Uncertainty Quantification of Sparse Travel Demand Prediction with Spatial-Temporal Graph Neural Networks

Dingyi Zhuang
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
dingyi@mit.edu

Shenhao Wang*
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
University of Florida
Gainesville, Florida, USA
shenhao@mit.edu
shenhaowang@ufl.edu

Haris Koutsopoulos
Northeastern University
Boston, Massachusetts, USA
h.koutsopoulos@northeastern.edu

Jinhua Zhao
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
jinhua@mit.edu

ABSTRACT

Origin-Destination (O-D) travel demand prediction is a fundamental challenge in transportation. Recently, spatial-temporal deep learning models demonstrate the tremendous potential to enhance prediction accuracy. However, few studies tackled the uncertainty and sparsity issues in fine-grained O-D matrices. This presents a serious problem, because a vast number of zeros deviate from the Gaussian assumption underlying the deterministic deep learning models. To address this issue, we design a Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network (STZINB-GNN) to quantify the uncertainty of the sparse travel demand. It analyzes spatial and temporal correlations using diffusion and temporal convolution networks, which are then fused to parameterize the probabilistic distributions of travel demand. The STZINB-GNN is examined using two real-world datasets with various spatial and temporal resolutions. The results demonstrate the superiority of STZINB-GNN over benchmark models, especially under high spatial-temporal resolutions, because of its high accuracy, tight confidence intervals, and interpretable parameters. The sparsity parameter of the STZINB-GNN has physical interpretation for various transportation applications.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

KEYWORDS

Spatial-temporal Sparse Data, Uncertainty Quantification, Graph Neural Networks, Travel Demand

ACM Reference Format:
Dingyi Zhuang, Shenhao Wang, Haris Koutsopoulos, and Jinhua Zhao. 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 (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539093

1 INTRODUCTION

It has attracted a lot of attention to predict the Origin-Destination (O-D) matrices of travel demand [10, 12, 40]. This task is challenging because the characteristics of the demand data vary with mobility services. For example, the O-D matrices for the subway stations are generally dense and probably satisfy the continuous Gaussian distribution, which implicitly underlies deterministic deep learning models. But for ride-hailing or bike-sharing services, the O-D zones could be much more granular, which lead to sparse and discrete entries in the O-D matrices. This sparsity issue becomes even more severe in the O-D matrices with high spatial-temporal resolution, because the initially dense O-D matrices could become much more diluted. As the mobility companies are increasingly adopting real-time interventions, the accurate prediction of the sparse and discrete O-D matrices at a high spatial-temporal resolution could significantly improve the quality of mobility services.

Although the prediction of dense and low-resolution O-D matrices has been extensively studied using deep learning (DL) models, few studies addressed the sparsity issue in the high-resolution travel demand data. The continuous data entries in the dense O-D matrices could follow Gaussian distributions [12, 13, 21, 30, 45, 46]. However, a large number of zero entries in a sparse O-D matrix evidently deviate from the Gaussian assumption. When the O-D matrix is sparse and dispersed with integer values, discrete distributions like the negative binomial distribution, would be more appropriate. Moreover, an enormous number of zeros could naturally emerge when the data set has a high spatial-temporal resolution. These zeros are important for transportation management, because they indicate areas with particularly low demand. Therefore, a successful prediction model should capture explicitly the zeros in the sparse

*Shenhao Wang is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '22, August 14–18, 2022, Washington, DC, USA.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539093

matrix and quantify their uncertainty, thus guiding the service allocation and management decisions.

To address sparsity, we propose a Spatial-Temporal Zero-Inflated Negative Binomial Graph Neural Network (STZINB-GNN) to quantify uncertainty and enhance prediction performance. We utilize zero-inflated negative binomial (ZINB) distributions to capture the enormous number of zeros in sparse O-D matrices, and the negative binomial (NB) distribution for each non-zero entry. Different from the variational autoencoder models, we design the spatial-temporal embedding with an additional parameter π to learn the likelihood of the inputs being zero. Our model utilizes the representation power of spatial-temporal graph neural networks to fit the parameters of probabilistic distributions. We compare a variety of probabilistic layers to assess the effectiveness of π in capturing sparsity. Empirically, we demonstrate the superiority of our model in the data set with fine-grained spatial-temporal resolution. Our main contributions include:

(1) We propose the STZINB-GNN to quantify the spatial-temporal uncertainty of O-D travel demand using a parameter π to learn data sparsity
(2) The parameters of the probabilistic GNNs successfully quantify the sparse and discrete uncertainty particularly in high-resolution data sets
(3) We demonstrate that the STZINB-GNN outperforms other models by using two real-world travel demand datasets with various spatial-temporal resolutions

The paper is organized as follows. Section 2 summarizes recent studies related to DL models for travel demand prediction, sparse data modeling, and uncertainty quantification. Section 3 defines the research question and develops the model. Section 4 introduces the dataset used for the case study, the evaluation metrics, and the experimental results. Section 5 concludes the paper and discuss future research.

2 LITERATURE REVIEW

2.1 Travel demand prediction with spatial-temporal deep learning

Travel demand prediction is a fundamental task in transportation applications. Based on the spatial division of the area of interest into zones, travel demand prediction is usually associated with the prediction of flow from the origin zone to the destination zone pairs, in the form of O-D matrices [12]. The main challenges for O-D matrix prediction relate to time series prediction and spatial correlation detection. Recently these challenges have been addressed using spatial-temporal deep learning techniques. Convolution Neural Network (CNN) based techniques are applied to extract the spatial patterns because O-D matrices are usually modelled on urban grids [22, 41, 43]. Noticeably, the CNN and its variants discover spatial patterns in Euclidean space [39]. The O-D matrices naturally possess the graph structure where origin or destination regions are usually regarded as the nodes and the O-D pairs are the edges. Recent work also applied graph neural networks (GNNs) on travel demand estimation to capture non-Euclidean correlations [4, 40]. On the other hand, Ke et al. [12] regarded O-D pairs, instead of origin and destination regions, as nodes and applied multi-graph convolutional network to predict ride-sourcing demand. Other studies used deep learning to predict the individual travel behavior by focusing on interpretation and architectural design [32, 34]. The power of deep learning arises from its representation learning capacity and its new statistical foundation [33]. However, both CNN and GNN-based approaches treat the O-D matrix entries as continuous and only focus on coarse temporal resolutions, like 60 minutes. Very few studies discussed the challenge to analyze the sparse O-D matrices with high spatial-temporal resolution.

Sparse travel demand data are different from the missing data. Zeros in sparse data mean no trips, but the sensors work properly. Missing data, on the other hand, are unobservable, potentially due to sensor malfunction. Wang et al. [35] applied the spatial-temporal Hankelization and tensor factorization to estimate traffic states using only 17% of the matrix entries. However, tensor factorization is a transductive method, which needs recalculation when new data come in. It is not suitable for short-term prediction. To differentiate our contributions from previous work, we will emphasize our model's spatial-temporal interpretability and uncertainty quantification in travel demand prediction.

2.2 Uncertainty of sparse travel demand prediction

Besides the average value, it is widely acknowledged that models should also predict uncertainty [26, 44]. The deterministic models that dominate the majority of the research implicitly assume homoskedasticity (i.e. common variance), which significantly simplifies the variance structure [2, 6]. However, it is also critical to capture uncertainty of sparse travel demand using deep learning. Rodrigues and Pereira [27] explored uncertainty quantification by training a CNN-LSTM model to fit both the mean and the quantiles. They showed the power of combining the spatial and temporal embedding to predict various distributions, but they did not investigate the sparse scenario. In non-DL models, Jang [8] showed that travel demand follows a zero-inflated negative binomial distribution. Rojas et al. [28] also noted that using a zero-inflation model is promising for modeling intermittent travel demand. However, recent demand prediction papers sidestepped this challenge by choosing a low resolution, such as 60 minutes [11, 12]. It is challenging to consider the higher resolution, such as 5min, because a deterministic DL model is no longer appropriate. Hence, it could be a viable alternative to combine the zero-inflated distribution and the deterministic DL to analyze the sparse O-D matrices.

It is also critical to design metrics to evaluate the quality of uncertainty quantification. Khosravi et al. [14] proposed mean prediction interval width and prediction interval coverage probability to quantify data uncertainty. Sankararaman and Mahadevan [29] introduced a likelihood-based metric, which is also an effective indicator to measure the alignment of the predicted and ground truth data distributions. Based on the existing research gap, our paper seek to design a model that formulates the sparse travel demand prediction problem with proper spatial-temporal pattern recognition and tight prediction interval bounds.

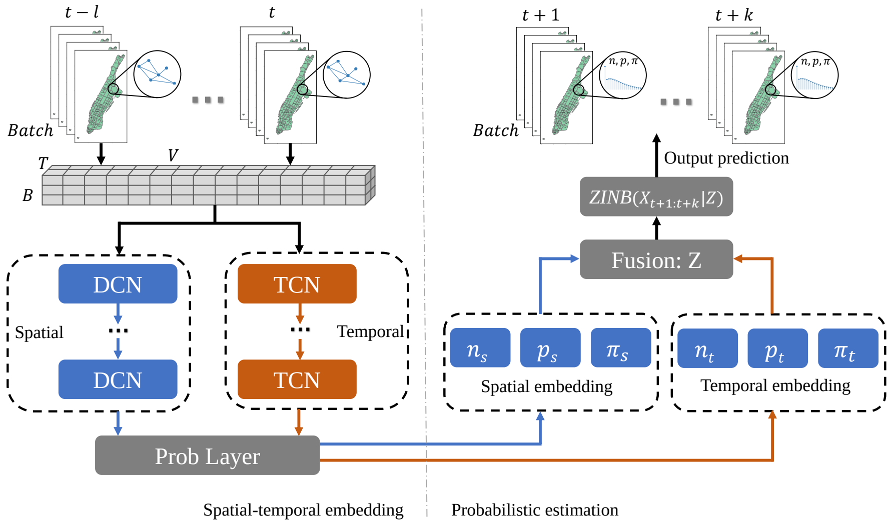


Figure 1: Framework of STZINB-GNN model.

3 METHODOLOGY

3.1 Problem Description

Our model predicts the future expected travel demand and confidence interval of each O-D pair with k time windows ahead, using m origins and u destinations along with the travel demand in time periods (windows) of length T minutes. It is a sequence-to-sequence prediction task. Different from the previous work that treated the locations of origins or destinations as vertices, we build the O-D graph $\mathcal{G} = (V, E, A)$ where V represents the O-D pair set, E denotes the edge set, and $A \in \mathbb{R}^{|V|\times|V|}$ is the adjacency matrix describing the relationship between O-D pairs [12]. It is clear that $|V| = m \times u$, and the O-D graph is fully connected. Let x_{it} denote the trips of the i^{th} O-D pair in the t^{th} time window, where $i \in V$, $x_{it} \in \mathbb{N}$. Then $X_t \in \mathbb{N}^{|V|\times T}$ denote the demand for all O-D pairs in the t^{th} time window, with x_{it} as its entry. Our goal is to leverage historical records $X_{1:t}$ as the data inputs to predict the distribution of $X_{t:t+k}$ (i.e. the demand for the next k time windows), thus analyzing the expectation and confidence intervals of the future demand.

3.2 Zero-Inflated Negative Binomial (ZINB) Distribution

We assume that the inputs follow the ZINB distribution [9, 28]. A random variable that follows NB distribution has a probability mass function f_{NB} as:

$$f_{NB}(x_k; n, p) \equiv Pr(X = x_k) = \begin{pmatrix} x_k + n - 1 \\ n - 1 \end{pmatrix} (1-p)^{x_k} p^n. \quad (1)$$

where n and p are the shape parameters that determine the number of successes and the probability of a single failure respectively. However, the real-world data often have many observations with zeros and overdispersion [18]. The exploded number of zeros exacerbates the parameter learning of the NB distribution. A new parameter π is therefore introduced to learn the inflation of zeros, leading to the ZINB distribution. Formally, its probability mass function can be described as:

$$f_{ZINB}(x_k; \pi, n, p) = \begin{cases} \pi + (1-\pi) f_{NB}(0; n, p) & \text{if } x_k = 0 \\ (1-\pi) f_{NB}(x_k; n, p) & \text{if } x_k > 0 \end{cases}. \quad (2)$$

Two steps are needed to generate the distributions of the data: they are either zeros with probability π or non zeros with probability $1-\pi$ following the NB distribution. The parameters π, n, p in the probability distributions are parameterized by spatial-temporal GNNs.

3.3 STZINB-GNN

We introduce STZINB-GNN, a generalizable deep learning architecture, to capture spatial-temporal correlations $Z = \mathcal{M}(X_{1:t}, A)$ under the ZINB assumption for each O-D matrix entry and predict the future k time windows ahead with $f_{ZINB}(X_{t+1:t+k}|Z)$. $\mathcal{M}$ is the proposed model framework that takes the historical demand $X_{1:t} \in \mathbb{Z}^{|V|\times tT}$ and adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ as inputs, and learns the parameter embedding $Z \in \mathbb{R}^{|V|\times k\times 3}$ of the future demand $X_{t+1:t+k} \in \mathbb{R}^{|V|\times k}$ with:

$$\begin{aligned} & f_{ZINB}(X_{t+1:t+k}|n_{t+1:t+k}, p_{t+1:t+k}, \pi_{t+1:t+k}) \\ & = f_{ZINB}(X_{t+1:t+k}|\mathcal{M}(X_{1:t}, A)) = f_{ZINB}(X_{t+1:t+k}|Z). \end{aligned} \quad (3)$$

The overall architecture of the STZINB-GNN is shown in Figure 1. We convert a batch of size B of input O-D-T tensor at j^{th} time window $X_{j:j+B} \in \mathbb{Z}^{|V|\times BT}$ into a tensor $\mathcal{X}_j \in \mathbb{Z}^{|V|\times B\times T}$ as the model input. We use Diffusion Graph Convolution Networks (DGCNs) to capture the spatial adjacency of O-D pairs, and use Temporal Convolutional Networks (TCNs) for temporal correlation. The outputs of the DGCNs and TCNs, including the spatial embedding n_s, p_s, π_s and temporal embedding n_t, p_t, π_t, are used to parameterize the ZINB distribution. These spatial and temporal embeddings contain the independent estimation of the ZINB parameters of their spatial and temporal locality. We then fuse the n_s, p_s, π_s and n_t, p_t, π_t into Z using the Hadamard product, which can be replaced by other non-linear operations like a fully-connected layer. By fusing the spatial and temporal embeddings of the distribution parameters, we obtain the final ZINB parameter set Z that fulfills

$$f_{ZINB}(X_{t+1:t+k}|n_{t+1:t+k}, p_{t+1:t+k}, \pi_{t+1:t+k}) = f_{ZINB}(X_{t+1:t+k}|Z). \tag{4}$$

Notice that the fused Z in Figure 1 can be interpreted as the parameter set of the future demand distribution. This procedure is similar to the variational autoencoders (VAEs) concept, where we learn the latent shape variables that formulate the distributions [7, 15, 16]. The theory of VAEs is used to introduce deep learning techniques into the statistical domain of variational inference, which provides more powerful representation of latent variables. Our encoding component uses the spatial-temporal embedding architecture with an additional sparsity parameter and the decoding part is related to the probabilistic estimation of future demand.

To address the problem that zeros will give infinite values for the KL-divergence based variational lower bound, we directly use the negative likelihood as our loss function to better fit the distribution into the data. Let y be the ground-truth values corresponding to one of the predicted matrix entries with parameters n, p, π from Z. The log likelihood of ZINB is composed of the $y = 0$ and $y > 0$ parts, and can be approximated as [23, 24]:

$$LL_y = \begin{cases} \log \pi + \log{(1-\pi)} p^n & \text{when} \quad y = 0 \\ \log 1 - \pi + \log \Gamma(n+y) - \log \Gamma(y+1) & \\ \quad - \log \Gamma(n) + n \log p + y \log(1-\pi) & \text{when} \quad y > 0 \end{cases}, \tag{5}$$

where π, n, p are also selected and calculated according to the index of $y = 0$ or $y > 0$, and Γ is the Gamma function. The final negative log likelihood loss function is given by:

$$NLL_{STZINB} = -LL_{y=0} - LL_{y>0}. \tag{6}$$

Note that our model can also be generalized to other distributions by modifying the probability layer into other distributions and using Z to represent the related shape parameter sets. For example, if we use Gaussian distribution as our model, we can parameterize the probability layer using the spatial and temporal embedding of the mean and variance, thus quantifying the data uncertainty that follows the Gaussian distribution. Using the flexibility of the probability layer, we design a variety of benchmark models to compare to the ZINB model. Our scripts can be found in Github[1].

3.4 Adjacency Matrix for O-D Pairs

This study uses O-D pairs as vertices, different from the previous GNN approaches that use regions as vertices [4, 5, 30, 41]. Therefore, we need to model spatial correlations of O-D pairs and construct the adjacency matrix in a different way. Intuitively, the O-D pairs with similar origins or destinations are also close in a graph representation, because passengers are likely to transfer between adjoining O-D pairs rather than remote ones. Inspired by the work of Ke et al. [12], we formulate our adjacency matrix as:

$$\begin{aligned} A^O_{i,j} &= haversine(lng^O_i, lat^O_i, lng^O_j, lat^O_j)^{-1}, \forall i, j \in V \\ A^D_{i,j} &= haversine(lng^D_i, lat^D_i, lng^D_j, lat^D_j)^{-1}, \forall i, j \in V \\ A_{i,j} &= \sqrt{\frac{1}{2}((A^O_{i,j})^2 + (A^D_{i,j})^2)}, \end{aligned} \tag{7}$$

where $lng^O_i, lat^O_i, lng^O_j, lat^O_j$ are the longitudes and latitudes of the origins of O-D pair i, j, and $lng^D_i, lat^D_i, lng^D_j, lat^D_j$ are for the destinations similarly. The function $haversine(\cdot)$ takes the longitudes and latitudes of two geographical points and calculate their distance on Earth. The basic idea is to leverage the O-D pairs' geographical adjacency by averaging the origin and destination similarity. It is clear that the order of i, j does not affect the output of the *haversine* function, which means A is symmetric. The final adjacency matrix A is the quadratic mean of A^O and A^D, where the distances between origins or destinations have the same influence in the adjacency matrix. Future studies can assign different weights to the origins or destinations in constructing the adjacency matrix, or even combine with demographic graphs to enrich the information of A. Since this paper focuses on uncertainty quantification and interpretability of the model, we use a simple construction of A to prevent any distraction.

3.5 Diffusion Graph Convolution Network

In order to capture the stochastic nature of flow dynamics among O-D pairs, we model the spatial correlations as a diffusion process [1, 20]. Introducing the diffusion process facilitates the learning of the spatial dependency from one O-D pair to another. The process is characterized by a random walk on the given graph with a probability $\alpha \in [0, 1]$ and a forward transition matrix $\tilde{W_f} = A/rowsum(A)$ [37]. After sufficiently large time steps, the Markovian property of the diffusion process guarantees it to converge to a stationary distribution $\mathcal{P} \in \mathbb{R}^{|V| \times |V|}$. Each row of $\mathcal{P}$ stands for the probability of diffusion from that node. The stationary distribution can be calculated in closed form [31]:

$$\mathcal{P} = \sum_{k=0}^{\infty} \alpha (1-\alpha)^k (D_O^{-1} A)^k, \tag{8}$$

where k is the diffusion step, which is usually set to finite number K. Variable k is only reused for demonstration of the diffusion process. We can similarly define the backward diffusion process with backward transition matrix $\tilde{W_b} = A^T/rowsum(A^T)$. Since our adjacency matrix A is symmetric, $\tilde{W_f} = \tilde{W_b}$. The forward and backward diffusion processes model the dynamics of passengers shifting from one O-D pair to another, like the shifting from school-home trip to home-market trip. The building block of DGCN layer can be written as [37]:

$$H_{l+1} = \sigma(\sum_{k=1}^{K} T_k(\tilde{W_f}) H_l \Theta^k_{f,l} + T_k(\tilde{W_b}) H_l \Theta^k_{b,l}), \tag{9}$$

where H_l represents the l^{th} hidden layer; the Chebyshev polynomial $T_k(X) = 2XT_{k-1}(X) - T_{k-2}(X)$ is used to approximate the convolution operation in DGCN, with boundary conditions $T_0(X) = I$ and $T_1(X) = X$ (note that the Chebyshev polynomial is used to approximate the diffusion $\mathcal{P}$ instead of using the closed form); learned parameters of the l^{th} layer $\Theta^k_{f,l}$ and $\Theta^k_{b,l}$ are added to control how each node transforms the received information; σ is the activation function (e.g. ReLU, Linear). In our model we stack 3 DGCN layers to better capture the O-D dynamics.

[1] https://github.com/ZhuangDingyi/STZINB

3.6 Temporal Convolutional Network

As Wu et al. [38] point out, the advantages of TCNs compared with recurrent neural networks (RNNs) include: 1) TCNs can use the sequences with varying length as inputs, which is more adaptive to different time resolutions and scales; 2) TCNs have a lightweight architecture and fast training [3]. The general idea of TCNs is to apply a shared gated 1D convolution with width w_l in the l^{th} layer in order to pass the information from w_l neighbors at the current time point. Each TCN layer H_l receives the signals from the previous layer H_{l-1} and is updated using [19]:

$$H_l = f(\Gamma_l * H_{l-1} + b), \tag{10}$$

where Γ_l is the convolution filter for the corresponding layer, $*$ is the shared convolution operation, and b stands for the bias. If previous hidden layer follows $H_{l-1} \in \mathbb{R}^{B\times|V|\times w_{l-1}}$, then the convolution filter is $\Gamma_l \in \mathbb{R}^{w_l \times w_{l-1}}$ so that $H_l \in \mathbb{R}^{B\times|V|\times w_l}$. If there is no padding in each TCN layer, it is clear that $w_l < w_{l-1}$. TCN is also a type of sequence-to-sequence model, which can directly output prediction for future target windows in a row. Furthermore, the TCN receptive field is flexible, which can be controlled by the number and kernels of Γ_l. It is useful in our following discussion about scaling our model with different temporal resolutions. We also stack 3 TCN layers.

4 NUMERICAL EXPERIMENTS

4.1 Data

In this session, we assess the model performance using two real-world datasets from Chicago Data Portal (CDP) [2] and Smart Location Database (SLD) [3]. The **CDP dataset** contains the trip records of Transportation Network Providers (ride-sharing companies) in the Chicago area. The city of Chicago is divided into 77 zones and the trip requests with pick-up and drop-off zone are recorded every 15min. We use 4-month observations from September 1st, 2019 to December 30th, 2019. The dataset is divided into various spatial resolutions to facilitate the discussion of model performance. We randomly select 10 × 10 O-D pairs with the same time period as spatially sparse data sample; The **SLD dataset** is also used in the work of Ke et al. [12]. Specifically, we select the For-Hire Vehicle (FHV) trip records in the Manhattan area (divided into 67 zones by administrative zip codes) from January 2018 to April 2018. This SLD data have similar features as the CDP dataset, including timestamps, and pick-up and drop-off zone IDs. As the dataset includes the timestamps of the individual trip information, we vary the temporal resolution (5min, 15min, and 60min intervals) to test our model performance. In addition to the full O-D matrix (67 × 67), we also sample 10 × 10 small O-D pair samples to align with the CDP dataset. We use 60% of the data for training, 10% for validation, and the last 30% for testing. Table 1 summarizes the data scenarios used in the analysis.

Name	Resolution	# of O-D pairs	Data size	Zero rate
CDP_SAMP10	15min	10 × 10	(100, 11521)	81%
SLD_SAMP10	15min	10 × 10	(100, 11520)	54%
SLD_5min	5min	67 × 67	(4489, 34560)	88%
SLD_15min	15min	67 × 67	(4489, 11520)	70%
SLD_60min	60min	67 × 67	(4489, 2880)	50%

Table 1: Data division summary.

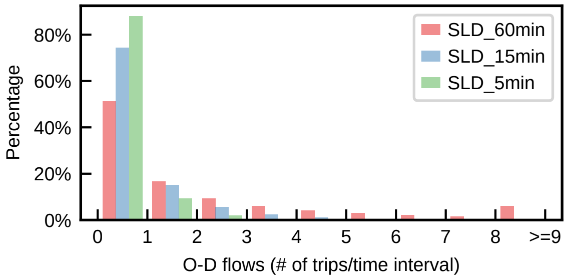


Figure 2: Distribution of travel demand in the SLD dataset per 5, 15, and 60min intervals.

The zero rate increases from 50% to 88% as the temporal resolution for SLD dataset increases from 60 minutes to 5 minutes. As shown in Figure 2, the distribution of O-D flows is quite skewed because most of the O-D flows are concentrated around small values. The O-D flows at a 60min resolution have several observations larger than 8, but at 5min resolution, the lowest value is 3 trips within the 5-minute interval. Such sparse (and small value) O-D matrix is a challenge for most of the deterministic deep learning models.

4.2 Evaluation Metrics

We use the metrics for both the point estimate statistics and distributional characteristics to evaluate and compare the performance of the various models. The prediction accuracy of the expected median value (i.e. point estimate accuracy) is evaluated using the Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{k|V|}\sum_{i=1}^{k|V|} |x_i - \hat{x}_i|, \tag{11}$$

where $\hat{x}_i$ and x_i are the predicted and ground-truth values of i^{th} data point respectively. To evaluate the estimated uncertainty, we use Mean Prediction Interval Width (MPIW) on the 10%-90% confidence interval [14]:

$$\text{MPIW} = \frac{1}{k|V|}\sum_{i=1}^{k|V|} (U_i - L_i). \tag{12}$$

where L_i and U_i correspond to the lower and upper bound of the confidence interval for observation i. The definition of MPIW can be extended to other quantiles of the data. Tighter (smaller) prediction intervals are more desirable. Apart from MPIW, we also use the Kullback-Leibler Divergence (KL-Divergence) to assess how close the model output distribution is to the test set distribution [17]. We define the KL-Divergence as:

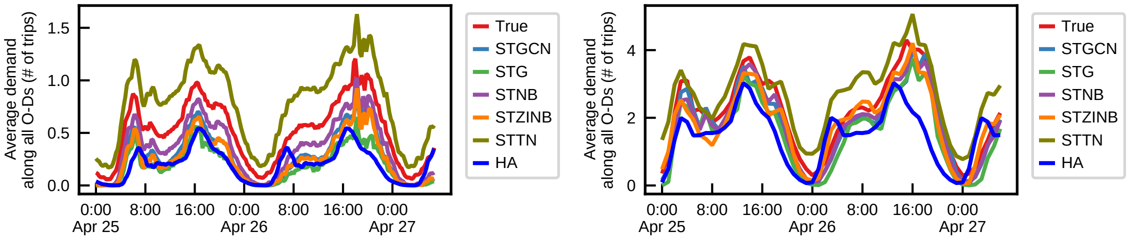
$$\text{KL-Divergence} = \frac{1}{k|V|}\sum_{i=1}^{k|V|} (\hat{x}_i \log \frac{\hat{x}_i + \epsilon}{x_i + \epsilon}), \tag{13}$$

[2] https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p

[3] https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Table 2: Model comparison under different metrics. X/Y values correspond to the mean/median values of the distribution.

Data scenario	Metrics	STZINB-GNN	STNB-GNN	STG-GNN	STTN-GNN	HA	STGCN
CDP_SAMP10	MAE	0.368/**0.366**	0.382/0.379	0.409/0.409	0.432/0.606	0.522	0.395
	MPIW	**1.018**	1.020	2.407	2.089	/	/
	KL-Divergence	**0.291**/0.424	0.342/0.478	0.435/0.435	1.058/0.928	1.377	0.897
	True-zero rate	0.796/0.788	0.796/0.788	0.790/0.790	0.758/0.764	0.759	**0.800**
	F1-Score	**0.848**/0.846	**0.848**/0.841	0.818/0.818	0.842/0.846	0.809	0.840
SLD_SAMP10	MAE	0.663/0.666	0.627/**0.616**	0.630/0.630	0.695/0.665	0.697	0.630
	MPIW	**1.310**	3.628	2.604	1.931	/	/
	KL-Divergence	0.518/**0.507**	0.980/1.662	1.022/1.022	3.578/3.052	0.978	0.768
	True-zero rate	0.499/**0.502**	0.465/0.418	0.461/0.461	0.308/0.336	0.364	0.478
	F1-Score	**0.567**/0.566	0.556/0.552	0.555/0.555	0.477/0.500	0.456	0.563
SLD_5min	MAE	0.149/0.150	0.147/**0.144**	0.155/0.155	0.155/0.155	0.149	0.159
	MPIW	**0.094**	1.249	0.922	0.741	/	/
	KL-Divergence	0.015/0.014	0.042/0.145	**0.001/0.001**	0.001/0.001	0.060	0.056
	True-zero rate	**0.879/0.879**	0.875/0.866	0.877/0.877	0.877/0.877	0.874	0.874
	F1-Score	**0.882/0.882**	0.880/0.878	0.879/0.879	0.879/0.879	0.876	0.879
SLD_15min	MAE	0.370/0.372	0.351/**0.342**	0.356/0.356	0.365/0.356	0.418	0.373
	MPIW	**0.603**	2.283	1.353	1.215	/	/
	KL-Divergence	0.167/**0.156**	0.357/0.704	0.353/0.353	1.445/1.211	0.445	0.395
	True-zero rate	0.725/**0.727**	0.710/0.684	0.709/0.709	0.632/0.648	0.703	0.708
	F1-Score	**0.751**/0.750	0.746/0.745	0.750/0.750	0.716/0.726	0.744	0.750
SLD_60min	MAE	1.040/1.067	0.958/**0.947**	1.199/1.199	1.275/1.254	1.014	0.997
	MPIW	3.277	5.753	2.282	**1.592**	/	/
	KL-Divergence	0.982/1.270	**0.926**/0.963	2.176/2.176	4.120/3.734	2.421	1.114
	True-zero rate	0.458/**0.476**	0.443/0.425	0.390/0.390	0.288/0.308	0.447	0.438
	F1-Score	0.536/0.537	**0.538**/0.534	0.479/0.479	0.407/0.423	0.490	**0.538**



(a) SLD_15min

(b) SLD_60min

Figure 3: Prediction results in the SLD_15min and SLD_60min scenarios. Results are averaged over the spatial dimension (i.e. all O-D pairs)

Since many x_i and $\hat{x}_i$ values are likely to be zeros, a small perturbation $\epsilon = 10^{-5}$ is used to avoid numerical issue because of division by 0. Since the KL-Divergence measures the difference between two distributions, smaller values are desirable.

To compare the model performance on the discrete O-D matrix entries, we use the true-zero rate and F1-score measurements. The true-zero rate quantifies how well the model replicates the sparsity in the ground-truth data. The F1-score, on the other hand, measures the accuracy of discrete predictions. Even though the F1-score is designed for classification models, we can still consider the discrete values as multiple labels and define the precision and recall accordingly [25]. Larger true-zero rate and F1-score values indicate better model performance.

4.3 Model Comparison

In order to explore the advantages of the STZINB-GNN, we compare the STZINB-GNN results against three other models: (1) **Historical Average (HA)** serves as the statistic baseline. It is calculated by averaging the demand in the same daily time intervals (e.g. 8:00AM-8:15AM) from the historical data to predict the one-step ahead future demand. (2) **Spatial-Temporal Graph Convolutional Networks (STGCN)**[4] is the state-of-the-art deep learning model for traffic prediction [42]. STGCN also uses graph convolution for spatial embedding and temporal convolution for temporal embedding. However, it fails to quantify the demand uncertainty and only produces point estimates; (3) **Models with probabilistic assumptions** in the spatial-temporal embedded probabilistic layer: negative binomial (STNB-GNN), Gaussian (STG-GNN), and truncated normal (STTN-GNN), as shown in Figure 1. These three

[4] https://github.com/FelixOpolka/STGCN-PyTorch

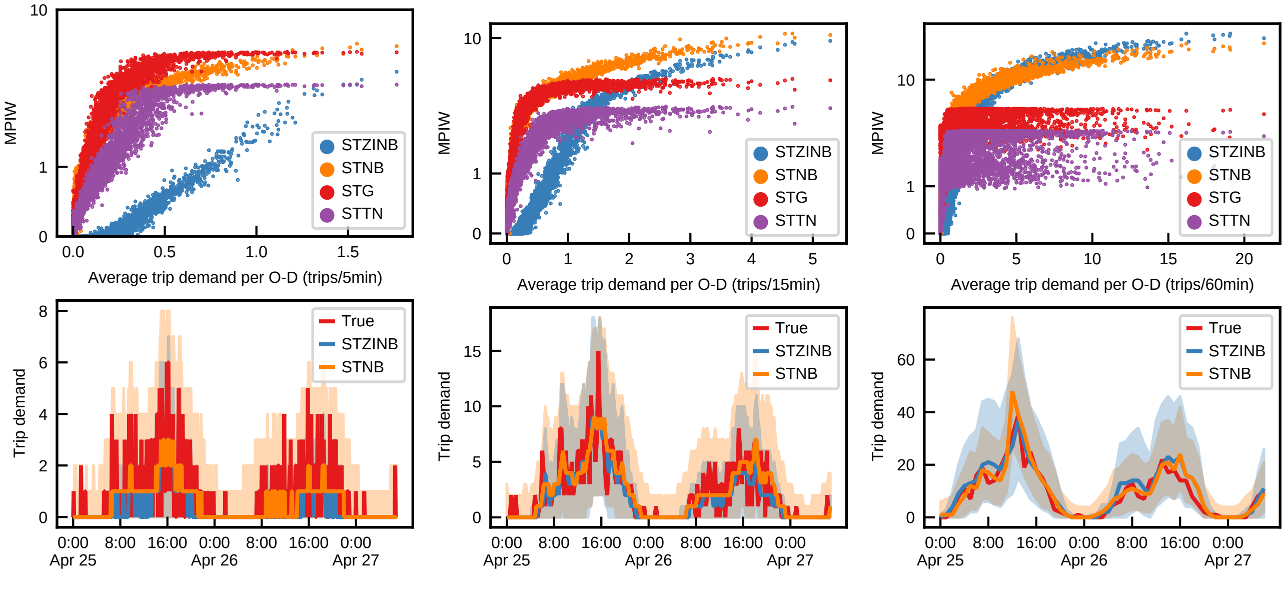
distributions have two parameters, less than the three parameters of STZINB-GNN. The other components and parameters are the same across these models.

STZINB-GNN outperforms other models when the O-D matrix resolution is high but performs worse when the resolution becomes coarser. Table 2 highlights the fields in bold to indicate the best performance for each data set. For fair comparison, the outputs from the continuous-output models, including STG-GNN, STTN-GNN, HA, and STGCN, were rounded to the closest integer to compare to the STNB-GNN and STZINB-GNN. As shown in Table 2, when the spatial and temporal resolutions are low, like the CDP_SAMP10 and SLD_SAMP10 cases, STZINB-GNN performs similarly to STNB-GNN. In the sparsest scenario, SLD_5min (with 88% zeros), STG-GNN, STTN-GNN, and STGCN fail to effectively capture the skewed data distribution, leading to low true-zero rates and F1-scores. The STZINB-GNN, on the other hand, successfully learns the sparsity of the data. Its prediction has a very small MPIW, nearly 10 times smaller than the other models. When the temporal resolution decreases, the STNB-GNN, STTN-GNN, and STGCN start to outperform STZINB-GNN. In the SLD_60min case, STNB-GNN fits the data better than the STGCN. The STZINB-GNN, on the other hand, only predicts the true-zero entries better. No single model dominates others under all resolution levels [36].

The negative binomial distribution in the probability layer can effectively capture the discrete values of the travel demand. Figure 3 illustrates the predicted and ground-truth average travel demand in the SLD (New York) data for two consecutive days, with Figure 3(a) for the SLD_15min and Figure 3(b) for the SLD_60min. With finer resolutions, the STNB-GNN and STZINB-GNN models are more likely to accurately predict the average travel demand. When the resolution decreases, like in the 60min case, all the deep learning models deliver similar performance. This 60-min resolution is commonly used in the majority of the deep learning studies, but our results demonstrate the importance of discrete probabilistic assumptions when the temporal unit is shorter than 60 minutes.

4.4 Uncertainty Quantification

The STZINB-GNN provides superior performance in quantifying uncertainty, because it uses much narrower MPIW for a fixed probability coverage, particularly when the temporal resolution is high. Figure 4 visualizes the scatter plots for the MPIW and the groundtruth travel demand of the 4489 O-D pairs at three temporal resolutions. Figure 4(a) demonstrates that the STZINB-GNN leads to significantly smaller MPIW than the other models in the granular 5min resolution case. This is because the average travel demand for each O-D pair is less than two in the SLD_5min case and most of them are zeros. The introduction of the sparsity parameter π in STZINB-GNN effectively captures the zeros. The STG-GNN and STTN-GNN are not able to capture the skewness of the data distribution, leading to large MPIWs. However, when the temporal resolution decreases, the zero-inflation with the sparsity parameter π becomes less important. Figures 4(b) and 4(c) show that the SLD_15min and SLD_60min cases have larger average travel demand per 15 and 60 minutes time interval. In these two cases, the



(a) SLD_5min case

(b) SLD_15min case

(c) SLD_60min case

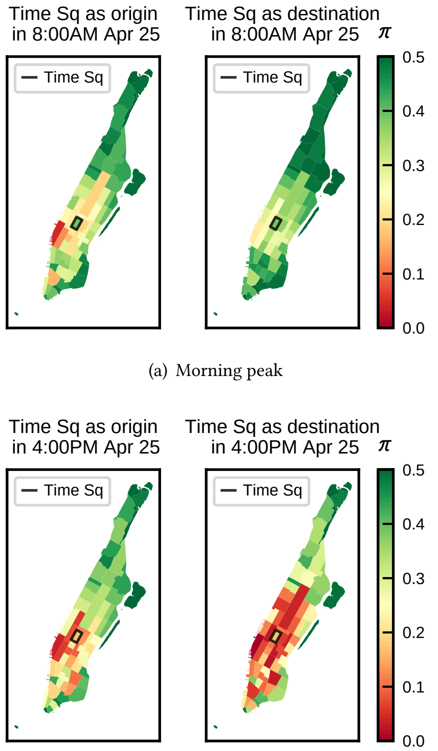
Figure 4: Model prediction uncertainty at various resolutions (SLD_5min to SLD_60min). The top row shows the MPIW needed to predict the demand of each O-D pair at different resolutions. The bottom row represents the prediction with uncertainty for a specific O-D pair, from Midtown Center to Union Square, in April 25 and 26, 2018. This O-D pair is selected because it has the largest trip flow.

STZINB-GNN has smaller MPIW only when the average travel demand is small. The STG-GNN and STTN-GNN outperform STZINB-GNN when the average demand is large. However, their output prediction distributions are not stable, because the same travel demand value corresponds to different MPIW with a large variance.

The bottom row of Figure 4 compares the expectation and MPIW between STZINB-GNN and STNB-GNN for a specific O-D pair that has the largest demand flow in the selected time range. In the SLD_5min and SLD_15min cases, the STZINB-GNN provides more compact confidence intervals than the STNB-GNN, even when its point prediction is less accurate. The results are consistent with the results in Table 2 and the MPIW comparison in the top row of Figure 4.

4.5 Interpretation of the Sparsity Parameter π

The sparsity parameter π in the ZINB distribution measures how likely a zone has zero demand. Note that each of our predicted O-D pair has the parameter π, which can capture the inflow and outflow sparsity level of a zone. Since O-D relationship is hard to illustrate in the map, we focus on a specific zone and project the O-D activities when the zone is selected as the origin or destination. Figure 5 shows the O-D activity (i.e. the parameter π) heatmap of Time Square, during the morning and evening peaks of April 25, 2018. It can be found that spatial locality exists, where communities are more likely to commute to their neighbors. Moreover, the temporal patterns also vary. As shown in Figure 5, many visitors explore the neighboring regions in the evening but are very inactive during the morning peak. Therefore, the sparsity parameter π renders the STZINB-GNN highly interpretable. It is very important in the transportation decision-making or the operation manager to use the sparsity parameter π to assign mobility service. Our framework has the potential extended to other prediction tasks that have many zeros and are sensitive to events, like incident precaution or paratransit service for the disabled people.



(a) Morning peak

(b) Evening peak

Figure 5: Morning and evening peaks outflow from and inflow to Time Square. We project the O-D pairs into the map according to Time Square as the origin or the destination zone. Red zones stand for small π values, meaning high possibility to have trips generated there and green regions otherwise.

5 CONCLUSION

In this paper, we propose a generalizable spatial-temporal GNN framework to predict the probabilistic distribution of sparse travel demand and quantify its uncertainty. We introduce the zero-inflated negative binomial distribution with a sparsity parameter π. We use spatial diffusion graph neural networks to capture spatial correlation and temporal convolutional networks to capture temporal dependency. The STZINB-GNN framework embeds the spatial and temporal representation of the distribution parameters respectively and fuses them to obtain the distribution for each spatial-temporal data point. The STZINB-GNN is evaluated using two real-world datasets with different spatial and temporal resolutions. We find that the STZINB-GNN outperforms the baseline models when the data are represented in high resolutions but performs worse when the resolution becomes coarser. This is also reflected in the prediction interval, where the STZINB-GNN has tight confidence intervals. Since the parameter π has physical interpretation, our model could help transportation decision makers to efficiently assign mobility services to zero or non-zero demand areas. This framework has the potential to be extended to other prediction tasks that use highly sparse data points, such as anomaly detection and accident prediction.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Vehicle Technology Program Award Number DE-EE0009211.

REFERENCES

[1] James Atwood and Don Towsley. 2016. Diffusion-convolutional neural networks. In *Advances in neural information processing systems*. 1993–2001.

[2] Chuan Ding, Jinxiao Duan, Yanru Zhang, Xinkai Wu, and Guizhen Yu. 2018. Using an ARIMA-GARCH Modeling Approach to Improve Subway Short-Term Ridership Forecasting Accounting for Dynamic Volatility. *IEEE Transactions on Intelligent Transportation Systems* 19, 4 (2018), 1054–1064. https://doi.org/10.1109/TITS.2017.2711046

[3] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proc. ICML*. 1243–1252.

[4] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. 2019. Spatiotemporal Multi-Graph Convolution Network for Ride-Hailing Demand Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019), 3656–3663. https://doi.org/10.1609/aaai.v33i01.33013656

[5] Xu Geng, Xiyu Wu, Lingyu Zhang, Qiang Yang, Yan Liu, and Jieping Ye. 2019. Multi-modal graph interaction for multi-graph convolution network in urban spatiotemporal forecasting. *arXiv preprint arXiv:1905.11395* (2019).

[6] Jianhua Guo, Wei Huang, and Billy M. Williams. 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies* 43 (2014), 50–64. https://doi.org/10.1016/j.trc.2014.02.006

[7] William L Hamilton. 2020. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning* 14, 3 (2020), 1–159.

[8] Tae Youn Jang. 2005. Count data models for trip generation. *Journal of Transportation Engineering* 131, 6 (2005), 444–450.

[9] Mengmeng Jiang and Hang Zhang. 2018. Sparse estimation in high-dimensional zero-inflated Poisson regression model. In *Journal of Physics: Conference Series*, Vol. 1053. IOP Publishing, 012128.

[10] Wenhua Jiang, Zhenliang Ma, and Haris N Koutsopoulos. 2022. Deep learning for short-term origin–destination passenger flow prediction under partial observability in urban railway systems. *Neural Computing and Applications* (2022), 1–18.

[11] Jintao Ke, Siyuan Feng, Zheng Zhu, Hai Yang, and Jieping Ye. 2020. Joint predictions of multi-modal ride-hailing demands: a deep multi-task multigraph learning-based approach. (11 2020). http://arxiv.org/abs/2011.05602

[12] Jintao Ke, Xiaoran Qin, Hai Yang, Zhengfei Zheng, Zheng Zhu, and Jieping Ye. 2021. Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network. *Transportation Research Part C: Emerging Technologies* 122 (2021), 102858.

[13] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen. 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 (2017), 591–608.

[14] Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. 2011. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks* 22, 3 (2011), 337–346. https://doi.org/10.1109/TNN.2010.2096824

[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[16] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[17] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[18] Diane Lambert. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1 (1992), 1–14.

[19] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.

[20] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[21] Fuqiang Liu, Jiawei Wang, Jingbo Tian, Dingyi Zhuang, Luis Miranda-Moreno, and Lijun Sun. 2022. A Universal Framework of Spatiotemporal Bias Block for Long-Term Traffic Forecasting. *IEEE Transactions on Intelligent Transportation Systems* (2022).

[22] Lingbo Liu, Zhilin Qiu, Guanbin Li, Qing Wang, Wanli Ouyang, and Liang Lin. 2019. Contextualized spatial–temporal network for taxi origin-destination demand prediction. *IEEE Transactions on Intelligent Transportation Systems* 20, 10 (2019), 3875–3887.

[23] Mihoko Minami, Cleridy E Lennert-Cody, Wei Gao, and M Román-Verdesoto. 2007. Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. *Fisheries Research* 84, 2 (2007), 210–221.

[24] Abbas Moghimbeigi, Mohammed Reza Eshraghian, Kazem Mohammad, and Brian Mcardle. 2008. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics* 35, 10 (2008), 1193–1202.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[26] Soora Rasouli and Harry Timmermans. 2012. Uncertainty in travel demand forecasting models: Literature review and research agenda. , 55–73 pages. https://doi.org/10.3328/TL.2012.04.01.55-73

[27] Filipe Rodrigues and Francisco C. Pereira. 2020. Beyond Expectation: Deep Joint Mean and Quantile Regression for Spatiotemporal Problems. *IEEE Transactions on Neural Networks and Learning Systems* (2020), 1–13. https://doi.org/10.1109/tnnls.2020.2966745

[28] Fernando Rojas, Peter Wanke, Giuliani Coluccio, Juan Vega-Vargas, and Gonzalo F Huerta-Canepa. 2020. Managing slow-moving item: a zero-inflated truncated normal approach for modeling demand. *PeerJ Computer Science* 6 (2020), e298.

[29] S Sankararaman and S Mahadevan. 2013. Distribution type uncertainty due to sparse and imprecise data. *Mechanical Systems and Signal Processing* 37, 1-2 (2013), 182–198.

[30] Junkai Sun, Junbo Zhang, Qiaofei Li, Xiuwen Yi, Yuxuan Liang, and Yu Zheng. 2020. Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[31] Shang-Hua Teng. 2016. Scalable algorithms for data and network analysis. *Foundations and Trends® in Theoretical Computer Science* 12, 1–2 (2016), 1–274.

[32] Shenhao Wang, Baichuan Mo, and Jinhua Zhao. 2020. Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions. *Transportation Research Part C: Emerging Technologies* 112 (2020), 234–251.

[33] Shenhao Wang, Qingyi Wang, Nate Bailey, and Jinhua Zhao. 2021. Deep neural networks for choice analysis: A statistical learning theory perspective. *Transportation Research Part B: Methodological* 148 (2021), 60–81.

[34] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. 2020. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies* 118 (2020), 102701.

[35] Xudong Wang, Yuankai Wu, Dingyi Zhuang, and Lijun Sun. 2021. Low-Rank Hankel Tensor Completion for Traffic Speed Estimation. *arXiv preprint arXiv:2105.11335* (2021).

[36] David H Wolpert and William G Macready. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation* 1, 1 (1997), 67–82.

[37] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. 2021. Inductive Graph Neural Networks for Spatiotemporal Kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4478–4485.

[38] Yuankai Wu, Dingyi Zhuang, Mengying Lei, Aurelie Labbe, and Lijun Sun. 2021. Spatial Aggregation and Temporal Convolution Networks for Real-time Kriging. *arXiv preprint arXiv:2109.12144* (2021).

[39] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.

[40] Xi Xiong, Kaan Ozbay, Li Jin, and Chen Feng. 2020. Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter. *Transportation Research Record* 2674, 8 (2020), 491–503.

[41] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[42] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks : A Deep Learning Framework For Traffic Forecasting. *ProceedLearningings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* (2018), 3634–3640. https://aaafoundation.org/american-driving-survey-2014-2015/

[43] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*.

[44] Yong Zhao and Kara Maria Kockelman. 2002. The propagation of uncertainty through travel demand models: An exploratory analysis. *Annals of Regional Science* 36, 1 (2002), 145–163. https://doi.org/10.1007/s001680200072

[45] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. 2021. Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transportation Research Part C: Emerging Technologies* 132 (2021), 103410.

[46] Dingyi Zhuang, Siyu Hao, Der-Horng Lee, and Jian Gang Jin. 2020. From compound word to metropolitan station: Semantic similarity analysis using smart card data. *Transportation Research Part C: Emerging Technologies* 114 (2020), 322–337.