2010.

[50] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *International conference on machine learning*. PMLR, 2018, pp. 4075–4084.

[51] F. Rodrigues and F. C. Pereira, "Beyond expectation: Deep joint mean and quantile regression for spatiotemporal problems," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 12, pp. 5377–5389, 2020.

[52] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[53] L. Lin, Z. He, and S. Peeta, "Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 258–276, 2018.

[54] G. Guo and T. Zhang, "A residual spatio-temporal architecture for travel demand forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 115, p. 102639, 2020.

[55] S. Hu and C. Xiong, "High-dimensional population inflow time series forecasting via an interpretable hierarchical transformer," *Transportation research part C: emerging technologies*, vol. 146, p. 103962, 2023.

[56] Q. Zhou, X. Lu, J. Gu, Z. Zheng, B. Jin, and J. Zhou, "Explainable origin-destination crowd flow interpolation via variational multi-modal recurrent graph auto-encoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 9422–9430.

[57] A. Alwosheel, S. van Cranenburgh, and C. G. Chorus, "Why did you predict that? towards explainable artificial neural networks for travel demand analysis," *Transportation Research Part C: Emerging Technologies*, vol. 128, p. 103143, 2021.

[58] T. Kim, S. Sharda, X. Zhou, and R. M. Pendyala, "A stepwise interpretable machine learning framework using linear regression (lr) and long short-term memory (lstm): City-wide demand-side prediction of yellow taxi and for-hire vehicle (fhv) service," *Transportation Research Part C: Emerging Technologies*, vol. 120, p. 102786, 2020.

[59] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[60] X. Jiang, L. Zhang, and X. M. Chen, "Short-term forecasting of high-speed rail demand: A hybrid approach combining ensemble empirical mode decomposition and gray support vector machine with real-world applications in china," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 110–127, 2014.

[61] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.

[62] M. Gu and Z. Duan, "Daily od demand prediction in urban metro transit system: A convolutional lstm neural network with multi-factor fusion channel-wise attention," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3398–3404.

[63] S. Ken-Iti, *Lévy processes and infinitely divisible distributions*. Cambridge university press, 1999, vol. 68.

[64] "TLC Trip Record Data," https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, NYC Taxi and Limousine Commission.

[65] "BART Hourly Ridership Data by Origin-Destination Pairs," https://www.bart.gov/about/reports/ridership, Bay Area Rapid Transit District.

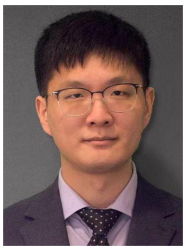
[66] New York City Department of City Planning, "New york city 2020 neighborhood tabulation areas (nta), edition 25c," BYTES of the BIG APPLE, Aug. 2025, pDF metadata sheet. [Online]. Available: https://s-media.nyc.gov/agencies/dcp/assets/files/pdf/data-tools/bytes/nynta2020_metadata.pdf

[67] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[68] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

Zihan Wan obtained the bachelor's degree in mathematics and statistics from University of Toronto, Canada, and the master's degree in applied mathematics from Columbia Univeristy, USA. Currently he is pursuing the Ph.D. degree in Department of Civil and Environmental Engineering, the Hong Kong Polytechnic University. His research interests include machine learning, statistical learning, and statistical modeling of transportation systems.

Zhenjie Zheng obtained the bachelor's degree in industrial engineering from Huazhong University of Science and Technology, China, and the Ph.D. degree in industrial engineering from Tsinghua University, China. He is currently a postdoctoral fellow with the Department of Civil and Environmental Engineering, the Hong Kong Polytechnic University. His research interests include machine learning in transportation management, AI-powered network-wide traffic state estimation, data-driven city emergency detection and management, and LoT-based intelligent traffic control.



Wei Ma (IEEE member) received bachelor's degrees in Civil Engineering and Mathematics from Tsinghua University, China, master degrees in Machine Learning and Civil and Environmental Engineering, and PhD degree in Civil and Environmental Engineering from Carnegie Mellon University, USA. He is currently an associate professor with the Department of Civil and Environmental Engineering at the Hong Kong Polytechnic University (PolyU). His research focuses on intersection of machine learning, data mining, and transportation network modeling, with applications for smart and sustainable mobility systems. He has received 2020 Mao Yisheng Outstanding Dissertation Award and best paper award (theoretical track) at INFORMS Data Mining and Decision Analytics Workshop. Dr. Ma also serves as the Associate Editor of IEEE T-ITS and OJ-ITS.