Uncertainty-aware probabilistic travel demand forecasting for Mobility-on-Demand services

Tao Peng*[1], Jie Gao[1], and Oded Cats[1]

[1]Department of Transport & Planning, Delft University of Technology, the Netherlands

Short summary

Past work has primarily focused on improving the accuracy of travel demand forecasting, often overlooking the inherent uncertainty in such predictions. To this end, we propose an innovative non-parametric uncertainty-aware probabilistic framework for travel demand forecasting in Mobility-on-Demand services. The framework employs a spatiotemporal graph neural network to learn and extract features from city-level travel demand data. These features are then processed through the designed variational autoencoder, which compresses the information and applies resampling and decoding operations to generate forecast samples. A kernel density estimation transforms these samples into a predictive distribution, producing accurate, confident, and well-calibrated predictions. Comprehensive experiments on a real-world dataset, evaluated across multiple metrics and benchmarked against four baseline models, demonstrate the superior performance of the proposed model in both point forecasting and probabilistic forecasting. This framework offers a robust and extensible tool for quantifying uncertainty in future travel demand.
Keywords: Mobility-on-Demand services, Probabilistic forecasting, Travel demand, Variational autoencoder

1 Introduction

Accurate travel demand forecasting is essential to effectively manage transportation systems, as it supports informed decisions that optimize resource allocation and improve service outcomes. However, travel demand is inherently uncertain, influenced by dynamic factors such as weather conditions, road incidents, and human behavior. These uncertainties present significant challenges, particularly for Mobility-on-Demand (MoD) platforms like Uber, DiDi, and Lyft, which process millions of demand-related queries daily. Forecasting methods that fail to account for these uncertainties produce unreliable predictions, which can lead to inefficient fleet management, delayed response times, and decreased user satisfaction, ultimately affecting the operational success of MoD platforms and the broader transportation systems.

Recognizing its significance, travel demand forecasting has been extensively studied in both industry and academia. For example, advanced neural network architectures have been proposed to improve prediction accuracy. In this stream of work, temporal trends such as seasonal, weekly, and time-of-day patterns are typically captured using Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU) layers Ke et al. (2017), L. Liu et al. (2020). Meanwhile, spatial dependencies are modeled using convolutional neural networks (CNNs), which analyze urban areas divided into zones or grids Wu et al. (2021). More recently, graph-based approaches have gained traction, representing zones as nodes and travel flows as edges, with Graph Convolutional Networks (GCNs) effectively extracting spatiotemporal features Yu et al. (2017). Despite these advancements in prediction accuracy, these methods focus on point forecasting, aiming to forecast a specific future demand value. However, such deterministic predictions fail to account for the inherent uncertainty in travel demand, stemming from factors like road incidents, weather conditions, and human behavior. Ignoring these natural uncertainties can lead to biased predictions, limiting the effectiveness of planning and decision-making, which, in turn, reduces user satisfaction levels and the system's long-term performance.

To address uncertainty, another line of research has focused on estimating travel demand distributions by combining deterministic and probabilistic components. For example, Wang et al. (2024)

propose a framework of probabilistic graph neural networks (Prob-GNN) to quantify the spatiotemporal uncertainty of travel demand with the assumption that travel demand follows Gaussian and Poisson distributions for distribution estimation. Similarly, Jin et al. (2024) propose a spatial-temporal uncertainty-aware graph network framework (STUP), featuring a dedicated uncertainty-aware block to estimate the parameters of an assumed Gaussian traffic distribution. Furthermore, with the assumption that travel demand follows a zero-inflated negative binomial (ZINB) distribution for the numerous zeros in sparse O-D matrices and a negative binomial (NB) distribution for each non-zero entry, Zhuang et al. (2022) propose a Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network (STZINBGNN) to quantify the uncertainty of the sparse travel demand. However, these methods often rely on strong parametric assumptions about the data distribution. For example, assuming that travel demand follows pre-defined distributions (e.g., Gaussian). This may introduce bias if the actual distribution deviates significantly from those assumed. Non-parametric methods, in contrast, avoid such assumptions, offering greater flexibility in modeling diverse data patterns. Applications of non-parametric techniques, such as kernel density estimation in energy prediction He & Li (2018), Gu et al. (2021), traffic flow prediction Li et al. (2024) and label-smoothing methods for travel time distribution H. Liu et al. (2023), highlight their potential for uncertainty quantification. Yet, their application in travel demand forecasting remains largely unexplored.

To this end, we design a novel uncertainty-aware probabilistic travel demand forecasting framework which is a non-parametric approach to quantify the uncertainty in travel demand. The designed framework integrates a spatiotemporal graph convolutional network (STGCN), a variational autoencoder (VAE) module and a kernel density estimation(KDE) technique to provide flexible, accurate and confident forecasting of travel demand. The STGCN effectively captures spatial and temporal dependencies in travel demand, while the VAE module compresses and resamples the learned features to generate diverse forecast samples. These samples are then transformed into a predictive distribution using KDE, which avoids restrictive parametric assumptions and ensures flexibility in capturing complex data patterns. This integrated design enables the framework to produce accurate, confident, and well-calibrated probabilistic forecasts. Extensive experiment on a real-world dataset shows that the designed approach outperforms state-of-the-art demand forecasting and uncertainty quantification methods across multiple evaluation metrics.

2 Methodology

In this section, we first formalize the learning problem of spatiotemporal travel demand distribution forecasting. Then we introduce the proposed spatiotemporal graph neural network-based VAE framework.

Problem description

We consider a travel demand distribution forecasting problem where the service area is partitioned into n distinct regions. Let $V = \{v_1, v_2, \cdots, v_n\}$ denote the set of all regions, where each $v_i \in V$ represents a unique region within the area of interest, e.g. a city. For simplicity, we drop the subscript i and represent a region as v. Based on this partitioning, we construct a graph representation $\mathcal{G} = (V, A)$, where V is a set of nodes (regions) and A is an adjacency matrix that represents the connections between them. Let x_v^t denote the number of trip orders in the region $v \in V$ during the t^{th} time interval, where $x_v^t \in \mathbb{R}$. We then define $X^t \in \mathbb{R}^{n \times 1}$ as the number of orders in all regions at the t^{th} time interval, with x_v^t as its entry for each region v. For a sequence of T time intervals, the travel demand sequence is denoted as $X^{t-T+1:t} = [X^{t-T+1}, \dots X^{t-1}, X^t]$. This sequence captures the historical demand over the past T intervals. Given the defined travel demand sequence, the demand forecasting problem is modeled as a function of the time-dependent historical demand sequence. Formally, given $X^{t-T+1:t}$, the goal is to forecast the *conditional probability distribution* P of travel demand at the next time step, X^{t+1}, which is represented as:

$$P(X^{t+1}|X^{t-T+1:t}) \tag{1}$$

STGCN+VAE design

Figure 1 presents the overall structure of our proposed approach, which consists of three key modules. The first module leverages a spatiotemporal graph convolutional network (STGCN) to

capture the complex spatial and temporal dependencies inherent in travel demand data, modeling both spatial interactions between regions and temporal patterns across time intervals. These learned features are then processed through a variational autoencoder (VAE) module, where the encoder compresses the high-dimensional spatiotemporal information into a latent representation, effectively preserving critical features while reducing dimensionality. The decoder subsequently resamples from this latent representation to generate multiple predictions, representing a range of possible outcomes that capture the uncertainty present in real-world travel demand. To model the complete probability distribution of travel demand, we apply a non-parametric kernel density estimation (KDE) technique. This module constructs a continuous, data-driven probability distribution based on the decoder's outputs, unconstrained by predefined distributional assumptions, thereby offering a flexible and accurate representation of demand uncertainty. In the following paragraphs, we provide an overall description of each module within our proposed approach.

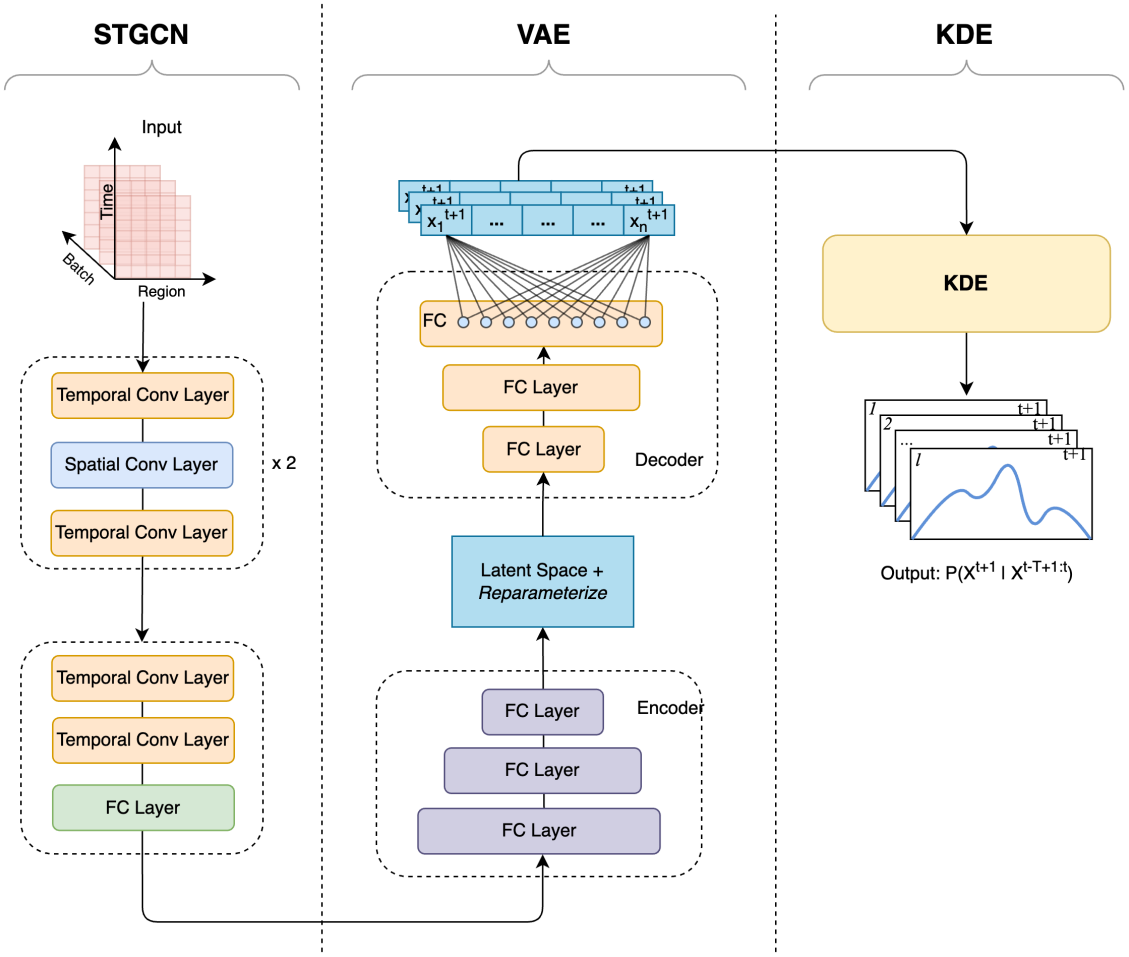

Figure 1: Overall architecture of the proposed method. During training, the STGCN and VAE modules are trained, while the latent space, decoder, and KDE are employed in the generation step.

Spatio-Temporal Graph Convolutional Networks (STGCN) As shown on the left side of Figure 1, the STGCN is designed to model both spatial dependencies (i.e., how regions are related in space through a graph) and temporal dependencies (i.e., how features of regions evolve over time) of travel demand data. It combines graph convolution layers to capture spatial relationships with temporal convolution layers to model dynamic changes over time. Our STGCN design follows the approach outlined in Yu et al. (2017).

The *temporal* convolution aspect of STGCN captures patterns in travel demand data over time, as shown in Figure 2. Specifically, given a time series sequence $X_v = [x_v^{t-T+1}, \ldots, x_v^t]$ at a region v, where $X_v \in \mathbb{R}^T$, the model applies a convolution kernel $\Gamma \in \mathbb{R}^{K \times 2C_o}$. Here, K represents the kernel size, indicating the length of the time window over which the convolution operates, and

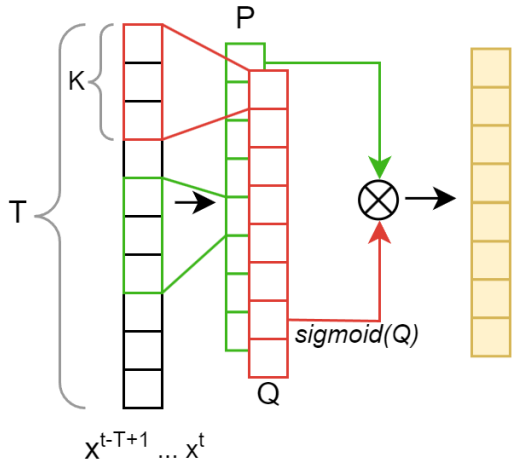

Figure 2: Temporal convolution block structure.

$2C_0$ defines the total number of output channels produced by the kernel. As the kernel slides over the sequence, it extracts temporal features, producing an output matrix $[PQ] \in \mathbb{R}^{(T-K+1)\times(2C_o)}$. This matrix is then split along the channel dimension into two parts, P and Q, each of dimension $(T-K+1)\times(C_o)$. Then, P, Q are applied with gated linear units (GLU) to control the information flows.

The *spatial* convolution block extracts spatial dependencies in travel demand data using a spectral-based graph convolution approach. This approach, introduced by Bruna et al. (2013), leverages the graph Fourier transform, which is defined through the eigenvectors of the graph Laplacian matrix. Although spectral convolution offers a mathematically grounded framework, it suffers from computational inefficiency due to eigen-decomposition. To address this limitation, Defferrard et al. (2016) propose an approximation using Chebyshev polynomials, significantly improving computational efficiency. This approximation ensures computational efficiency while preserving accuracy. To integrate spatial and temporal features, the spatio-temporal convolutional block (ST-Conv block) is designed to jointly process graph-structured time series, facilitating coherent exploration of dependencies. An output layer synthesizes these features as the input of the following VAE module.

Variational autoencoder (VAE) The VAE architecture in the model compresses information and generates new data samples. As shown in Figure 1, the encoder, consisting of three fully connected layers, processes features learned from the STGCN module into a dense latent representation *Lat*, which captures essential travel demand patterns while discarding redundancies. To enhance diversity and account for real-world uncertainty, the reparameterize operation is applied to *Lat* before decoding, enabling the model to explore the latent space and produce a distribution of possible outcomes. The decoder reconstructs the next-step travel demand, denoted as $\hat{X}^{(t+1)}$, with training aimed at minimizing the mean squared error (MSE) loss between reconstructed travel demand and ground truth. After training, the model generates multiple demand samples efficiently using the decoder module of VAE. These samples are used to construct a travel demand distribution through KDE, providing a probabilistic view of demand for each region.

3 Application, results and discussion

To evaluate the proposed approach, we conduct experiments using the Yellow Taxi Trip Records dataset from the New York City TLC Trip Record [1], a public dataset that includes all MoD travel orders in the city. As shown in Figure 3, New York City is divided into 263 regions, with Manhattan having the highest travel demand. Our experiments use data from January to March 2024. Among the dataset's 19 features, we retain only pickup dates, times, and locations, resulting in 9,554,778

[1] `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

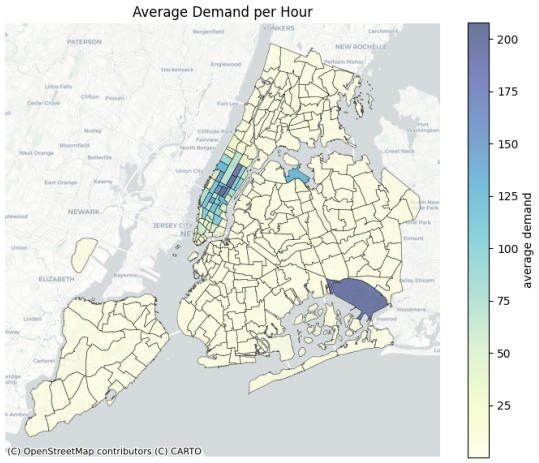

Figure 3: Average travel demand heatmap across New York City's regions. The dark regions represent higher travel demand.

records. These records are aggregated into one-hour prediction intervals, with a new field, "demand number", introduced to represent the aggregated demand. After the data processing, we split the entire data into training set (Jan. 1st - Mar. 13th), validation set (Mar. 14th - Mar. 21st) and test set (Mar. 22nd - Mar. 31st) for training, validation, and testing, respectively.

Evaluation metrics and benchmarks We adopt five commonly used metrics to evaluate the performance of our approach, namely mean absolute error (MAE), root mean squared error (RMSE), Mean Prediction Interval Width (MPIW), Continuous Ranked Probability Score (CRPS), and Interval Score (IS). MAE and RMSE assess point forecasting accuracy, while MPIW, CRPS, and IS measure probabilistic forecasting performance, where lower values indicate better results. To analyze performance variations, regions with demand below 10 are classified as low demand, and those with 10 or more are classified as high demand. This categorization reduces the dominance of high demand regions in the evaluation. We compare our proposed approach with widely used benchmarks, including a point forecasting model, STGCN, and three probabilistic forecasting models: STGCN combined with a normal distribution (STGCN + Normal), STGCN combined with a log-normal distribution (STGCN + Log-normal) and Deep Autoregressive recurrent network (DeepAR) proposed by Salinas et al. (2020). This comparison enables us to evaluate our method's performance in both single-value and distributional prediction performance.

Results The prediction results of different models in different demand scenarios are shown in Table 1, with the best result highlighted in bold. The proposed approach, STGCN+VAE, consistently achieves the best performance across all metrics when evaluated for all regions combined as well as for high-demand regions in particular, delivering the lowest MAE, RMSE, MPIW, CRPS, and IS values. This highlights the unique strength of STGCN+VAE in providing both precise predictions and reliable uncertainty quantification. In low-demand regions, STGCN+VAE shows slightly lower performance on certain metrics compared to DeepAR. This difference is likely due to data sparsity in low-demand areas, where frequent zero values and limited variability reduce the effectiveness of the sample-based learning approach used by STGCN+VAE. In contrast, DeepAR uses recurrent neural networks (RNNs), which can process sequences with varying numbers of non-zero values. However, low-demand regions generally have a limited impact on system-level decisions compared to high-demand regions. Therefore, the strengths of STGCN+VAE in capturing both accuracy and uncertainty in high-demand scenarios highlight its effectiveness and practical relevance for real-world applications.
To visually demonstrate the effectiveness of our proposed models, Figure 4 compares the ground truth with the ordered prediction intervals for six selected regions in the downtown area. The orange line represents the observed values, while the blue line and shaded area correspond to the predicted values and the 90% predictive interval of the proposed model, respectively. The

close alignment of the blue line with the orange line, along with the well-calibrated blue-shaded area, highlights a more accurate and precise prediction interval, showcasing the advantages of our proposed approach.

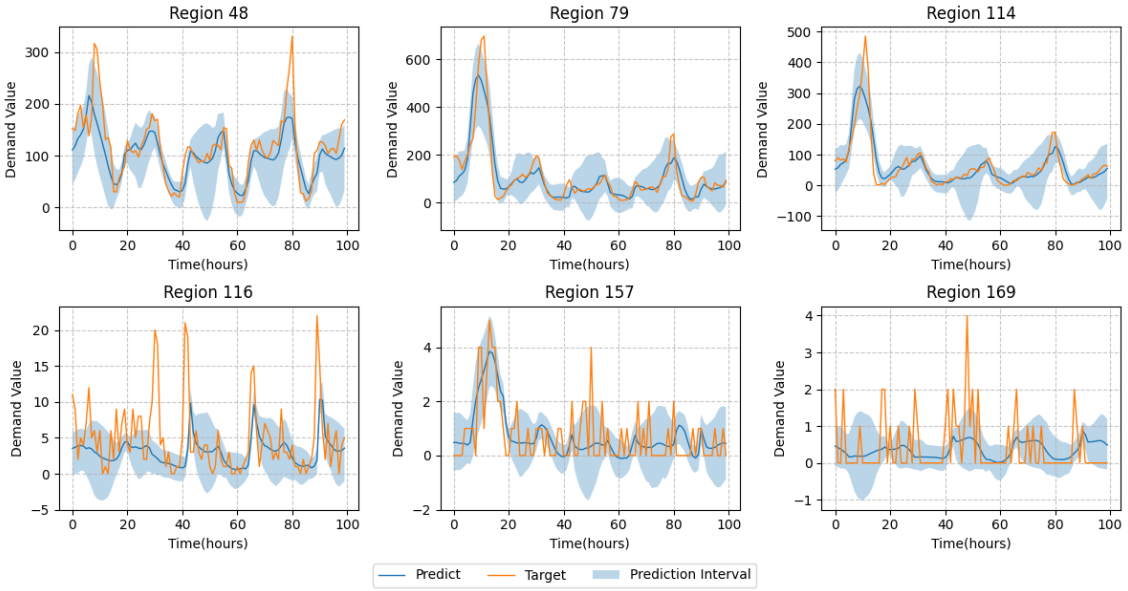

Figure 4: Comparison between observations and predictions for six example regions.

Table 1: Comparison of predictive results across models

Metric	Region	Point Forecasting	Probabilistic Forecasting			
		STGCN	STGCN+Normal	STGCN+Lognormal	DeepAR	STGCN+VAE
MAE	All	5.7487	5.8504	24.3598	11.0107	**5.4094**
	Low demand	0.9431	0.9548	1.7435	**0.7750**	0.8815
	High demand	27.2737	27.7788	125.6621	56.8581	**25.6908**
RMSE	All	17.1034	17.1731	64.9564	35.6128	**16.1368**
	Low demand	2.0968	0.9771	4.0361	2.7630	**0.9389**
	High demand	39.7884	5.2705	151.8075	83.1558	**5.0686**
MPIW	All	-	29.4285	13761.1612	29.1572	**23.8823**
	Low demand	-	3.3382	16234.3577	**1.2152**	6.0574
	High demand	-	146.2911	2683.3019	154.3143	**103.7229**
CRPS	All	-	22.8911	24.4993	20.9385	**17.1134**
	Low demand	-	3.5585	2.2019	**0.9861**	1.0643
	High demand	-	109.4855	124.3728	110.3097	**88.9996**
IS	All	-	76.7607	13761.1699	41.9638	**19.4128**
	Low demand	-	35.5485	16234.3681	**3.5794**	3.8665
	High demand	-	261.3572	2683.3020	213.8939	**89.0475**

Next, we demonstrate our findings in terms of uncertainty quantification by displaying the prediction intervals for three representative regions in the Manhattan area over two consecutive days, as shown in Figure 5. The selected days include a Sunday, representing a weekend, and a Monday, representing a typical workday. These regions—a workplace area, a tourism region, and a residential region—were chosen for their distinct travel demand patterns, providing a comprehensive view of how demand and uncertainty vary across different contexts.

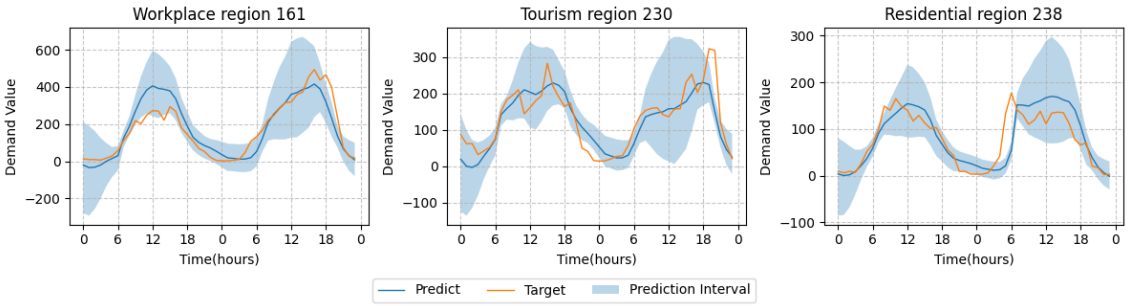

Figure 5: Uncertainty quantification across three representative regions over two consecutive days (March 23, 2024, Sunday, and March 24, 2024, Monday).

In the **workplace region**, travel demand is notably lower on Sunday compared to Monday, with the peak shifting from around 12 PM on Sunday to approximately 5 PM on Monday, aligning with the end of the workday. The highest uncertainty is observed around 12 PM on both days, likely because of diverse activities and variable travel behavior during midday. In the **tourism region**, the demand pattern remains relatively stable across both days, with demand stays consistently high during daylight hours and declines after 6 PM. The uncertainty is consistently high throughout the day, which can be attributed to the diverse nature of tourist activities. Unlike commuting patterns, which are predictable and structured, tourist behavior varies widely based on individual schedules, preferences, and external factors such as weather and events. The **residential region** exhibits similar overall demand levels on both days but with differences in peak times. On Sunday, peak demand occurs around 9–10 AM, driven by leisure activities, whereas on Monday, it shifts to 6 AM, corresponding to the morning commute. Uncertainty is notably lower during commuting periods across all regions, suggesting that these patterns are more consistent and predictable.
A common trend across all three regions is the highest uncertainty occurring around 12 PM. This may be attributed to the diversity of travel purposes during midday, introducing greater variability. Conversely, during commuting hours, the uncertainty is notably lower, likely because commuting patterns are more consistent and predictable. These findings demonstrate the robustness of our probabilistic approach in balancing predictive accuracy with uncertainty quantification, offering valuable insights for stakeholders to make informed decisions under varying demand conditions.

4 Conclusions

Recognizing the critical role of uncertainty in travel demand prediction, we propose a nonparametric probabilistic travel demand forecasting framework for MoD services, eliminating the need for strict data assumptions. The framework leverages the STGCN model to efficiently learn and extract features from historical data, followed by the designed VAE module to compress these features. Multiple forecast samples are generated through resampling and decoding to explore the latent travel demand distribution, which are then fitted using KDE to construct a predictive distribution for uncertainty quantification. Experimental results demonstrate that the proposed framework achieves high predictive accuracy and robust uncertainty quantification, with predictions closely aligning with real-world observations. In addition, its flexible design allows the STGCN module to be replaced with alternative feature extraction methods, highlighting its adaptability and potential for diverse applications. Currently, the designed model utilizes only demand-side data. However, in practice, travel demand is highly correlated with supply-side factors, such as vehicle availability and driver behavior. Integrating these factors to jointly predict demand and supply represents a promising direction for future research.

References

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, *29*.

Gu, B., Zhang, T., Meng, H., & Zhang, J. (2021). Short-term forecasting and uncertainty analysis of wind power based on long short-term memory, cloud model and non-parametric kernel density estimation. *Renewable Energy*, *164*, 687–708.

He, Y., & Li, H. (2018). Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. *Energy conversion and management*, *164*, 374–384.

Jin, X., Wang, J., Guo, S., Wei, T., Zhao, Y., Lin, Y., & Wan, H. (2024). Spatial–temporal uncertainty-aware graph networks for promoting accuracy and reliability of traffic forecasting. *Expert Systems with Applications*, *238*, 122143.

Ke, J., Zheng, H., Yang, H., & Chen, X. M. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation research part C: Emerging technologies*, *85*, 591–608.

Li, M., Guo, J., & Zhong, X. (2024). Real-time traffic flow uncertainty quantification based on nonparametric probability density function estimation. *Journal of Transportation Engineering, Part A: Systems*, *150*(11), 04024074.

Liu, H., Jiang, W., Liu, S., & Chen, X. (2023). Uncertainty-aware probabilistic travel time prediction for on-demand ride-hailing at didi. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining* (pp. 4516–4526).

Liu, L., Chen, J., Wu, H., Zhen, J., Li, G., & Lin, L. (2020). Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems*, *23*(4), 3377–3391.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, *36*(3), 1181–1191.

Wang, Q., Wang, S., Zhuang, D., Koutsopoulos, H., & Zhao, J. (2024). Uncertainty quantification of spatiotemporal travel demand with probabilistic graph neural networks. *IEEE Transactions on Intelligent Transportation Systems*.

Wu, Y., Zhuang, D., Labbe, A., & Sun, L. (2021). Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 4478–4485).

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Zhuang, D., Wang, S., Koutsopoulos, H., & Zhao, J. (2022). Uncertainty quantification of sparse travel demand prediction with spatial-temporal graph neural networks. In *Proceedings of the 28th acm sigkdd conference on knowledge discovery and data mining* (pp. 4639–4647).