Communications in Transportation Research 3 (2023) 100099

Contents lists available at ScienceDirect

Communications in Transportation Research

journal homepage: www.journals.elsevier.com/communications-in-transportation-research

Full Length Article

AGNP: Network-wide short-term probabilistic traffic speed prediction and imputation

Meng Xu [a], Yining Di [a], Hongxing Ding [a], Zheng Zhu [b,c,d,*], Xiqun Chen [b,c,d], Hai Yang [a,e]

[a] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[b] Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, 310058, China
[c] Zhejiang Provincial Engineering Research Center for Intelligent Transportation, Hangzhou, 310058, China
[d] Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, Hangzhou, 310058, China
[e] Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511458, China

ARTICLE INFO

Keywords:
Prediction and imputation
Neural processes
Congestion prediction
Graph neural networks

ABSTRACT

The data-driven Intelligent Transportation System (ITS) provides great support to travel decisions and system management but inevitably encounters the issue of data missing in monitoring systems. Hence, network-wide traffic state prediction and imputation is critical to recognizing the system level state of a transportation network. Abundant research works have adopted various approaches for traffic prediction and imputation. However, previous methods ignore the reliability analysis of the predicted/imputed traffic information. Thus, this study originally proposes an attentive graph neural process (AGNP) method for network-level short-term traffic speed prediction and imputation, simultaneously considering reliability. Firstly, the Gaussian process (GP) is used to model the observed traffic speed state. Such a stochastic process is further learned by the proposed AGNP method, which is utilized for inferring the congestion state on the remaining unobserved road segments. Data from a transportation network in Anhui Province, China, is used to conduct three experiments with increasing missing data ratio for model testing. Based on comparisons against other machine learning models, the results show that the proposed AGNP model can impute traffic networks and predict traffic speed with high-level performance. With the probabilistic confidence provided by the AGNP, reliability analysis is conducted both numerically and visually to show that the predicted distributions are beneficial to guide traffic control strategies and travel plans.

1. Introduction

With advances in the Intelligent Transportation System (ITS), traffic information can be quickly collected, analyzed, and transmitted, supporting travel decision and system management (Ganin et al., 2019; Ran et al., 2012; Sumalee and Ho, 2018). The ITS improves traffic mobility by incorporating various technological devices into vehicles and road infrastructure (Deb et al., 2017; El Hamdani et al., 2020). However, the data-driven ITS suffers from the problem of missing data (Kaur et al., 2022; Li et al., 2020). Due to a tight budget or privacy concerns, the monitoring system generally has a limited coverage that leads to missing data issues on some road segments; the occasional system breakdown can also deteriorate the problem, e.g., hardware breakdown, network connectivity issues, power supply shortage, and extreme weather (Bae et al., 2018; Qu et al., 2009). Patterns of traffic data missing can be divided into three categories with respect to randomness. Random missing is the case where missing values are independent of each other; temporally/spatially correlated missing is the case in which missing values are temporally/spatially correlated; when missing values have both temporal and spatial correlation, it is referred to as block missing (Chen et al., 2018; Li et al., 2018; Yang et al., 2021a). Solving these three categories of missing data and their combinations are the main research targets of traffic data imputation.

Given the prevalence of missing data, traffic state analysis, prediction, and control become quite challenging. A complete dataset of historical traffic observations contributes more to the performance of prediction, compared to a low-quality dataset with abundant missing data. In other words, imputation of miss values is critical to enhance data quality for

* Corresponding author. Institute of Intelligent Transportation Systems, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, 310058, China.
E-mail address: zhuzheng89@zju.edu.cn (Z. Zhu).

https://doi.org/10.1016/j.commtr.2023.100099
Received 22 March 2023; Received in revised form 4 June 2023; Accepted 4 June 2023
Available online 25 July 2023

predictions. Hence, it is of great benefit to handle data missing issues when designing predictive models. Some studies addressed these two problems separately and hence developed independent data-driven models for them (Che et al., 2018), while some studies proposed their well-designed data-driven models to make predictions based on incomplete datasets with missing values (Cui et al., 2020). These studies typically adopt deep learning models that map inputs to outputs with deterministic imputed/predicted values. However, traffic state cannot be fully captured by the deterministic predictive values. That is, transportation metrics are stochastic in real time; traffic states could experience sudden changes due to episodic emergencies or unexpected events, such as traffic accidents, vehicle sudden breakdowns, extreme weather (snow, ice, and fog) or temporary maintenance of road facilities. Thereby, previous models could fail to give latent information about transportation systems. Contrarily, probabilistic models can provide reliability information on its predictions; the information can be further summarized as information on accident-prone road sections or bottleneck of traffic networks based on historical knowledge. This process is referred to as reliability analysis in this work.

A typical example of probabilistic models is the Gaussian process (GP) model. GP models offer closed-form expressions of outputs' probability distributions, parameterized by the Gaussian mean and variance. In this way, GP models the prediction problem as stochastic processes, and thereby can give reliability information of the predicted variable. In ITS, GPs are frequently utilized to predict traffic speed and simulate traffic trends (Liu et al., 2013; Rodrigues et al., 2018a; Rodrigues and Pereira, 2018b). However, GPs suffer from computational complexity issues and dramatically constrain the scale of studied scenarios and the size of utilized datasets. Considering the scale of transportation networks and massive amount of traffic data, GPs are greatly weakened to model the transportation systems. Recently, some researchers proposed an attentive neural process (ANP) model (Kim et al., 2019) which combines the strengths of GPs and neural networks. It shows linear computational complexity and is competent to handling large-size datasets. However, as a novel method, ANPs have not received much attention in ITS fields, and is not capable of handling graph-structured data, i.e., traffic networks.

In this work, we incorporate ANPs with graph neural networks, and propose a novel attentive graph neural process (AGNP) method for short-term network-level traffic speed prediction and imputation. After summarizing spatiotemporal patterns of the traffic data, we convert the transportation network into graph structures, and model the graph-wise traffic speed as a stochastic process. The AGNP will be utilized to approximate the process, i.e., predicting future graph-structured traffic speeds via historical datasets with partial missing values. The AGNP model will output a Gaussian distribution, parameterized by the mean and variance, for each predicted lane of roadway segments. This allows for reliability analysis and confidence in the predictions. To test its performance, we use real-world traffic data from Xuancheng, China and conducted experiments with increasing levels of missing data. The results show that the AGNP model had strong predictive abilities and the predicted distributions are useful for en route guidance and path planning, as demonstrated both numerically and visually through reliability analysis.

The remainder of this study is organized as follows. **Section 2** gives a comprehensive literature review on traffic state imputation and prediction. **Section 3** introduces preliminary knowledge for building the AGNP model, including the research problem, the spatiotemporal features of traffic data, and the ANP model. In **Section 4**, the AGNP model is formulated which is based on the ANP models and tailored specially for prediction and imputation problems. **Section 5** introduces the data and settings, shows the predictive results, and conduct reliability analysis based on a real-world case study. **Section 6** concludes the study.

2. Literature review

Given temporal or/and spatial data missing patterns, traffic data imputation is critical to enhancing the operational performance of ITSs (Ni et al., 2005; Tang et al., 2021). Temporal-pattern-based imputation relies on temporal information and regards traffic data as time series. For example, Smith et al. (2002) adopted the autoregressive integrated moving average (ARIMA) model, assuming that the imputed data are linearly dependent on the prevailing data. Chan et al. (2021) imputed the missing information by a weighted combination of short-term and long-term historical data, and incorporated spatiotemporal traffic information into a deep learning neural network for traffic congestion prediction. Since the imputation techniques from the temporal view only focus on correlations among timestamps but fail to use the correlation among sensors, spatial information is further incorporated into these models to improve performance and accuracy (Chen et al., 2019; Zhang et al., 2021a; Zhu et al., 2016). Li et al. (2013) found that spatiotemporal dependencies can reduce imputing errors in both probabilistic principal component analysis (PPCA) and kernel probabilistic principal component analysis (KPPCA) methods. Moreover, the KPPCA-based imputation method performs better when spatiotemporal dependence is nonlinear.

Various studies incorporate data imputation and short-term traffic prediction into an integrated task (Tian et al., 2018; Yang et al., 2021b; Zhang et al., 2021b). Cui et al. (2020) designed a graph Markov network for traffic flow prediction with missing values. The traffic state transition process is regarded as a Markov process in a topological network. Yang et al. (2021b) developed a spatiotemporal data imputation and traffic prediction framework; the long short-term memory (LSTM) network is utilized to capture temporal dependencies, and the graph Laplacian (GL) captures spatial dependence. Zhang et al. (2021b) proposed a graph convolutional bidirectional recurrent neural network to simultaneously address data imputation and traffic prediction problems; by passing spatiotemporal messages and topology information in the road network, missing traffic data can be estimated. Liu et al. (2018) developed an attention convolutional neural network (CNN) for short-term traffic prediction and extracted spatiotemporal traffic features at 15-min intervals to conduct predicting tasks. They showed that their proposed model has advantages in predicting, and meanwhile can reduce the adverse impact of missing data. Many researchers also introduced abundant external factors to these advanced models (Jia et al., 2017; Ke et al., 2017; Liu and Chen, 2017), e.g., events, holiday, and rainfall. These works aim at leveraging intrinsic dependencies between these factors and traffic conditions to better capture complex nonlinear dependencies and make informed predictions. These external factors serve as attributes or features, and are encoded either in one-hot format (Zhang et al., 2017) or entity embedding (Liu et al., 2019).

Since a graph structure can effectively model the topology and spatial correlations among road links, advanced spatiotemporal relationships can be regressed by graph neural networks (GNNs) (de Medrano and Aznarte, 2020; Salamanis et al., 2016). Hence, the graph-based deep learning method has been widely used for traffic prediction and imputation (Luan et al., 2022; Peng et al., 2021; Ta et al., 2022; Xu et al., 2023; Zhang et al., 2019; Zhu et al., 2022a). Spatiotemporal GNN techniques rely on the graph convolution operation to handle graph-structured and temporal traffic data, like simple temporal convolution (Lee and Rhee, 2022; Zhu et al., 2022a), gated temporal convolution layers (Ta et al., 2022; Yu et al., 2017), LSTM (Peng et al., 2021), and attention-based encoder/decoder networks (Huang et al., 2022; Ye et al., 2021). Cui et al. (2019) proposed a traffic graph convolutional LSTM-based neural network (TGC-LSTM) model to learn spatiotemporal features and conduct predictions. Ye et al. (2021) combined attention encoder networks (AEN) with GNN to extract spatial and temporal features from traffic flow data. Huang et al. (2022) incorporated the inhomogeneous Poisson process into a GNN to predict traffic demand. Tang and Zeng (2022) applied graph attention network to capture spatial dependencies and gated recurrent unit to extract the temporal evolution of traffic states. However, previous methods focus on traffic prediction and data imputation but ignore the reliability analysis of the predicted/imputed values.

As aforementioned, reliability analysis of predicted/imputed traffic states is vital for the decision-making in ITSs because of flexibilities for unexpected events. However, the existing data-driven techniques often fail to provide reliability in traffic prediction/imputation accuracy (Rodrigues and Pereira, 2018b; Yuan et al., 2021). The GP can be a desirable alternative to address the issue of reliability in traffic prediction/imputation. The GP can provide closed-form expressions of the posterior distribution where the mean and covariance functions characterize the stochastic process of the observations (Rasmussen and Williams, 2006; Zhu et al., 2022b). Zhu et al. (2022c) built a grid GP model to simultaneously predict pickups, returns, and idle bikes for a large-scale bike-sharing system in Manhattan, New York. Spana and Du (2022) exploited GPs to extract potential relationships from training data and further incorporated them into a coordinated routing mechanism. The key in GPs is to identify the mean and covariance (i.e., kernel) function, which can be computationally expensive. Thus, Garnelo et al. (2018) proposed a neural process (NP) that maintains the properties of GPs to model distributions over functions and estimate uncertainties over predictions, and has linear computation complexity. Kim et al. (2019) further improved the training efficiency of NP by introducing the attention mechanism and proposed the attentive neural process (ANP) model.

Although ANP models show high-level performance in computer vision fields, it is not suitable for graph-structure datasets. To the best of our knowledge, this work is among the first to propose NP architecture for traffic speed imputation/prediction in graph-structured transportation networks. Moreover, it also provides a promising method to give reliability analysis for traffic networks and obtain insightful reference for real-world operational decision.

3. Problem descriptions and preliminary knowledge

In this section, we introduce problem descriptions and preliminary knowledge for building the AGNP model. **Section 3.1** defines the research problem for this work, and **Section 3.2** introduces the attentive neural process model.

3.1. Research problem

Generally, a road segment (or a road) is defined as a directed connection from one intersection to another. In other words, each road segment has a start intersection and an ending intersection. As depicted in Fig. 1, from the starting intersection to the ending intersection, each road segment can be abstracted to one mixed lane connected with several channelized lanes (up to three depending on the turning movement directions of the ending intersection). In the mixed lane, vehicles can switch lanes freely; at the channelized lane, vehicles must join one of the channelized lanes to complete a specified movement, i.e., a left turn, a straight-ahead movement, or a right turn. Given sufficient detecting equipment at the specific road segment, the speeds of vehicles in the mixed lane and the channelized lanes are recorded separately.

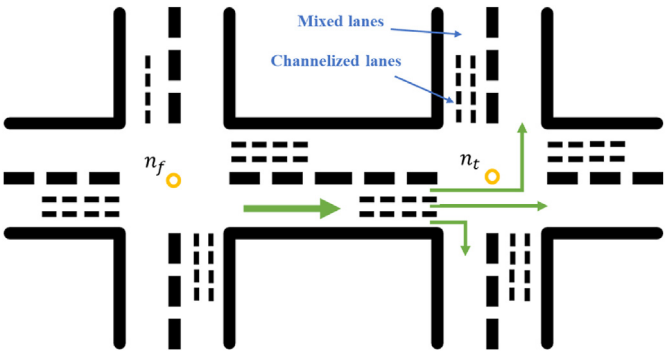

Fig. 1. Road illustration.

Based on the aforementioned definitions of road segments, urban traffic speed networks for any arbitrary time period t can be modeled as a constant directed graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V}$ denotes the node set and $\mathscr{E}$ denotes the link set. From then on, nodes denote (mixed and channelized) lanes of road segments and links denote the connectivity of nodes (road lanes).

Specifically, nodes represent lanes, either mixed lanes or channelized lanes. According to Fig. 1, each node (lane) is denoted according to the start intersection of its road segment, the ending intersection of its road segment, and its movement direction; thereby, the expression of the node set V is given as Eq. (1):

$$\mathscr{V} = \{v(n_f, n_t, d)\}, \forall n_f, n_t \in N, v \in \mathscr{V}, d \in \{0, 1, 2, 3\} \tag{1}$$

where $v(\bullet)$ is a node, defined by the start intersection n_f, the ending intersection n_t, and the target direction d (mixed 0, left 1, straight ahead 2, and right 3). N is the intersection set. Moreover, link set E is given by

$$\mathscr{E} = \{e_{ij}\}, \forall i, \quad j \in \mathscr{V} \tag{2}$$

Furthermore, node features x_i for node v_i include start intersection n_{f_i}, ending intersection n_{t_i}, heading intersection n_{h_i}, direction d_i, lane width a_i, free-flow speed s_i, length of the channelized lane $l_{i,c}$, length of the mixed lane $l_{i,m}$, turning rate w_i, green time ratio g_i (or equivalent terms), daytime t_d, day of week t_w, precipitation r, and holiday label h. Specifically, holiday label indicates whether the current timestamp is public holiday or not.

$$x_i = \left[n_{f_i}, n_{t_i}, n_{h_i}, d_i, a_i, s_{ffi_i}, l_{i,c}, l_{i,m}, w_i, g_i, t_d, t_w, r, h\right] \tag{3}$$

These feature values are related to the node (i.e., road segment) itself and are constant values that do not change over training procedure. Moreover, we utilize $\boldsymbol{X}$ to denote the feature set, i.e., $\boldsymbol{X} = \{x_i | \forall i \in \mathscr{V}\}$.

Furthermore, at each time period t, a partially-observed adjacency matrix A_t is constructed to denote concurrent traffic speeds obtained by monitoring systems. Due to network topology, some part of the adjacency matrix is discontinuous. That is, one mixed lane node has traffic flow to its corresponding channelized lane nodes; one channelized lane node has traffic flow to the next mixed lane node that it turns into; a pair of nodes in other cases are discontinuous (referred to as discontinuity). Moreover, because of missing data issues, some continuous part of the adjacency matrix is still unknown (referred to as unknowability). The discontinuity or unknowability between two nodes are expressed as zero, respectively, and other values represent valid observed traffic flow. Thereby, the formulation of $A_t^{(i,j)}$ is summarized as Eq. (4):

$$A_t^{(i,j)} = \begin{cases} a_t^{(i,j)}, \text{for observed traffic flows} \\ 0, \text{else} \end{cases} \tag{4}$$

where $a_t^{(i,j)}$ denotes the traffic flow speed from node i to node j at timestamp t.

Apart from A_t, we define $\overline{A}_t$ to represent the fully-observed (or complete) adjacency matrix which only contains discontinuity. A simplified illustration of A_t and $\overline{A}_t$ is depicted in Fig. 2. The detailed

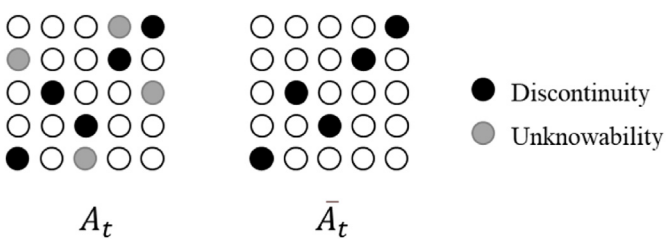

Fig. 2. Illustration of A_t and $\overline{A}_t$.

formulation of $\overline{A}_t^{(i,j)}$ is given as Eq. (5):

$$\overline{A}_t^{(i,j)} = \begin{cases} a_t^{(i,j)}, \text{for existing traffic flows} \\ 0, \text{for unobserved or unaccessible traffic flows} \end{cases} \quad (5)$$

Referring to Fig. 3, the network prediction and imputation problem is to utilize previous m partially-observed adjacency matrix $\{A_{t-T-m}, A_{t-T-m+1}, \ldots, A_{t-T}\}$ in the previous day and the directed network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to predict the next day complete adjacency matrix $\overline{A}_t$, including known and unknown nodes. Note that T represents total time intervals in one day; thus it is to utilize information of the previous day to predict current traffic states. The AGNP model which is proposed for this prediction and imputation problem will be introduced in detail in **Section 4**.

3.2. Attentive neural processes

As an essential extension of NPs (Garnelo et al., 2018), ANPs together with NPs are crucial models to capture stochastic processes with neural networks. They both display linear computational efficiency which compensates for high time complexities of stochastic processes due to large matrix calculations. Specifically, ANPs equip NPs with well-known attention mechanisms and exhibit better goodness-of-fit. Moreover, trained ANPs can provide reliability analysis on predicted outputs, which can be further served as a basis for establishing real-world operational plans.

ANPs define an infinite family of conditional distributions to fit stochastic processes, i.e., distributions over functions, mapping an input X to a random variable Y. In other words, ANPs give a predicted Gaussian distribution for the output Y, parameterized by the mean and variance values. Specifically, it conditions on an arbitrary size of observed context points $(X_C, Y_C) = \{(x_i, y_i)\}_{i \in C}$ to capture an arbitrary size of target points $(X_T, Y_T) = \{(x_i, y_i)\}_{i \in T}$. Note that the notations, X_C, x_i, X_T here only denote the inputs of ANP and thus are different from the node feature x_i or the feature set $X = \{x_i | \forall i\}$ in **Section 3.1**. The structural details of ANPs are depicted in Fig. 4. ANPs exploit an encoder-decoder framework. To encode the inputs, ANPs design a deterministic path to generate an explicit representation r_i out of each inputted context point (x_i, y_i), and meanwhile a latent path to capture the latent distribution of inputs. In the deterministic path, each context point (x_i, y_i) will be stacked together and fed into a multi-head self-attention layer to get an r_i; then in the subsequent cross-attention module, each target point serves as the query Q whereas context points act as keys K and their corresponding representations as values V. In this way, the deterministic path outputs an explicit target-specific embeddings r_* for each target (x_*, y_*). Similarly, in the latent path, context point pairs go through multi-head self-attention layer to get s_i, and aggregate by taking the average to get s_C and capture latent distributions of context points. The distribution is a parameterized Gaussian distribution for z. In the decoder, target points along with z and r_* will go through a multi-layer perceptron to provide mean and variance for modeling the likelihood $p(Y_T | X_T, r_*, z)$. The detailed conditional distribution $p(Y_T | X_T, X_C, Y_C)$ is given as Eq. (6):

$$p(Y_T | X_T, X_C, Y_C) = \int p(Y_T | X_T, r_*, z) q(z | s_C) dz \quad (6)$$

where $q(\bullet)$ is the posterior distribution.

Via reparameterization tricks (Kingma and Welling, 2014), the parameters of ANPs are learned by maximizing the evidence lower bound which is showed as Eq. (7):

$$\log p(Y_T | X_T, X_C, Y_C) \geq \mathbb{E}_{q(z|s_T)}[\log p(Y_T | X_T, r_*, z)] - D_{KL}(q(z|s_T) \| q(z|s_C)) \quad (7)$$

where $D_{KL}(q_1(\bullet) \| q_2(\bullet))$ denotes the KL divergence between the distribution q_1 and the distribution q_2.

It is worth noting that ANPs cannot handle graph-structured inputs as stated in **Section 3.1**. Thereby, it will dismiss spatial patterns hidden in the transportation network. In **Section 4**, we introduce the proposed AGNP model which utilizes a graph encoder and an ANP encoder to jointly extract spatiotemporal patterns.

4. Attentive graph neural processes

In this section, we comprehensively introduce the proposed AGNP model. We display the overall model framework in **Section 4.1**, and **Section 4.2** describes each module in detail. The overall learning procedure is introduced in **Section 4.3**.

4.1. Overview of the model framework

The proposed attentive graph neural processes (AGNPs), which exploit an encoder-decoder structure, are depicted in Fig. 5.

Generally, the model is to utilize preceding m partially-observed adjacency matrices $\{A_{t-T-m}, A_{t-T-m+1}, \ldots, A_{t-T}\}$ in the previous day and the directed network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to predict the next day complete adjacency matrix $\overline{A}_t$. In other words, the training task is to utilize subgraphs in the previous day to predict the current complete graph. Thereby, we introduce a mask operation to generate sub-subgraphs A_t' for subgraphs A_t (Fig. 6(a)), and exploit A_t' to predict A_t so that the model learns how to infer complete graph from subgraphs. As depicted in Fig. 6(b), the masked graphs A_t' can be viewed as subgraphs of A_t. In other words, the masked graphs A_t' and partially-observed graphs A_t are counterparts for A_t and complete graphs $\overline{A}_t$.

Then the masked graphs $\{A'_{t-T-m}, \ldots, A'_{t-T}\}$ and partially-observed graphs $\{A_{t-T-m}, \ldots, A_{t-T}\}$ all go through a graph encoder module separately; it helps to aggregate node features with their neighbors' features according to the concurrent adjacency matrix, and convert node features into latent representations. The representations generated by masked graphs will be stacked together and fed into an attentive neural process encoder module; the encoder finally outputs an explicit representation and a latent distribution of previous m graphs. In the decoder module, the representations generated by partially-observed graphs $\{A_{t-T-m}, \ldots, A_{t-T}\}$ together with the explicit representation and distributions go through a decoder with multi-layer perceptron (MLP) and inner product modules to calculate the complete adjacency matrix $\overline{A}_t$ for the complete graph. The detailed training procedure will be introduced in **Section 4.3**.

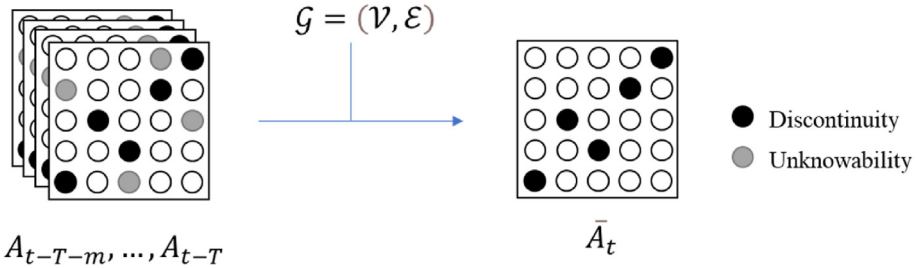

Fig. 3. Illustration of prediction and imputation problem.

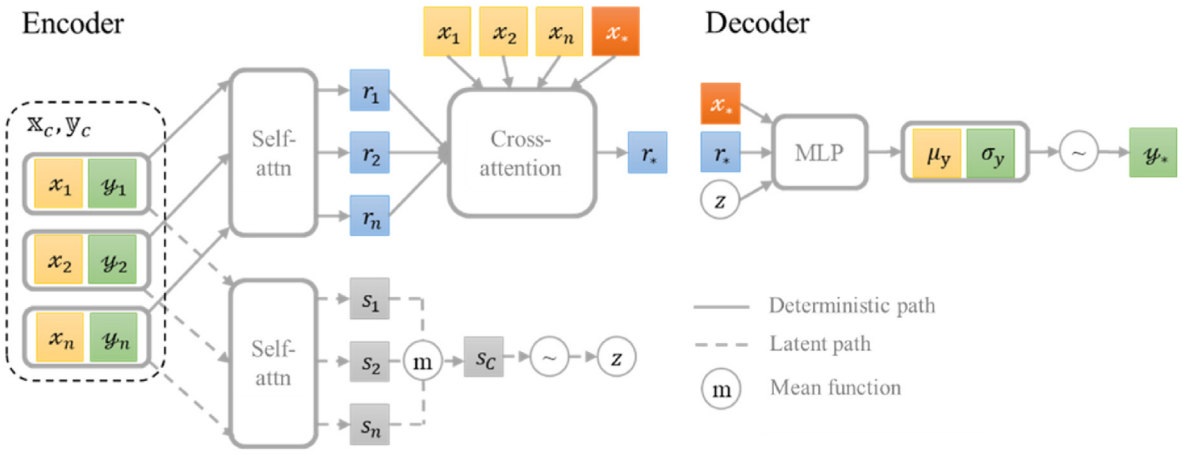

Fig. 4. Architecture of ANPs.

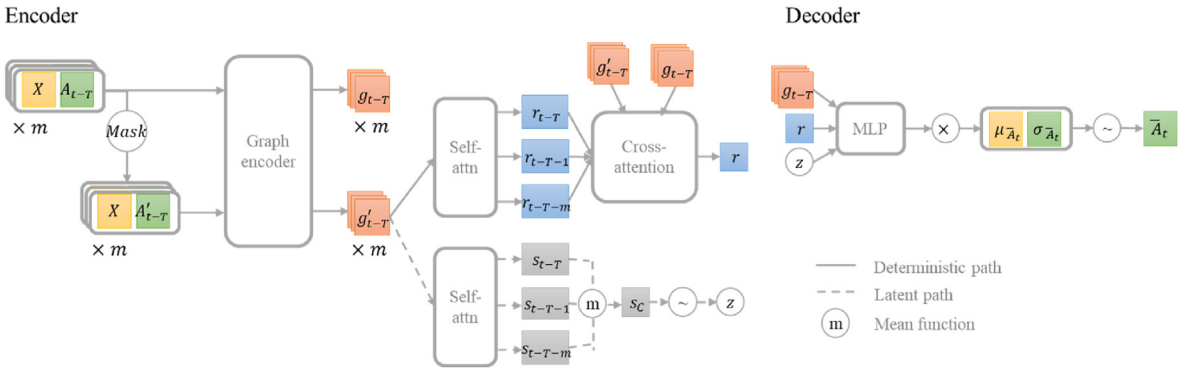

Fig. 5. Architecture of AGNPs.

4.2. Introduction of modules

4.2.1. Mask operation

The AGNP model starts with a mask operation which selects certain links and sets them to be zero in the corresponding adjacency matrix. It is to imitate data missing in the monitoring system. Specifically, some roadway segments are randomly selected to be masked; all links that start with its mixed lane or all related channelized lanes will be set to zero in the concurrent adjacency matrix. This relates to uncovered road situation, so that no information is recorded in the system. Apart from roadway segments, we also randomly select some lanes, either channelized lanes or mixed lanes, to mask. This corresponds to the random data missing, such as occasional monitor failure. Given m graphs, the masked roads and lanes are the same. However, for different training epochs, the model will mask different positions randomly.

4.2.2. Graph encoder

Next in the AGNP framework is a graph encoder module, primarily a combination of MLP and graph convolutional layers. This module aims to extract information from past graph-structural traffic speed states, either masked or unmasked.

For each time interval, the node features are fed into a one-layer MLP to generate a general representation for each node; these representations together with the corresponding adjacency matrix then go through a two-layer graph convolutional network (Kipf and Welling, 2017), followed by another one-layer MLP to further transform the graph information into a one-dimensional representation. Specifically, each graph convolutional layer has the following layer-wise expression:

$$\boldsymbol{H}^{(l+1)} = \mathrm{ReLU}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{8}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}$ denotes the revised adjacency matrix with added self-connections. Note that $\boldsymbol{A}$ here represents a general adjacency matrix that is fed into this layer, not necessarily denoting the real-world speed adjacency matrix; and Eq. (8) applies to all graphs with respect to time index from $t-T-m$ to $t-T$. $\boldsymbol{I}$ represents the identity matrix, $\tilde{\boldsymbol{D}}^{(i,i)} = \sum_j \tilde{\boldsymbol{A}}^{(i,j)}$ denotes the degree matrix of $\tilde{\boldsymbol{A}}$, $\boldsymbol{W}^{(l)}$ is the trainable weight matrix of l-th layer and $\mathrm{ReLU}(\bullet) = \max(0, \bullet)$ is the adopted activation function. $H^{(l)}$ is the matrix fed into the l-th layer and $\boldsymbol{H}^{(0)} = \boldsymbol{X}$.

Thereafter, the m one-dimensional representations for traffic speed states of previous day's m time intervals are stacked together and outputted for following modules in our proposed AGNP framework. Conclusively, this graph encoder module takes advantage of graph structural information, helps each node integrate information from itself and its neighbor nodes, and thus generates more informative representations.

4.2.3. Attentive neural process encoder module

After encoding graphs, the generated representations will be sent into an attentive neural process encoder module. This module is similar to the ANP encoder module introduced in **Section 3.2**, except for the inputs to be the graph embeddings generated in the graph encoder module. Specifically, the representations will go through a deterministic path to obtain an explicit target-specific embedding r, and a latent path to output an intrinsic representation z sampled from inherent distributions s_C, simultaneously.

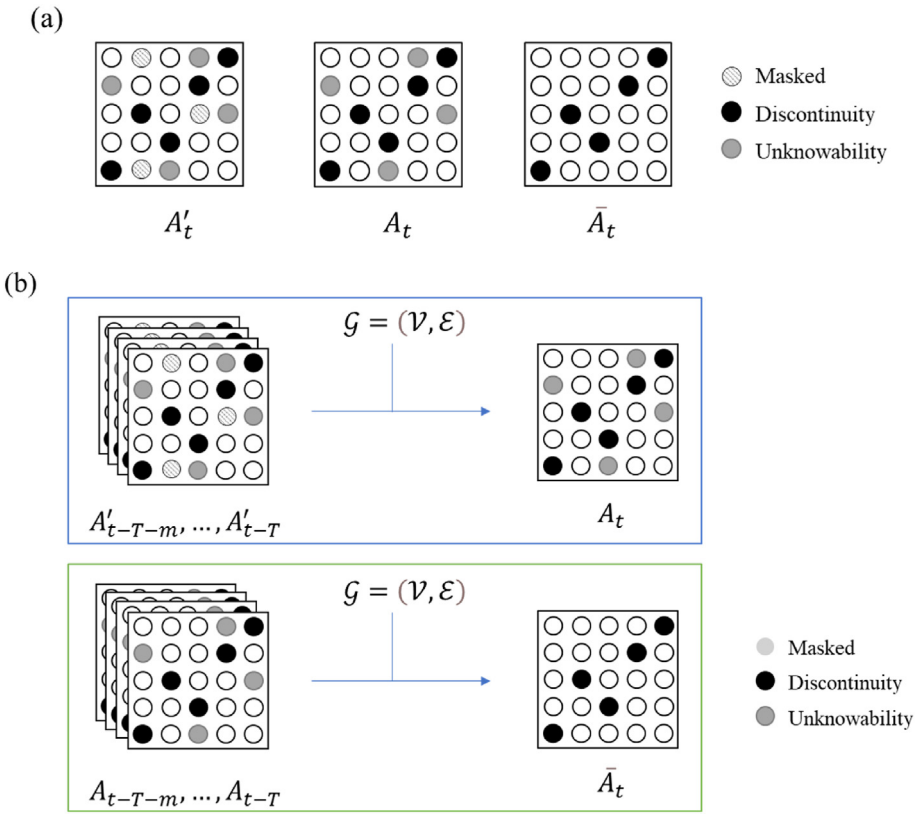

Fig. 6. (a) Illustration of mask operations; (b) illustration of inferring from subgraphs to complete graphs.

After this module, the representation can be further extracted and converted based on interactions between features with the aid of the attention mechanisms in the attentive neural process encoder. The self-attention and cross-attention mechanisms will guide the model to obtain a more informative representation of features that encode latent interactions.

4.2.4. Graph decoder module

In the decoder phase, the latent representations generated by unmasked graphs together with the target-specific embedding r and the intrinsic representation z will be further passed through a two-layer MLP and an inner product operation. The output is the distribution for each position in the complete graph $\bar{A}_t$. Specifically, given real-world connectivity situations, it is pointless to make predictions for nonexistent links. In other words, only part of the adjacency matrix needs predicting. Thereby, after the inner product operation, positions related to nonexistent links will be assigned with zero.

4.3. Overall training procedure

Referring to NPs, the variational lower bound of AGNP can be expressed as Eq. (9):

$$\mathscr{L} = \mathbb{E}_{q(z|\boldsymbol{X}, A_{t-T-m}, \ldots, A_{t-T})}[\log p(\bar{A}_t|\boldsymbol{X}, z)] - D_{KL}\left[q(z|\boldsymbol{X}, A'_{t-T-m}, \ldots, A'_{t-T}) \| q(z|\boldsymbol{X}, A_{t-T-m}, \ldots, A_{t-T})\right] \quad (9)$$

where $p(\bar{A}_t|\boldsymbol{X}, z)$ is the output of the decoder introduced in **Section 4.2**, $q(z|\boldsymbol{X}, A_{t-T-m}, \ldots, A_{t-T})$ is parameterized by estimated mean and variance based on preceding m partially-observed adjacency matrices in the previous day, and $q(z|\boldsymbol{X}, A'_{t-T-m}, \ldots, A'_{t-T})$ is similarly parameterized based on sub-subgraphs. In this way, A_t and A'_t phase can be viewed as the whole graph and subgraph, and thus are counterparts for the whole graph $\bar{A}_t$ and subgraph A_t, respectively. Hence, maximizing the variational lower bound $\mathscr{L}$ is to maximize the conditional log likelihood and meanwhile, force the latent distribution inferred from $\{A'_{t-T-m}, \ldots, A'_{t-T}\}$ to be close to the latent distribution inferred from $\{A_{t-T-m}, \ldots, A_{t-T}\}$. Generally, parameters of AGNPs are learned by maximizing Eq. (9) via the reparameterization trick (Kingma and Welling, 2014).

5. Numerical studies

5.1. Data descriptions and setup

The utilized traffic speed dataset is collected between 6 a.m. and 9 p.m., from April 1st to 27th in 2019, at Xuancheng City, Anhui Province, China. The monitored road network in Xuancheng is equipped with dense cameras and thus generate good-quality datasets. Moreover, since the road network topology is relatively simple and in regular shape, a small amount of missing data at certain camera can be completed by using technical methods, like license plate recognition, through the neighboring cameras. Thereby, a comprehensive traffic flow status dataset can be obtained. Specifically, the dataset contains key attributes of traffic states and geographic information with corresponding timestamps. Through dense checkpoints set up in the road network, this dataset contains 374 valid road segments, each of which relates to a free-flow speed, lengths of the mixed zone, and the channelized zones (constant between different lanes). The road sections were transformed into 1,196 valid lanes, including mixed and channelized lanes. The average speed was aggregated at a 5-min interval for each lane, producing $12 \times 15 \times 27 = 4,860$ valid timestamps. Thereby, there are 180 timestamps per day; that is $T = 180$. The statistics also summarize lane width, the turn rate, and the green time ratio for each channelized lane. In our experiment, based on the processing described in **Section 3.1**, the model is trained and tested based on a single matrix layer with 1,196 nodes, 14 node features, and 1,563 valid links at any given timestamp.

We assume that the dataset is completely-observed, although it may

not record every single road in Xuancheng. In other words, it serves as the ground truth, i.e., complete graphs $\overline{A}_t$, for the proposed AGNP model. In order to simulate a partially-observed monitoring system, removal operations, similar to mask operations, are utilized to generate A_t out of complete graphs $\overline{A}_t$. In training phase, we also generate $A_t^{'}$ out of A_t with mask operations. Specifically, we conduct three experiments with increasing missing data ratio, i.e., 10% missing data, 40% missing data, and 70% missing data. In the 10% missing data experiment, data for 30 roads and 30 other lanes are randomly selected to remove (unknowability), relating to two kinds of masks introduced in **Section 4.2.1**. And in the 40% missing data experiment, 100 roads and 100 other lanes are randomly removed; in the 70% missing data experiment, 200 roads and 200 other lanes are randomly removed. When further generating the masked adjacency matrix $A_t^{'}$ based on the partially-observed datasets, we further randomly mask 15 to 25 roads and 10 to 30 lanes to during training, regardless of the missing data ratio. All the experiments are conducted with 60 min as historical time window, i.e., $m = 12$.

5.2. Predictive results

According to the 5-min aggregation, there are $12 \times 15 = 180$ traffic state graphs per day, and 4,860 graphs in total for 27 days. We take the first 26 days as training data, and the last one day for testing. All models for comparison and corresponding knowledge are introduced in Table 1. Three traditional machine learning models, two advanced deep learning models, one GP method, and ANP model are chosen for comparison with our proposed AGNP model. All experiments are coded with Python 3.7, and trained on a workstation with a single Intel(R) Xeon(R) Gold 5218 R CPU and 64 GB of RAM. The traditional machine learning models, i.e., RF, MLP, and LASSO, are compiled with the python package sklearn (Pedregosa et al., 2011); the Gaussian-process-related model, i.e., SGP is built with package GPflow (Matthews et al., 2017); deep learning models, ANP model, and our proposed AGNP model are coded with PyTorch.

The predictive results of all models are listed in Table 2; the results of NP-related and GP-related methods are outputted based on the corresponding predicted mean values. To measure and evaluate the performance of different methods, mean absolute errors (MAEs) and root mean squared errors (RMSEs) are adopted. Best-performance results are marked with bold font. It is worth noting that, the MAE and RMSE loss are both the average across all data on April 27th. In terms of the AGNP model, the errors are not the average of the adjacency matrices, but valid positions related to existent lanes in the adjacency matrices. Our proposed AGNP model shows high-level predictive performance, and is less effected by missing rate compared to other models. Fig. 7 depicts how predictive performance changes over time with missing rate as 10%. With regards to the high MAE at noon and night, one possible explanation is that traffic congestion during rush hour periods (which often occur around noon and in the early evening) causes higher uncertainty in traffic conditions than other time periods with light congestion and thus leads to higher error. Therefore, it is critical to conduct reliability analysis to provide comprehensive evaluation of traffic states, and help account for the uncertainties and variations in traffic conditions.

Table 1
Baseline models.

Model	Description
RF	Random forest model, and the number of trees is 100 in this study.
MLP	Multi-layer perception model, which is a basic NN containing two hidden layers and 256 neurons in each hidden layer.
LASSO	LASSO model in which the L1 regularization term equals 0.1.
LSTM	Long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997), that contains one LSTM layer and two fully-connected layers. The hidden size of LSTM is 64 and the activation function is tanh.
GRU	The gated recurrent neural (GRU) structure (Chung et al., 2014), designed with one GRU layer and two fully connected layers. The hidden size of GRU is 128 and the activation function is ReLU.
SGP	Sparse GP (Candela and Rasmussen, 2005), which is a probabilistic sparse approximation method for GP regression.
ANP	Attentive neural process (ANP) model introduced in **Section 3.2**.
AGNP	The proposed AGNP model.

Table 2
Predictive results.

Model	Missing 10%		Missing 40%		Missing 70%	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MLP	4.307	6.249	4.474	6.364	4.520	6.437
RF	4.594	6.888	5.019	7.360	5.640	8.165
LASSO	5.574	7.918	5.558	7.925	5.704	7.995
LSTM	4.084	6.407	4.301	6.565	4.359	6.601
GRU	4.176	6.430	4.322	6.534	4.403	6.598
SGP	5.198	7.989	5.679	8.621	5.854	8.827
ANP	4.340	6.742	4.481	7.088	4.594	7.412
AGNP	**4.013**	**6.190**	**4.132**	**6.328**	**4.236**	**6.368**

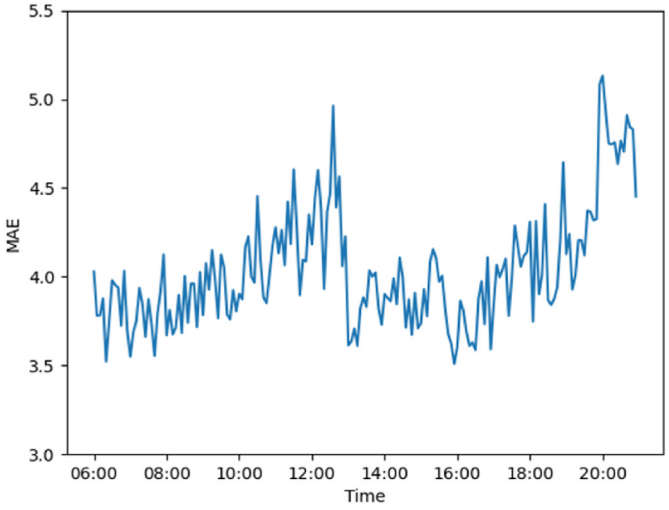

Fig. 7. Temporal MAE results.

To this end, we have showed the predictive power based on mean value in predicted distributions. It is worth noting that although the improvement of our AGNP model is not massive compared to other models, it achieves competitive performance. Apart from predictive results, AGNP has additional benefits beyond its performance in prediction itself. Specifically, the model is designed to conduct reliability analysis and provide statistical knowledge about traffic patterns that can be useful for traffic management and control. It provides insight into the uncertainty of traffic speed predictions and the potential for traffic congestion. The reliability analysis experiment will be introduced in the following subsection, **Section 5.3**.

5.3. Reliability analysis and discussions

In this work, the proposed AGNP model outputs a distribution for each lane in the transportation network, which is parameterized by a mean value and a variance value. Statistically, this distribution can be exploited to explore current traffic states. It is intuitive to use the cumulative distribution function, which is given in Eq. (10), to calculate the possibility of traffic speed of any line under any condition.

$$\mathrm{P}\{y<\theta\}=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\theta} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\mathrm{d}t \tag{10}$$

Specifically, we take as an example the possibility of traffic speeds to be smaller than 20 km/h at 9 a.m., April 27th. Table 3 lists reliability analysis results for specific lanes, including possibility and coefficient of variance. First, the specific channelized lane can be analyzed separately.

Table 3
Reliability analysis results.

Starting Node	Direction	End Node	Possibility (%)	CV
4562-4536-1	Left	4536-4635-0	44.79	0.3630
4662-4651-2	Straight Ahead	4651-4635-0	48.61	0.5813
4536-4635	—	—	53.33	0.3633

The first two rows in Table 3 list analysis in this case; for example, the first row indicates that there is a chance of 44.79% of the specific lane speed being smaller than 20 km/h. Hence, it provides information when designing urban road networks, e.g., road diversion, temporal one-way street, and new lane construction. Second, the channelized and mixed lanes can be viewed together, i.e., a road, and predict the average possibility of its traffic being slower than 20 km/h, or any other speed limit. This case can be viewed in the last row of Table 3. This is beneficial for traffic management, like issuing congestion warnings or designing toll charges. The coefficient of variance (CV), which is given in Eq. (11) can also be summarized from the outputted distributions, indicating the difficulty of prediction. That is, the higher the CV, the more variability (in terms of speed) the road/lane has. Table 3 also lists CV values for corresponding lanes. It illustrates the bottleneck of predictions that need extra focus and attention when monitoring the urban traffic.

$$\mathrm{CV} = \frac{\sigma}{\mu} \tag{11}$$

Furthermore, Fig. 8 visually depicts the heat map of reliability analysis results, including robust traffic speeds and coefficient of variance. Fig. 8(a) illustrates the predicted heat maps of road network states at 9 a.m. on April 27th, i.e., the morning peak, which was randomly selected for this experiment. The heat value ψ is combined with two indicators of road average speed and variance to comprehensively evaluate the real state of road networks.

$$\psi = \mu - \sigma \tag{12}$$

where ψ is a robust representation of the predicted traffic speed. Fig. 8(b) is the heat map of network-wise coefficient of variance. It visually reflects the bottleneck of prediction, which is prone to congestion or other traffic problems, because areas with high-level CV values have high variability of traffic states. The heat map helps to quantify the level of uncertainty in traffic predictions, which can help to prioritize areas where additional data collection or monitoring efforts are needed.

Conclusively, for traffic managers and travelers, reliability analysis and corresponding road network heat maps can numerically quantify and visually represent the state of the road network, and identify potential traffic congestion, thus assisting them to reasonably design traffic control strategies and make travel plans. For example, based on forecasted traffic speeds and confidence, navigation platform may construct an intelligent path planning system that proactively redirects travelers via en route guidance, mitigating congestion before it occurs. That is, reliability analysis can provide various support for traffic control strategy formulation and induced traffic.

6. Conclusions

In urban areas, it is unrealistic for monitoring systems to cover every road due to limited budget issues and privacy concerns. Hence, utilizing previous partially-observed traffic state information to comprehensively predict the current traffic states is of great importance. In other words, network-wise short-term traffic speed imputation and prediction is essential for real-world transportation systems. In this work, we originally propose the AGNP model for network-wise short-term traffic speed prediction and imputation problems. Specifically, with summarized spatiotemporal traffic patterns, we construct graph structures for transportation networks and treat the graph-wise traffic speed state as a stochastic process. The proposed AGNP model is utilized to approximate the progress based on historical datasets with missing values. For each traffic lane, the AGNP model will provide a Gaussian distribution with a mean and a variance value. Thereby, it is capable of conducting reliability analysis and providing confidence in its predictions. Real-world traffic speeds collected in Xuancheng, Anhui, China, are utilized to test the performance of AGNPs. We conduct three experiments with increasing missing data ratio, and the results demonstrate a high-level predictive performance in network-wise traffic speed prediction and imputation. Reliability analysis is analyzed both numerically and visually to show that the predicted distributions are beneficial to guide traffic control strategies and travel plans both numerically and visually.

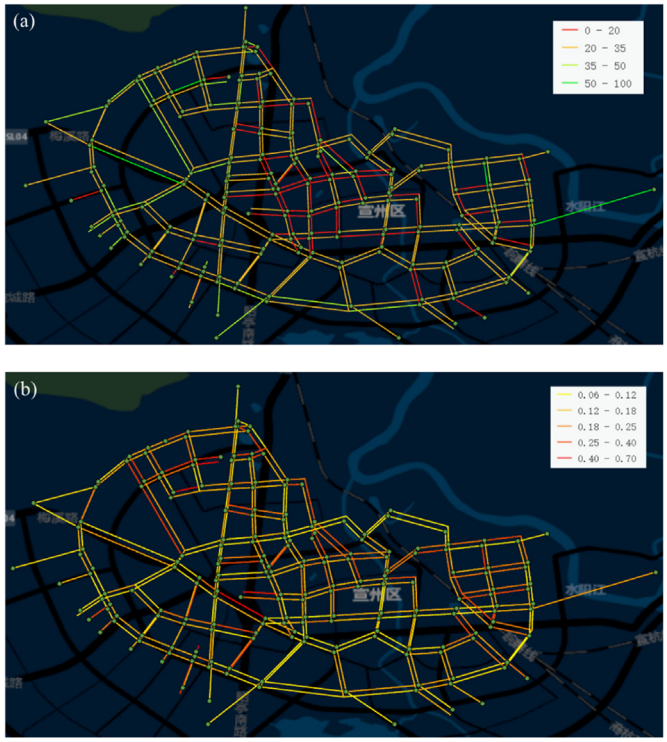

Fig. 8. Heat map of reliability analysis results: (a) heat map of robust traffic speed; (b) heat map of coefficient of variance.

The major contributions of this study include: (1) this study integrally consider prediction and imputation problems, which requires fitting with incomplete datasets; compared to other methods that separately consider the two problems, this model makes them an integral framework, and thus powerful in real-world applications; (2) we incorporate graph convolutional layers into ANP models, and propose the AGNP model, which exhibits powerful predictive performance despite the increasing missing data ratio; (3) the predicted distributions also allow numerical and visual reliability analysis, making it practical for real-world applications, e.g., evaluation of possible traffic congestion and design of traffic control strategies.

There are some potential future works to extend this study. First, this work conducts a one-step prediction. We will consider extending this problem to a scenario with multi-step predictions and reliability analysis (Wang and Wu, 2021; Zhu et al., 2023), to see whether more insightful knowledge could be obtained. Second, with the predicted traffic speed and individual's perception on travel time reliability (Gao et al., 2021; Ortuzar, 2021; Zhu et al., 2021), navigation platforms can consider developing an intelligent path planning system and diverting travelers in advance, i.e., en route guidance, so that congestion can be avoided beforehand. Advanced technologies, e.g., reinforcement learning and control theory, can be integrated for in-depth investigation of this kind of induced traffic. Third, future work can continue to investigate how

predictions together with reliability analysis can be further utilized to guide real-world services and traffic control systems. We will carry out simulation experiments to further explore reliability-oriented operations or strategies, and discuss its real-world advantages for application. Fourth, we can enhance neural process models by incorporating physical laws, for instance, by fitting a physical regularized GP model (Yuan et al., 2021; Zhu et al., 2022c) with neural networks. In this way, the model can leverage physical laws and knowledge to constrain relationships between variables, thereby augmenting its predictive accuracy and modeling capability.

Replication and data sharing

The author has uploaded the code package to an explicit Github repository at https://github.com/xm123155/AGNP_submission_commtr. The replication package was approved by the replication editor.

Declaration of competing interest

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

Acknowledgements

This work was partially supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2023C03155), Hong Kong Research Grants Council (Nos. HKUST16208920 and T41-603/20R), National Natural Science Foundation of China (Nos. 71922019 and 72171210), and the Smart Urban Future (SURF) Laboratory, Zhejiang Province.

References

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L.D., Freeze, P.B., 2018. Missing data imputation for traffic flow speed using spatio-temporal cokriging. Transport. Res. C Emerg. Technol. 88, 124–139.
Candela, J.Q., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.
Chan, R.K.C., Lim, J.M.Y., Parthiban, R., 2021. A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system. Expert Syst. Appl. 171, 114573.
Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent neural networks for multivariate time series with missing values. Sci. Rep. 8 (1), 1–12.
Chen, X., He, Z., Wang, J., 2018. Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition. Transport. Res. C Emerg. Technol. 86, 59–77.
Chen, X., He, Z., Chen, Y., Lu, Y., Wang, J., 2019. Missing traffic data imputation and pattern discovery with a Bayesian augmented tensor factorization model. Transport. Res. C Emerg. Technol. 104, 66–77.
Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of Neural Information Processing Systems (NeurIPS).
Cui, Z., Henrickson, K., Ke, R., Wang, Y., 2019. Traffic graph convolutional recurrent neural network: a deep learning framework for network-scale traffic learning and forecasting. IEEE Trans. Intell. Transport. Syst. 21 (11), 4883–4894.
Cui, Z., Lin, L., Pu, Z., Wang, Y., 2020. Graph Markov network for traffic forecasting with missing data. Transport. Res. C Emerg. Technol. 117, 102671.
de Medrano, R., Aznarte, J.L., 2020. A spatio-temporal attention-based spot-forecasting framework for urban traffic prediction. Appl. Soft Comput. 96, 106615.
Deb, S., Strawderman, L., Carruth, D.W., DuBien, J., Smith, B., Garrison, T.M., 2017. Development and validation of a questionnaire to assess pedestrian receptivity toward fully autonomous vehicles. Transport. Res. C Emerg. Technol. 84, 178–195.
El Hamdani, S., Benamar, N., Younis, M., 2020. Pedestrian support in intelligent transportation systems: challenges, solutions and open issues. Transport. Res. C Emerg. Technol. 121, 102856.
Ganin, A.A., Mersky, A.C., Jin, A.S., Kitsak, M., Keisler, J.M., Linkov, I., 2019. Resilience in intelligent transportation systems (ITS). Transport. Res. C Emerg. Technol. 100, 318–329.
Gao, K., Yang, Y., Qu, X., 2021. Diverging effects of subjective prospect values of uncertain time and money. Commun. Transport. Res. 1, 100007.
Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D.J., Eslami, S.M., Teh, Y.W., 2018. Neural processes. In: Proceedings of the International Conference on Machine Learning (ICML).
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Huang, F., Yi, P., Wang, J., Li, M., Peng, J., Xiong, X., 2022. A dynamical spatial-temporal graph neural network for traffic demand prediction. Inf. Sci. 594, 286–304.
Jia, Y., Wu, J., Xu, M., 2017. Traffic flow prediction with rainfall impact using a deep learning method. J. Adv. Transport., 6575947
Kaur, M., Singh, S., Aggarwal, N., 2022. Missing traffic data imputation using a dual-stage error-corrected boosting regressor with uncertainty estimation. Inf. Sci. 586, 344–373.
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. Transport. Res. C Emerg. Technol. 85, 591–608.
Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., et al., 2019. Attentive neural processes. In: Proceedings of the International Conference on Learning Representations (ICLR).
Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: Proceedings of the International Conference on Learning Representations (ICLR).
Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (ICLR).
Lee, K., Rhee, W., 2022. DDP-GCN: multi-graph convolutional network for spatiotemporal traffic forecasting. Transport. Res. C Emerg. Technol. 134, 103466.
Li, L., Li, Y., Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transport. Res. C Emerg. Technol. 34, 108–120.
Li, L., Zhang, J., Wang, Y., Ran, B., 2018. Missing value imputation for traffic-related time series data based on a multi-view learning method. IEEE Trans. Intell. Transport. Syst. 20 (8), 2933–2943.
Li, H., Li, M., Lin, X., He, F., Wang, Y., 2020. A spatiotemporal approach for traffic data imputation with complicated missing patterns. Transport. Res. C Emerg. Technol. 119, 102730.
Liu, L., Chen, R.C., 2017. A novel passenger flow prediction model using deep learning methods. Transport. Res. C Emerg. Technol. 84, 74–91.
Liu, S., Yue, Y., Krishnan, R., 2013. Adaptive collective routing using Gaussian process dynamic congestion models. In: Proceedings of the 19th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining.
Liu, Q., Wang, B., Zhu, Y., 2018. Short-term traffic speed forecasting based on attention convolutional neural network for arterials. Comput. Aided Civ. Infrastruct. Eng. 33 (11), 999–1016.
Liu, Y., Liu, Z., Jia, R., 2019. DeepPF: a deep learning based architecture for metro passenger flow prediction. Transport. Res. C Emerg. Technol. 101, 18–34.
Luan, S., Ke, R., Huang, Z., Ma, X., 2022. Traffic congestion propagation inference using dynamic Bayesian graph convolution network. Transport. Res. C Emerg. Technol. 135, 103526.
Matthews, A.G.D.G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., et al., 2017. GPflow: a Gaussian process library using TensorFlow. J. Mach. Learn. Res. 18 (40), 1–6.
Ni, D., Leonard, J.D., Guin, A., Feng, C., 2005. Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. J. Transport. Eng. 131 (12), 931–938.
Ortuzar, J.D.D., 2021. Future transportation: sustainability, complexity and individualization of choices. Commun Transport Res 1, 100010.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Peng, H., Du, B., Liu, M., Liu, M., Ji, S., Wang, S., et al., 2021. Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning. Inf. Sci. 578, 401–416.
Qu, L., Li, L., Zhang, Y., Hu, J., 2009. PPCA-based missing data imputation for traffic flow volume: a systematical approach. IEEE Trans. Intell. Transport. Syst. 10 (3), 512–522.
Rasmussen, C.E., Williams, C.K.I., 2006. Classification. In: *Gaussian Processes for Machine Learning*. MIT press, Cambridge, MA, pp. 7–32.
Ran, B., Jin, P.J., Boyce, D., Qiu, T.Z., Cheng, Y., 2012. Perspectives on future transportation research: impact of intelligent transportation system technologies on next-generation transportation modeling. J. Intell. Transport. S. 16 (4), 226–242.
Rodrigues, F., Pereira, F.C., 2018. Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data. Transport. Res. C Emerg. Technol. 95, 636–651.
Rodrigues, F., Henrickson, K., Pereira, F.C., 2018. Multi-output Gaussian processes for crowdsourced traffic data imputation. IEEE Trans. Intell. Transport. Syst. 20 (2), 594–603.
Salamanis, A., Kehagias, D.D., Filelis-Papadopoulos, C.K., Tzovaras, D., Gravvanis, G.A., 2016. Managing spatial graph dependencies in large volumes of traffic data for travel-time prediction. IEEE Trans. Intell. Transport. Syst. 17 (6), 1678–1687.
Smith, B.L., Williams, B.M., Oswald, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transport. Res. C Emerg. Technol. 10 (4), 303–321.
Spana, S., Du, L., 2022. Optimal information perturbation for traffic congestion mitigation: Gaussian process regression and optimization. Transport. Res. C Emerg. Technol. 138, 103647.
Sumalee, A., Ho, H.W., 2018. Smarter and more connected: future intelligent transportation system. IATSS Res. 42 (2), 67–71.
Ta, X., Liu, Z., Hu, X., Yu, L., Sun, L., Du, B., 2022. Adaptive Spatio-Temporal Graph Neural Network for Traffic Forecasting. Knowledge-Based Systems, 108199.
Tang, J., Zeng, J., 2022. Spatiotemporal gated graph attention network for urban traffic flow prediction based on license plate recognition data. Comput. Aided Civ. Infrastruct. Eng. 37 (1), 3–23.
Tang, J., Zhang, X., Yin, W., Zou, Y., Wang, Y., 2021. Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory. J. Intell. Transport. S. 25 (5), 439–454.

Tian, Y., Zhang, K., Li, J., Lin, X., Yang, B., 2018. LSTM-based traffic flow prediction with missing data. Neurocomputing 318, 297–305.
Wang, W., Wu, Y., 2021. Is uncertainty always bad for the performance of transportation systems? Commun. Transport. Res. 1, 100021.
Xu, M., Di, Y., Yang, H., Chen, X., Zhu, Z., 2023. Multi-task supply-demand prediction and reliability analysis for docked bike-sharing systems via transformer-encoder-based neural processes. Transport. Res. C Emerg. Technol. 147, 104015.
Yang, B., Kang, Y., Yuan, Y., Huang, X., Li, H., 2021a. ST-LBAGAN: spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation. Knowl. Base Syst. 215, 106705.
Yang, J.M., Peng, Z.R., Lin, L., 2021b. Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and Graph Laplacian regularized matrix factorization. Transport. Res. C Emerg. Technol. 129, 103228.
Ye, J., Xue, S., Jiang, A., 2021. Attention-based spatio-temporal graph convolutional network considering external factors for multi-step traffic flow prediction. Digital Communications and Networks 8 (3), 343–350.
Yu, B., Yin, H., Zhu, Z., 2017. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).
Yuan, Y., Zhang, Z., Yang, X.T., Zhe, S., 2021. Macroscopic traffic flow modeling with physics regularized Gaussian process: a new insight into machine learning applications in transportation. Transp. Res. Part B Methodol. 146, 88–110.
Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Zhang, Y., Cheng, T., Ren, Y., 2019. A graph deep learning method for short-term traffic forecasting on large road networks. Comput. Aided Civ. Infrastruct. Eng. 34 (10), 877–896.
Zhang, K., He, Z., Zheng, L., Zhao, L., Wu, L., 2021a. A generative adversarial network for travel times imputation using trajectory data. Comput. Aided Civ. Infrastruct. Eng. 36 (2), 197–212.
Zhang, Z., Lin, X., Li, M., Wang, Y., 2021b. A customized deep learning approach to integrate network-scale online traffic data imputation and prediction. Transport. Res. C Emerg. Technol. 132, 103372.
Zhu, Z., Peng, B., Xiong, C., Zhang, L., 2016. Short-term traffic flow prediction with linear conditional Gaussian Bayesian network. J. Adv. Transport. 50 (6), 1111–1123.
Zhu, Z., Mardan, A., Zhu, S., Yang, H., 2021. Capturing the interaction between travel time reliability and route choice behavior based on the generalized Bayesian traffic model. Transp. Res. Part B Methodol. 143, 48–64.
Zhu, K., Zhang, S., Li, J., Zhou, D., Dai, H., Hu, Z., 2022a. Spatiotemporal multi-graph convolutional networks with synthetic data for traffic volume forecasting. Expert Syst. Appl. 187, 115992.
Zhu, Z., Xu, M., Di, Y., Chen, X., Yu, J., 2022b. Modeling ride-sourcing matching and pickup processes based on additive Gaussian process models. Transport. Bus.: Transport Dynamics 11 (1), 590–611.
Zhu, Z., Xu, M., Di, Y., Yang, H., 2022c. Fitting spatial-temporal data via a physics regularized multi-output grid Gaussian process: case studies of a bike-sharing system. IEEE Trans. Intell. Transport. Syst. 23 (11), 21090–21101.
Zhu, Z., Xu, M., Ke, J., Yang, H., Chen, X., 2023. A Bayesian clustering ensemble Gaussian process model for network-wide traffic flow clustering and prediction. Transport. Res. C Emerg. Technol. 148, 104032.

Meng Xu received the B.S. degree in Logistics Management from Nankai University and the M.S. degree in Information Technology from The Hong Kong University of Science and Technology. She is currently pursuing the Ph.D. degree with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology. Her research interests include economic analysis in multi-modal transportation systems, statistical learning, and machine learning methodologies in transportation systems and reinforcement learning for transport simulation and optimization.

Yining Di received the B.S. degree in Civil Engineering from Zhejiang University, China, and the B.S. degree of science in Civil Engineering from University of Illinois at Urbana-Champaign, USA. He is currently pursuing the Ph.D. degree in Civil Engineering from The Hong Kong University of Science and Technology. His primary research interests are data-driven transportation analysis, intelligent traffic system, computer-aided visual analysis, and machine learning.

Hongxing Ding received the B.S. degree in Construction Management from Hunan University and the M.S. degree in Management Science and Engineering from Southeast University. She is currently pursuing the Ph.D. degree with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology. Her research interests include economic analysis in multi-modal transportation systems and transportation network modeling and optimization.

Zheng Zhu is a Professor and Assistant Department Head in College of Civil Engineering and Architecture at Zhejiang University. Before that, he had been working as a Research Assistant Professor at The Hong Kong University of Science and Technology (2018–2021). He got the Ph.D. degree (2018) and M.S. degree (2014) in Transportation Engineering, the M.A. degree in Statistics (2017) at the University of Maryland, and the B.S. degree in Hydraulic Engineering (2012) at Tsinghua University. He has been working on the research area of multi-modal transportation systems for years, which covers major scientific questions and development needs, such as understanding supply-demand mechanisms, discovering efficient operational and management strategies, integrating artificial intelligence methodologies, and supporting real-world applications. He has published over 40 SCI/SSCI indexed research papers, and his current Google Scholar h-index is 18.

Xiqun (Michael) Chen is Tenured Professor of Zhejiang University, Director of Institute of Intelligent Transportation Systems, Vice Dean of ZJU-UIUC Institute. His research interests include intelligent transportation systems, shared mobility on demand, simulation-based optimization, and transportation big data analytics. He received the B.E. and Ph.D. degrees from the Department of Civil Engineering, Tsinghua University, in 2008 and 2013, respectively. He has published over 100 peer-reviewed journal articles and over 70 conference papers. He received the National Excellent Young Scholars Award of National Natural Science Foundation of China, Science and Technology Innovation Youth Award of China Communications and Transportation Association, Science and Technology Award of China Intelligent Transportation Systems Association, the 2013 IEEE Intelligent Transportation Systems Society Best Ph.D. Dissertation Award, and Best Paper Awards at six international conferences.

Hai Yang is currently a Chair Professor at The Hong Kong University of Science and Technology. He is internationally known as an active scholar in the field of transportation, with more than 250 papers published in SCI/SSCI indexed journals and a Google Scholar h-index citation rate of 83. Most of his publications appeared in leading international journals, such as *Transportation Research*, *Transportation Science*, *and Operations Research*. He received a number of national and international awards, including Frank M. Masters Transportation Engineering Award, American Society of Civil Engineers (2020), and National Natural Science Award bestowed by the State Council of PR China (2011). He served as the Editor-in-Chief of *Transportation Research Part B: Methodological* from 2013 to 2018 and is now a distinguished editorial board member of this journal.