SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2024, Vol. 67, Iss. 5, 152303:1–152303:15
https://doi.org/10.1007/s11432-023-3899-1

A credible traffic prediction method based on self-supervised causal discovery

Dan WANG, Yingjie LIU & Bin SONG*

State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, *China*

Received 17 April 2023/Revised 2 August 2023/Accepted 17 November 2023/Published online 26 April 2024

Abstract Next-generation wireless network aims to support low-latency, high-speed data transmission services by incorporating artificial intelligence (AI) technologies. To fulfill this promise, AI-based network traffic prediction is essential for pre-allocating resources, such as bandwidth and computing power. This can help reduce network congestion and improve the quality of service (QoS) for users. Most studies achieve future traffic prediction by exploiting deep learning and reinforcement learning, to mine spatio-temporal correlated variables. Nevertheless, the prediction results obtained only by the spatio-temporal correlated variables cannot reflect real traffic changes. This phenomenon prevents the true prediction variables from being inferred, making the prediction algorithm perform poorly. Inspired by causal science, we propose a novel network traffic prediction method based on self-supervised spatio-temporal causal discovery (SSTCD). We first introduce the Granger causal discovery algorithm to build a causal graph among prediction variables and obtain spatio-temporal causality in the observed data, which reflects the real reasons affecting traffic changes. Next, a graph neural network (GNN) is adopted to incorporate causality for traffic prediction. Furthermore, we propose a self-supervised method to implement causal discovery to to address the challenge of lacking ground-truth causal graphs in the observed data. Experimental results demonstrate the effectiveness of the SSTCD method.

Keywords wireless network traffic prediction, causal discovery, self-supervised

1 Introduction

With the increasing popularity of smart applications and services, such as virtual reality (VR)/augmented reality (AR) [1], cloud games and panoramic videos [2], existing communication network carries explosive amounts of data [3], and massive computation demand [4]. According to Statista's report, global mobile devices are expected to grow from 15 billion in 2021 to 18.22 billion in 2025 [5]. This phenomenon will bring sudden traffic congestion and intolerable service delay problems to the network degrading the overall network performance. To avoid these problems, applying artificial intelligence (AI) technology to the sixth-generation wireless communication network for traffic prediction has been studied, which helps the network adopt the resource pre-allocation strategy to enhance the quality of service (QoS) to the users [6]. According to the studies in [7–10], most traffic prediction methods can be achieved by mining the spatio-temporal correlation between the prediction variables and the target variable of network traffic [11], which assume that the target variable is correlated with all prediction variables. However, these existing methods do not identify decisive prediction variables and no direct causality is between prediction variables and the target variable. In fact, much causal information exists in network traffic, such as regional characteristics. Specifically, the network traffic of the target area may related to the network traffic of the adjacent area, but not to the more distant area [12]. Therefore, if such causal information is used, the decisive variable can be captured, improving the accuracy and credibility of predictions.

Recent advances in causal inference theory [13] make it possible to mine causal information for traffic prediction. According to [14], causal inference is generally divided into two parts. The first part is the causal discovery which discovers causalities among variables based on observational data. The second

* Corresponding author (email: bsong@mail.xidian.edu.cn)

part is the causal effect that implements treatments to observe whether the change in the outcome is significantly based on the assumption of causality. The above two methods can be used to mine causalities among variables. Inspired by causal thinking, we consider adopting causal inference to capture the causalities between prediction variables and target variables in traffic prediction, thereby enhancing the accuracy and credibility of predictions.

Nevertheless, it is a great challenge to exploit causality among variables for traffic prediction. On the one hand, it is difficult to explore causality among variables through observed traffic data. Since the ground-truth causal graph during the prediction process is difficult to obtain, the causality among the variables is difficult to capture. On the other hand, incorporating causality among the variables to obtain accurate prediction results imposes challenges on the algorithm. To address these challenges, we propose a wireless traffic prediction method based on self-supervised causal discovery. Specifically, the method contains two steps: causal discovery and causal prediction. We address the lack of ground-truth causal graphs in the observed data by implementing Granger causal discovery [15] in a self-supervised manner to obtain the spatio-temporal causality between prediction variables and target variables. Furthermore, we use a graph neural network (GNN)-based variational autoencoder to incorporate the spatio-temporal causality into traffic prediction.

In the proposed work, we divide the entire prediction area into cells, and the historical data of each cell are used as prediction variables. We construct a causal graph using the encoder, which contains the spatio-temporal causality among the prediction variables. Combining the obtained spatio-temporal causality, we predict future traffic using the decoder. In our work, a self-supervised method is designed to conduct causal discovery for solving the non-existent ground-truth causal graph problem in the observed data. This method can improve the accuracy of causal discovery by minimizing the prediction error, further improving prediction accuracy.

Specifically, the main contributions of this paper can be summarized as follows:

• We are the first to propose a wireless traffic prediction method based on Granger causal discovery to explore spatio-temporal causality between prediction variables and target variables. Then, the prediction model can eliminate the effects of non-causal variables, thereby improving traffic prediction accuracy and credibility.

• We then design a self-supervised approach to achieve causal discovery. It improves the prediction performance of causal discovery in traffic. This is done by improving the accuracy of the causal traffic prediction task, which allows us to maximize the accuracy of causal discovery without ground-truth causal graphs.

• We validate the effectiveness of the algorithm by testing it on real data and comparing it with the current state-of-the-art methods. The experimental results demonstrate the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 introduces related works on wireless traffic prediction and causal inference. Section 3 presents the traffic prediction model based on causal discovery. Then, we present and analyze the experimental results in Section 4. Finally, Section 5 concludes the paper.

2 Related work

We investigate existing work on both wireless traffic prediction and causal inference in this section.

2.1 Wireless traffic prediction

With the surge in traffic brought about by the popularization of VR/AR, panoramic video and other applications, wireless traffic prediction has attracted much attention in recent years. These applications need to implement more data transmission and lower latency requirements than previous applications. This phenomenon usually brings sudden network congestion and unsatisfied user QoS problems. Hence, it is important to research accurate traffic modeling and prediction capabilities to avoid these problems. Wireless traffic prediction is a time series prediction that includes traditional and deep learning methods.

According to the investigation, autoregressive integrated moving average (ARIMA) [16] is a classic traffic prediction method. It predicted future time series from the perspective of the probability theory and incorporated multiple time series models, including autoregression, moving average, and autoregressive moving average (ARMA). It can achieve better prediction results in wireless traffic with different

characteristics. However, due to the computational complexity, it was only applied to predictions of a single variable. Meanwhile, researchers have also demonstrated that it can only predict regular variables in wireless traffic. Therefore, Ref. [17] proposed a multivariate prediction method, i.e., the vector autoregressive model (VAR). VAR predicted the future time series values of multiple variables at once. However, since it cannot handle nonlinear relationships among variables, it performed poorly in real traffic prediction.

On the other hand, there are various prediction methods based on deep learning that enable training a good predictive model through a large amount of historical traffic data. In [18], long-short term memory networks (LSTM) were introduced to capture temporal dependencies among different variables through memory mechanisms. To consider dynamic space interactions between regions, Ref. [19] proposed to combine the relative-flow gating mechanism (RGM) and convolutional neural network (CNN) to learn dynamic space dependency. In addition, the author in [20] used multiple convolution kernels to perform convolution operations on time series to achieve feature extraction. Ref. [21] proposed a traffic prediction method based on transfer learning for urban cellular traffic. It divided the city into different groups and used the similarity among groups to reuse knowledge. A dual attention-based federated learning method [22] was proposed to train a high-quality predictive model with multiple edge clients. Specifically, a quasi-global model was shared among clients, thus addressing the statistical heterogeneity of federated learning.

The existing studies above are based on data correlation to predict future traffic data. All possible relevant variables are used in the model to improve prediction accuracy. However, some of them do not have direct causality with the target variable, leading to low accuracy and unreliable prediction results. Therefore, to solve the above problem, we consider employing a causal discovery algorithm to mine the causality among the prediction variables, thereby improving the prediction accuracy and credibility.

2.2 Causal inference

As an important technology in AI, the deep learning algorithm is trained by a large amount of multi-angle data and a deep learning model with high performance can be obtained. However, there are still some problems. Due to the lack of credibility of the deep learning network, the model cannot distinguish decisive prediction variables, and the model cannot be applied to important occasions. More recently, causal science, described by Judea Pearl, the father of Bayesian networks, has been used in deep learning to address the transfer and credibility problems of deep learning [23, 24]. It is a powerful statistical analysis tool that includes causal discovery and causal effect.

As an important part of causal inference, causal discovery has the ability to discover causal relationships among variables that can be applied to prediction problems. In [25], the authors proposed a causal discovery algorithm that utilized an attention mechanism to understand the predictor variables of focus by CNNs when making predictions and intervene in that variable to discover relationships among variables. Additionally, when the pre-sampling method lost the shared information among the samples, the authors found a causal relationship among the predictors according to the dynamics among the shared samples [26]. In [27], to reduce the computational cost and calculate only the causal relationships among the variables of interest, the authors introduced a local causal discovery algorithm to learn the parents and the children of the target variable according to the backward frame. This method can obtain variable relationships by finding the invariant V structure. In addition, for nonlinear relationships, the authors in [28] proposed a causal discovery algorithm based on nonlinear independent component analysis (ICA), which can infer the causal direction through a series of independence tests.

However, all of the above studies perform causal discovery on datasets with clear ground-truth causalities, and there is no known ground-truth causal graph in real traffic data. We cannot obtain a ground-truth causal graph for traffic prediction. To tackle this issue, we propose a self-supervised way to perform causal discovery to achieve causal traffic prediction.

3 Traffic prediction model based on causal discovery: SSTCD

This section introduces the proposed network traffic prediction model based on causal discovery. The model is composed of two parts: causal discovery and causal prediction. Here, causal discovery is implemented in a self-supervised manner, and causal prediction leverages GNN to achieve. The overall model is illustrated in Figure 1.

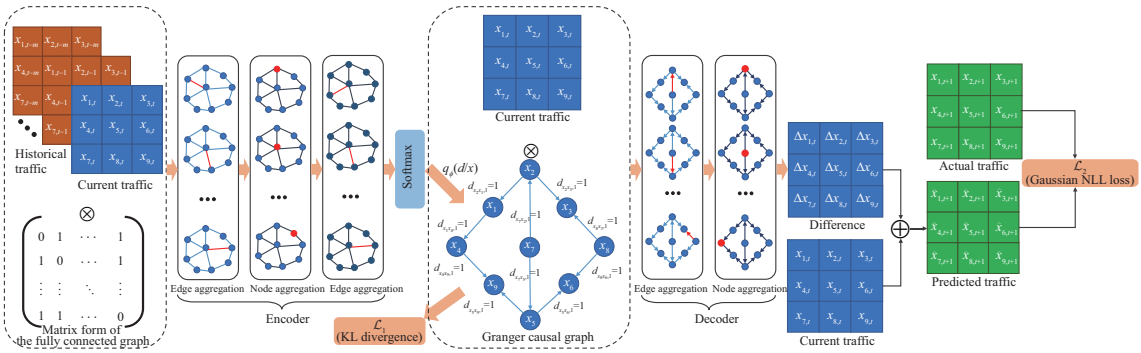

Figure 1 (Color online) Traffic prediction model based on self-supervised causal discovery. The left side of the figure contains multiple 3×3 squares, where each element represents a cell. The blue squares represent the current traffic, and the orange squares represent the historical traffic. The number of cells in the figure is an example, and the cell size used in the experiment is 100×100. The red nodes and edges in the encoder and decoder represent aggregated edges and nodes.

3.1 Background: Granger causality

Granger causality is a method that infers the causality between observed variables from the observed time series data. The central idea of Granger causality is that cause and effect have a sequence in the time dimension: if the past elements of time series X can improve the prediction results of time series Y in the future, then X "Granger causes" Y.

3.2 Problem formulation

In this work, we consider the problem of urban network traffic prediction. Given an urban region D, we divide it into $N \times N$ cells. Each cell has its own network traffic data. We define the network traffic data of the region as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N^2}\}$. Each element in $\boldsymbol{X}$ represents the network traffic data in one cell, formulated as $\boldsymbol{x}_i = \{x_{i,t}, x_{i,t-1}, \ldots, x_{i,t-m}, \ldots, x_{i,1}\}$, $x_{i,t}$ denotes the current network traffic and the other elements denote the historical network traffic. This paper focuses on the one-step ahead network traffic prediction problem of cell i and improves the prediction performance by modeling in time and space. The network traffic prediction problem is described as predicting future network traffic based on the current and historical network traffic data of cell i and its adjacent cells $\mathcal{N}_i$. Here, the predicted future network traffic of cell i is defined as $\hat{x}_{i,t+1}$. The network traffic prediction problem can be formulated as

$$\hat{x}_{i,t+1} = \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}; \theta), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the prediction model, $\boldsymbol{x}_{\mathcal{N}_i}$ is the network traffic data of all cells adjacent to cell i, and θ denotes the learnable parameters in the prediction model. To reduce the error between the predicted and the real future network traffic, the objective function is defined as follows:

$$\arg\min_{\theta} \left\{ \sum_{i=1}^{\mathcal{N}_i} \sum_{t=1}^{T} \mathcal{L}\left(\mathcal{F}(\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}; \theta), x_{i,t+1}\right) \right\}, \tag{2}$$

where $\mathcal{L}$ is the loss function and $x_{i,t+1}$ is the real future network traffic.

3.3 Proposed method: SSTCD

This subsection proposes a network traffic prediction model, SSTCD (self-supervised spatio-temporal causal discovery), as shown in Figure 1. Suppose network traffic in regions is correlated in time and space. The prediction performance improves by modeling spatio-temporal information in network traffic [29]. However, predicting network traffic based on correlations causes the input variables without direct causality to negatively affect the prediction results in current prediction methods. To solve this problem, this paper designs a spatio-temporal network traffic prediction model based on the causal discovery that can predict future network traffic by mining the causality between the temporal and spatial network traffic.

Granger causality is one of the most common methods to mine causal relationships among variables through observational data. We first use Granger causality to infer the spatio-temporal causal relationship among the network traffic of different cells. Since the network traffic changes are complex and nonlinear

relationships in reality, we adopt the nonlinear Granger causality method to obtain the causal variables that affect the prediction results. Then, we predict the future traffic based on the causal variables. The nonlinearity Granger causality is defined as follows [30].

Definition 1. For the predicted cell i, its corresponding network traffic is defined as a variable $\boldsymbol{x}_i$. Similarly, the network traffic variable of its adjacent cells is $\boldsymbol{x}_{\mathcal{N}_i}$. It is assumed that there are N adjacent cells, denoted by $\boldsymbol{x}_{\mathcal{N}_i} = \{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_N}\}$, where $i \notin \mathcal{N}_i$. $\boldsymbol{x}_i$ and $\boldsymbol{x}_{\mathcal{N}_i}$ across time-steps $T = \{1, \ldots, t-1\}$. Given a nonlinear autoregressive function $\mathcal{F}$, the predicted network traffic is

$$\hat{x}_{i,t+1} = \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}) + \varepsilon_i^{t+1}, \tag{3}$$

where ε_i^{t+1} is independent noise. If $\mathcal{F}$ depends on $\boldsymbol{x}_i$ and $\boldsymbol{x}_{\mathcal{N}_i}$, the network traffic variables of adjacent cells are Granger causes the network traffic variable of cell i.

Our method consists of two parts: causal discovery and causal prediction, which are achieved through a variational autoencoder (VAE). Incorporating the definition of the Granger causality, we formulate the two parts as follows:

$$\hat{x}_{i,t+1} = \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}, \mathcal{G}) + \varepsilon_i^{t+1}, \tag{4}$$

where $\mathcal{G}$ is the inferred causal graph. The two parts of our method are illustrated in (4). Here, the causal graph is inferred through a GNN, and the future network traffic can be predicted through a function with parameters. Therefore, Eq. (4) can be reformulated as

$$\hat{x}_{i,t+1} = \mathcal{F}_\theta(\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}, \mathcal{F}_\phi) + \varepsilon_i^{t+1}, \tag{5}$$

where $\mathcal{F}_\phi$ and $\mathcal{F}_\theta$ are modeled as the encoder and decoder of VAE, respectively.

Next, we will elaborate on the causal discovery and causal prediction in our method.

3.4 Causal discovery

To build a causal graph among variables, the encoder must infer the type of relationships among the variables. In our paper, we infer the spatio-temporal causality among prediction variables and target variables through the observed data. In the observed data, we use recent traffic data to predict future traffic data due to historical traffic data with a large time span having little impact on future traffic data. Here, historical traffic data of different lengths are obtained by setting the size of the sliding window.

To tackle the problem of groundtruth causal graphs being non-existent in the observed data, we use GNN to infer the relationship among variables on a fully connected graph. We set each variable as a node, and an edge connects each node. Given the input variables $\boldsymbol{x}_i$, $\boldsymbol{x}_{\mathcal{N}_i}$ of the encoder $\mathcal{F}_\phi$, the message pass process can be expressed as

$$\boldsymbol{h}_i^1 = f_{\text{emb}}(\boldsymbol{x}_i), \tag{6}$$

$$\boldsymbol{h}_{i,j}^1 = f_e^1(\boldsymbol{h}_i^1, \boldsymbol{h}_j^1), \tag{7}$$

$$\boldsymbol{h}_i^2 = f_v^1\left(\sum_{i \neq j} \boldsymbol{h}_{i,j}^1\right), \tag{8}$$

$$\boldsymbol{h}_{i,j}^2 = f_e^2(\boldsymbol{h}_i^2, \boldsymbol{h}_j^2), \tag{9}$$

$$q_\phi(d_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = \text{softmax}(\boldsymbol{h}_{i,j}^2), \tag{10}$$

where ϕ represents the parameters of the neural network, $\boldsymbol{h}_i^l$ is the embedding of node in layer l, $\boldsymbol{h}_{i,j}^l$ is an embedding of the edge and $j \in \mathcal{N}_i$. We use fully-connected networks (multi-layer perceptrons, MLPs) as the functions $f(\cdots)$ (i.e., f_{emb}, f_e^1, f_v^1, and f_e^2). The edge type d_{ij} is sampled from $q_\phi(d_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)$. In our work, there are two edge types: "no edge"-type $d_{ij,0}$ and "directed edge"-type $d_{ij,1}$. More specifically, $d_{ij,1} = 1$ represents a directed edge between variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ and the causal direction is from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$. In addition, $d_{ij,0} = 1$ means no direct causality exists between variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Lemma 1 ([26])**.** If all variables are observable and there are no instantaneous connections between variables, Granger causality is equivalent to causality in the underlying directed acyclic graph (DAG).

In the constructed causal graph, the variables are all observable. Also, we consider that all variables have a sequence in the time dimension, indicating that there are no instantaneous connections between variables. Therefore, according to the above lemma, the inferred causal direction is equivalent to the causal relations.

Algorithm 1 Traffic prediction based on a self-supervised causal discovery algorithm

Input: Random initialization parameters ϕ, θ of the encoder and decoder; the current and historical traffic of cells $\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}$.
Output: Optimal traffic prediction model.

for $j \leftarrow 1, \ldots, J$ **do**
  Input $\boldsymbol{x}_i, \boldsymbol{x}_{\mathcal{N}_i}$ into the encoder network;
  Calculate the embedding vector $\boldsymbol{h}_i^1$ according to (6);
  Aggregate edges and nodes of GNN according to (7)–(9);
  Construct a causal graph between variables according to (10);
  Input current traffic x_i^t and causal graph into decoder;
  Edge aggregation and node aggregation according to (11) and (12);
  Output the predicted traffic value $\hat{x}_{i,t+1}$;
  Calculate loss function $\mathcal{L}$ according to (15);
  Update $\phi \leftarrow \phi - \nabla_\phi \mathcal{L}(\phi)$;
  Update $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}(\theta)$;
end for

3.5 Causal prediction

After we get the causal graph between variables, we use a decoder to combine the causal graph to predict future network traffic. It takes the network traffic at time t and the causality d_{ij} among the variables as input. The message is passed according to the causal direction of the constructed causal graph. We denote the message pass process as

$$\boldsymbol{h}_{i,j}^{t\prime} = \sum_k d_{ij,k} f_e'(x_i^t, x_j^t), \tag{11}$$

$$\hat{x}_{i,t+1} = x_{i,t} + f_v' \left(\sum_{i \neq j} \boldsymbol{h}_{i,j}^{t\prime} \right), \tag{12}$$

where k is the edge type. Note that our model predicts the difference between the network traffic at times t and $t+1$.

3.6 Loss function

To accurately construct the causal graph among variables and predict future traffic, we construct loss functions in causal discovery and causal prediction of the model, respectively. We use KL divergence as the loss function for causal discovery and Gaussian negative log-likelihood (NLL) as the loss function for causal prediction. The total loss function is

$$\mathcal{L} = \mathcal{L}_{\text{NLL}}(\hat{x}_{i,t+1}, x_{i,t+1}) - \text{KL}\left[q_\phi(\boldsymbol{d} \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \| p_\theta(\boldsymbol{d})\right]. \tag{13}$$

The above equation is the evidence lower bound (ELBO) of the VAE. Note that $p_\theta(\boldsymbol{d})$ denotes the prior information of the ground truth causal graph. Nevertheless, there is no groundtruth causality in the observed traffic data in reality. Therefore, we use a reparameterizable approximation to estimate the KL term. In the KL term, $p_\theta(\boldsymbol{d})$ is uniformly distributed prior to encouraging sparser graphs if there is "no edge"-type of edge between prediction variables. The KL term can be estimated as the sum of entropy plus a constant, which is denoted by

$$\sum H\left(q_\phi(\boldsymbol{d} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)\right) + \epsilon, \tag{14}$$

where ϵ is a constant. Then, Eq. (13) is rewritten as

$$\mathcal{L} = \mathcal{L}_{\text{NLL}}(\hat{x}_{i,t+1}, x_{i,t+1}) - \sum H\left(q_\phi(\boldsymbol{d} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)\right) - \epsilon. \tag{15}$$

It can be seen from the equation that no terms with ground truth causality are included. We improve the accuracy of causal discovery by causal prediction tasks instead of groundtruth causal graphs, which is a self-supervised manner. Meantime, improving causal discovery accuracy will also enhance causal prediction performance.

In detail, the traffic prediction based on a self-supervised causal discovery algorithm is illustrated in Algorithm 1. In this algorithm, the model iterates J times in one epoch.

4 Experiments

In this section, we evaluate the effectiveness of the proposed SSTCD method through extensive experiments. Specifically, the superior predictive performance of our method is demonstrated by a comparison with state-of-the-art methods. Moreover, we compared the baseline methods and evaluated the influence of different parameters on the experimental results.

4.1 Dataset and evaluation metrics

The datasets used in this paper are recorded by Telecom Italia [31], which contain "Call" detail records (CDR) for the city of Milan and the province of Trentino. The city of Milan is divided into 10000 cells, while the province of Trentito is divided into 11466 cells. They cover traffic monitoring logs of devices from different sources collected from 11/01/2013 to 01/01/2014. This paper uses these two datasets to simulate traffic changes in different regions. Each piece of data in the datasets is the traffic data of each cell, and their time interval is 10 min. The datasets have three types of services: "SMS", "Call", and "Internet".

To evaluate prediction performance, we employ two metrics commonly used to evaluate prediction performance: root mean square error (RMSE) and mean absolute error (MAE). RMSE and MAE are used to measure the difference between the predicted values and the groundtruth. Here, the RMSE and MAE can be formulated as $\sqrt{\frac{\sum_i(\hat{x}_{i,t+1}-x_{i,t+1})^2}{\sum_i}}$ and $\frac{\sum_i|\hat{x}_{i,t+1}-x_{i,t+1}|}{\sum_i}$, where $\hat{x}_{i,t+1}$ and $x_{i,t+1}$ are the predicted and groundtruth future traffic for cell i on time $t+1$, respectively.

4.2 Baseline methods

In this paper, we compare the proposed method with the seven baseline methods to verify the effectiveness of the algorithm, as follows:

(a) Linear regression (LR) [32]**.** LR is the simplest model to predict future network traffic with a linear relationship.

(b) Support vector regression (SVR) [33]**.** SVR is an important application branch of support vector machine (SVM). It predicts future traffic through nonlinear dependencies in time.

(c) LSTM [34]**.** LSTM is a temporal recurrent neural network that is suitable for processing and predicting time series.

(d) Spatial-temporal cross-domain neural network (STCNet) [21]**.** STCNet utilizes spatio-temporal modeling and transfer learning to predict traffic among different types of cellular networks.

(e) Adaptive multi-receptive field spatial-temporal graph convolutional networks (AMF-STGCN) [35]**.** AMF-STGCN models spatio-temporal dependencies in mobile networks and applies attention to capture various receptive fields of heterogeneous base stations.

(f) Multi-view spatial-temporal graph network (MVSTGN) [36]**.** MVSTGN integrates attention and convolution mechanisms into traffic pattern analysis for mining spatio-temporal information of network traffic.

(g) Our method without causal discovery (Ours-CD). To highlight the critical role of causality in traffic prediction, we conduct an ablation experiment.

4.3 Algorithm predictive performance

We randomly select 16 adjacent cells in the datasets for traffic prediction. To facilitate comparison with other works, we randomly select cells in the region (40, 60). Here, sliding windows are used for training, validation, and test dataset construction. We set the sliding window size to 3. We set both the encoder and the decoder to have 16 hidden neurons. In addition, our model is trained for 500 epochs. Following the datasets partitioning results of [21], we use the traffic data from the first seven weeks for training and validation and the data from the last week for testing.

Table 1 [21,32–36] shows the prediction performance of the different methods. It can be seen from the table that our method outperforms the baseline methods on the Milan and Trentino datasets in RMSE and MAE. Compared to MVSTGN, our method achieves an average performance improvement of 30.4% and 19.1% in RMSE and MAE, respectively. The reasons why our model works better are as follows.

• LR achieves the worst performance among all methods because the real traffic changes are very complex, and predicting future traffic with a simple linear model is not feasible. However, the proposed

Table 1 The prediction performance of different methods is measured using RMSE and MAE

Method	Milan						Trentino					
	RMSE			MAE			RMSE			MAE		
	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet
LR [32]	93.5556	50.8538	233.8462	34.7213	26.0667	154.3343	37.4835	12.1744	31.1899	13.6867	6.6347	23.3793
SVR [33]	72.0058	45.9312	214.6154	29.3615	22.0680	129.3548	28.8495	10.9959	28.6250	11.5739	5.6169	19.5952
LSTM [34]	68.8889	44.5381	235.7692	39.8537	22.0013	141.0236	20.3863	6.9821	22.7604	10.5001	4.9783	20.9484
STCNet [21]	54.1664	33.3415	172.3077	28.6564	15.8556	98.1035	21.7020	7.9819	22.9820	11.2960	4.0357	14.8612
AMF-STGCN [35]	49.5324	32.0237	169.5224	27.7017	15.2928	98.0234	19.8454	7.6665	22.6105	10.9197	3.8924	14.8490
MVSTGN [36]	49.0515	30.9443	165.0445	24.9796	14.6816	88.6983	19.6527	7.4081	22.0133	9.8466	3.7369	13.4364
Ours-CD	66.6155	40.5864	210.8694	33.4869	15.6451	113.1154	26.6898	9.7164	28.1253	13.2001	3.9821	17.1352
Ours	**39.2290**	**21.3864**	**98.6580**	**19.5993**	**13.3644**	**64.9545**	**15.7173**	**5.1199**	**13.1588**	**7.7258**	**3.4016**	**9.8396**

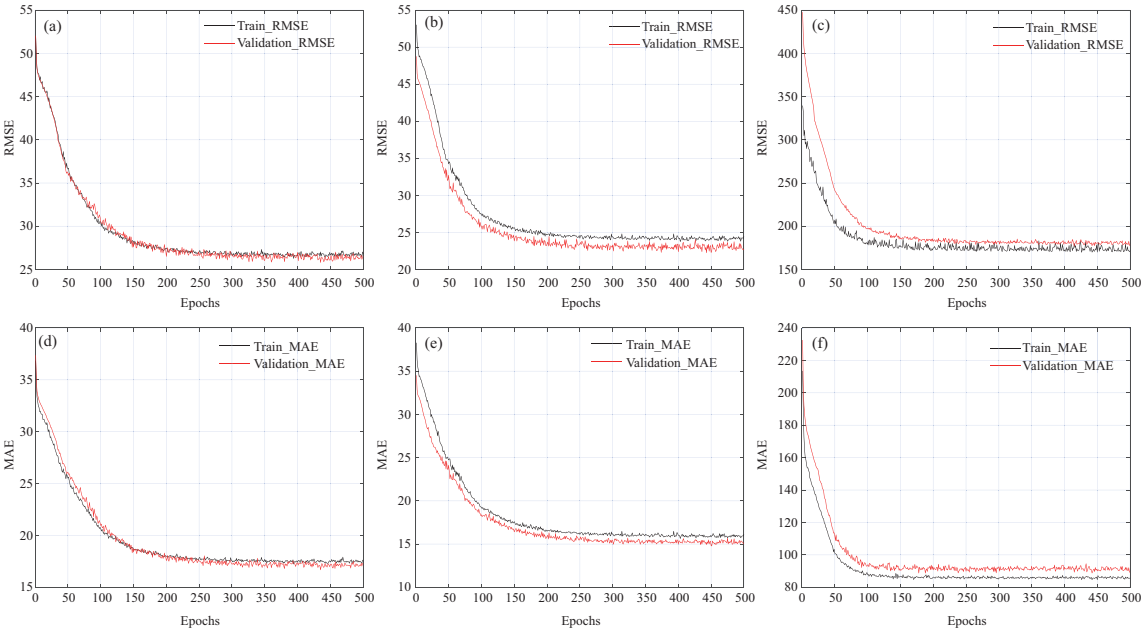

Figure 2 (Color online) Training and validation loss of our method on the Milan dataset. Training and validation RMSE loss of our method on (a) "SMS", (b) "Call", and (c) "Internet". Training and validation MAE loss of our method on (d) "SMS", (e) "Call", and (f) "Internet".

SSTCD adopts the nonlinear Granger causal discovery method to mine the nonlinear causality in traffic. Then, applying the decoder based on the graph network achieves a nonlinear prediction of future traffic.

• SVR and LSTM are nonlinear models that can handle nonlinear dependencies in time. However, since they only consider temporal dependencies and do not consider spatial dependencies simultaneously, the prediction results obtained are not the best. The proposed SSTCD can simultaneously mine the spatial causality and temporal causality of traffic through causal discovery. Further, the method improves traffic prediction performance by designing direct causality with multiple perspectives.

• Although STCNet, AMF-STGCN, and MVSTGN consider the spatio-temporal dependence of the traffic, their prediction performances are still poor performance to our method. The reason is that STCNet, AMF-STGCN, and MVSTGN utilize all correlated variables to predict the target variables rather than the causal variables. The non-causal variables change the distribution of predicted results, thereby decreasing the performance of traffic prediction. In contrast, SSTCD makes traffic predictions by finding causal relationships, thus preventing the interference of non-causal variables.

• To verify the impact of causal discovery on the overall predictive performance of the model, we perform experiments by removing the causal discovery step in the proposed method. That is, we assume that all variables are correlated (as in other studies). As can be seen from the results in the table, the prediction accuracy of Ours-CD is smaller than STCNet and much smaller than the proposed method, demonstrating that not all variable correlations positively affect prediction performance. In addition, due

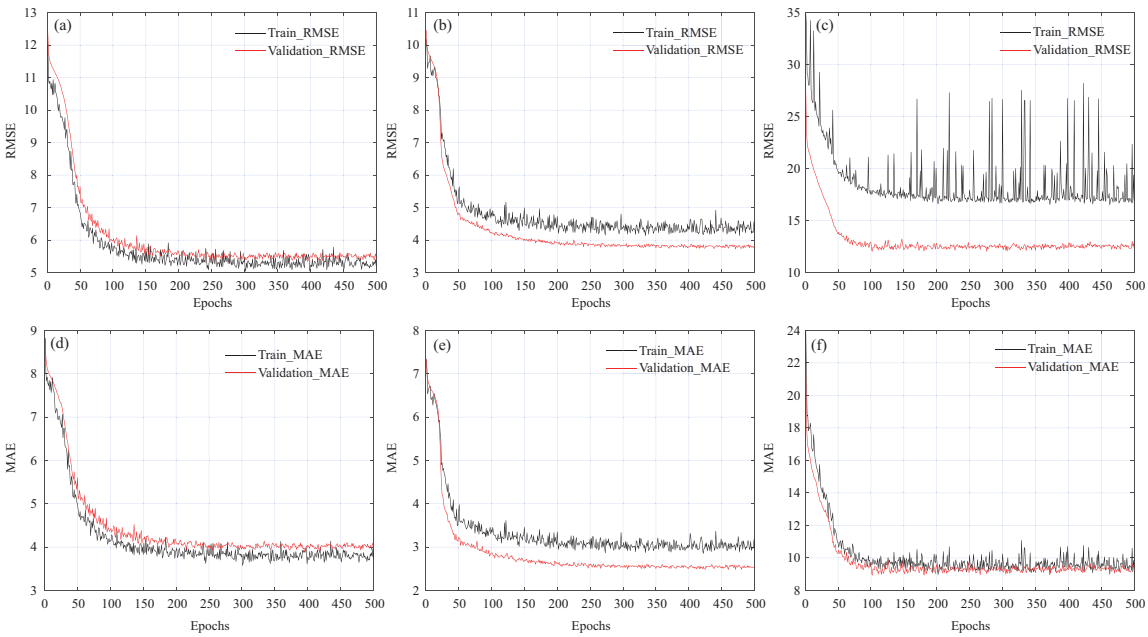

Figure 3 (Color online) Training and validation loss of our method on the Trentino dataset. Training and validation RMSE loss of our method on (a) "SMS", (b) "Call", and (c) "Internet". Training and validation MAE loss of our method on (d) "SMS", (e) "Call", and (f) "Internet".

to the transfer learning used in STCNet, which reduces the bias in the data, its prediction accuracy is higher than Ours-CD.

Figures 2 and 3 show the RMSE and MAE curves of our proposed method on different datasets, respectively. In each figure, the red curve represents the validation set loss, and the blue curve represents the training set loss. The training set loss for each figure first decreases and then stabilizes. This shows that our model continuously learns the network parameters throughout training and eventually stabilizes. The validation set loss for each figure is very close to the training set loss curve, indicating that the overall structure of the model is not problematic and that there is no overfitting. This shows that our prediction results are credible.

The prediction performance and cumulative distribution function (CDF) of the absolute error of our method on the different datasets are shown in Figures 4 and 5. It can be seen intuitively from the figures that the performance of the proposed method is better than MVSTGN. As seen from the CDF results of the Milan dataset, 95% of the "SMS", "Call" and "Internet" prediction errors are less than 23, 35, and 32, respectively. In the Trentino dataset, 94% of the "SMS", "Call" and "Internet" prediction errors are less than 12.331, 8.993, and 9.067, respectively. The errors are mostly concentrated on the left side of the x-axis. It is shown that our method has high prediction accuracy on the two datasets.

4.4 Causal graphs visualization

To visually show the influence of causality on the prediction results, we give the causal graphs generated during the training process in Figure 6. We plot the causality of 3×3 cells for the last epoch during training. It can be seen that Figure 6(c) has the most complicated causality, and Figure 6(a) has the simplest causality, which means "SMS" and "Call" were the main communication services at the time. In addition, based on the results shown in the figure, we can see that not all variables have a causality with each other. Therefore, if the correlation of all variables is considered, it will inevitably lead to wrong prediction results.

4.5 Influence of parameters

To evaluate the predictive performance of our method under different parameters, we conduct a series of experiments on cell, batch, and learning rate, and record the corresponding results.

Here, to verify the effect of different cell sizes on the prediction performance of the proposed method, we set the cell size to 9, 16, 25, and 36 on the two datasets, respectively. Note that the larger the number

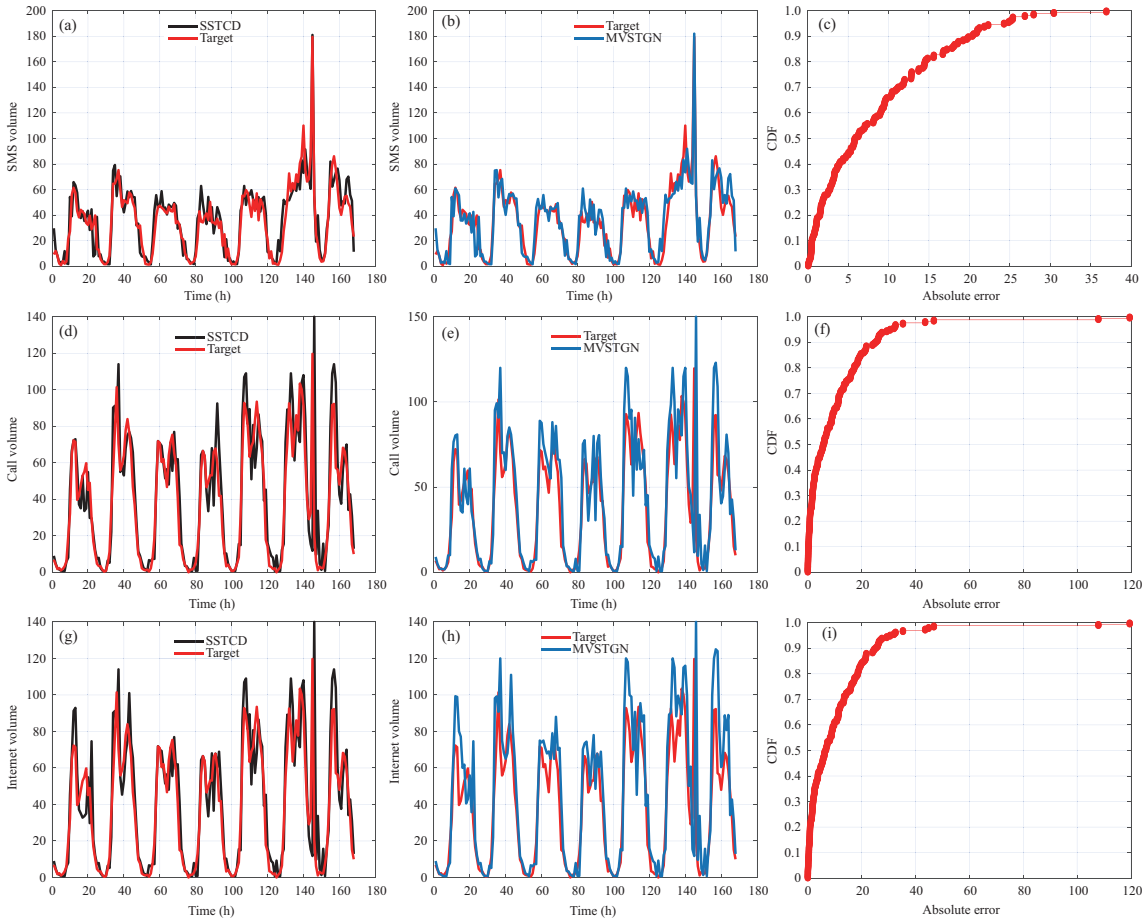

Figure 4 (Color online) Prediction results of our method on the Milan dataset. Prediction results of our method on (a) "SMS", (d) "Call", and (g) "Internet". Prediction results of MVSTGN on (b) "SMS", (e) "Call", and (h) "Internet". The CDF of our method on (c) "SMS", (f) "Call", and (i) "Internet".

of cells, the larger the cumulative error of the model. Here, we use data normalization to eliminate this error, and the RMSE and MAE results are shown in Figures 7 and 8. The figures show that the optimal cell size is different by dataset. The optimal cell size for the "SMS" and "Internet" is 16 in the Milan dataset, while the same for the "Call" is 25. The corresponding RMSEs are 0.0472, 0.0452, and 0.0094, respectively. Their lowest MAE are 0.0264, 0.0245, and 0.0063, respectively. Differently, the optimal cell size for "Internet" in the Trentino dataset is 16. This indicates that different regions have different service usage preferences, and our method mines the causalities of traffic between regions generated by this preference. Therefore, we use different cell sizes in different datasets to achieve optimal prediction performance. In addition, neither the larger cell size nor the smaller cell size obtained the optimal prediction results. This is because the prediction results do not consider causality beyond the region for small numbers of cells, while larger numbers of cells introduce spurious causality into the predictions.

In Figures 9 and 10, we evaluate the effect of different batch sizes on the model. Here, to find the batch size corresponding to the best prediction result, we set the batch size to 32, 64, 128, and 256. As can be seen in the figures, the optimal batch sizes corresponding to different services are different. Although the prediction performance of the model when the batch size is 32 and 64 is higher, the model is underfitting on "SMS" and "Call". The reason is that the spatio-temporal causalities between regions are very complex for "SMS" and "Call" services. As shown in Figure 6, the model cannot mine all the causalities between the variables with a smaller batch size. On the contrary, a larger batch size can provide the model with more data, allowing the model to discover more causalities within the data. Therefore, training the prediction model with a large batch size is necessary for the optimal traffic prediction results. The optimal RMSE values for the three services in the Milan dataset are 39.2290, 21.3864 and 98.6580, and 13.0518, 4.0880 and 13.1588 for the Trentino dataset.

Similarly, the learning rate is also an important parameter of our model. Different from other studies,

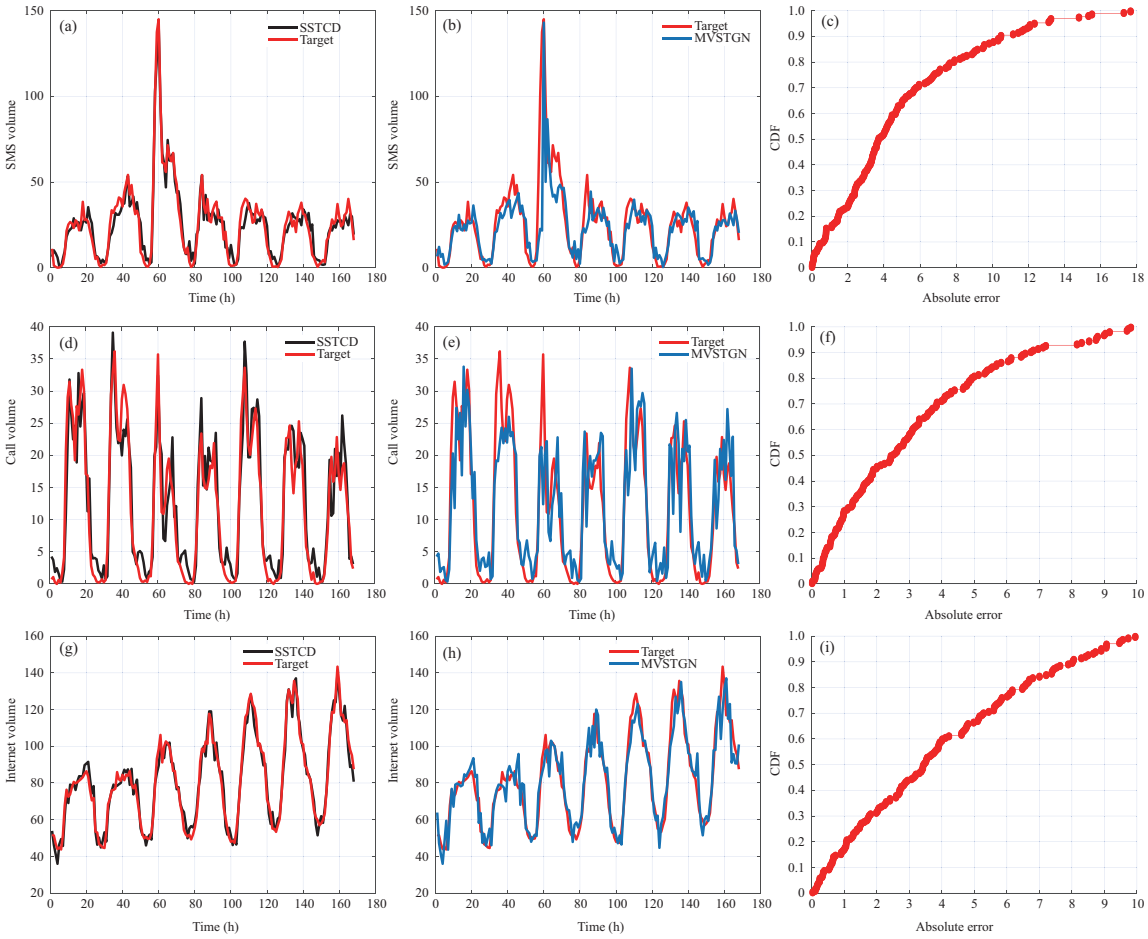

Figure 5 (Color online) Prediction results of our method on the Trentino dataset. Prediction results of our method on (a) "SMS", (d) "Call", and (g) "Internet". Prediction results of MVSTGN on (b) "SMS", (e) "Call", and (h) "Internet". The CDF of our method on (c) "SMS", (f) "Call", and (i) "Internet".

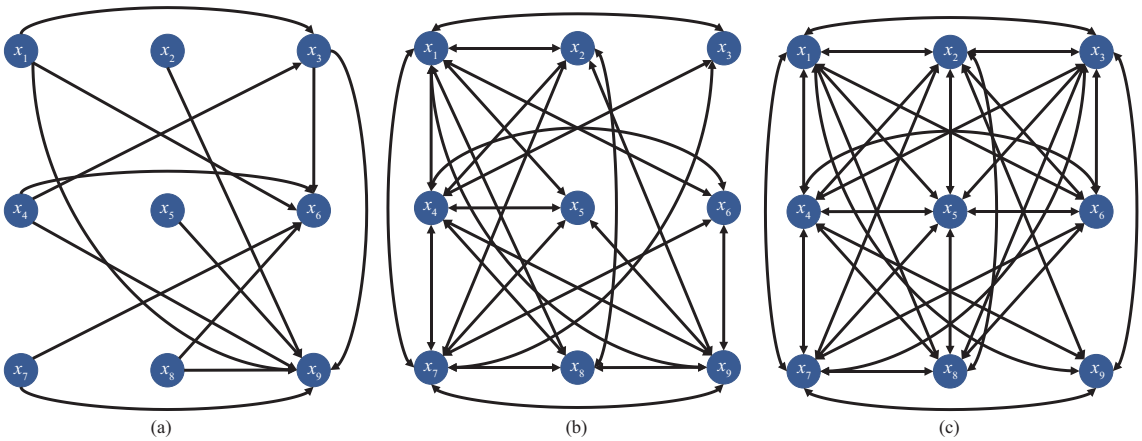

Figure 6 (Color online) Causal graph of our method on the Milan dataset. (a) "Internet"; (b) "Call"; (c) "SMS".

we explore the effect of different learning rates on the prediction results and illustrate them in Figures 11 and 12. Here, the learning rates are set to be 0.00005, 0.0005, 0.001, and 0.005, respectively. As seen from the figures, different learning rates greatly impact prediction performance. Also, different services correspond to different optimal learning rates. The optimal learning rate for the "SMS" and "Call" services in the Milan and Trentino datasets is 0.001, with corresponding RMSE values of 27.5620,

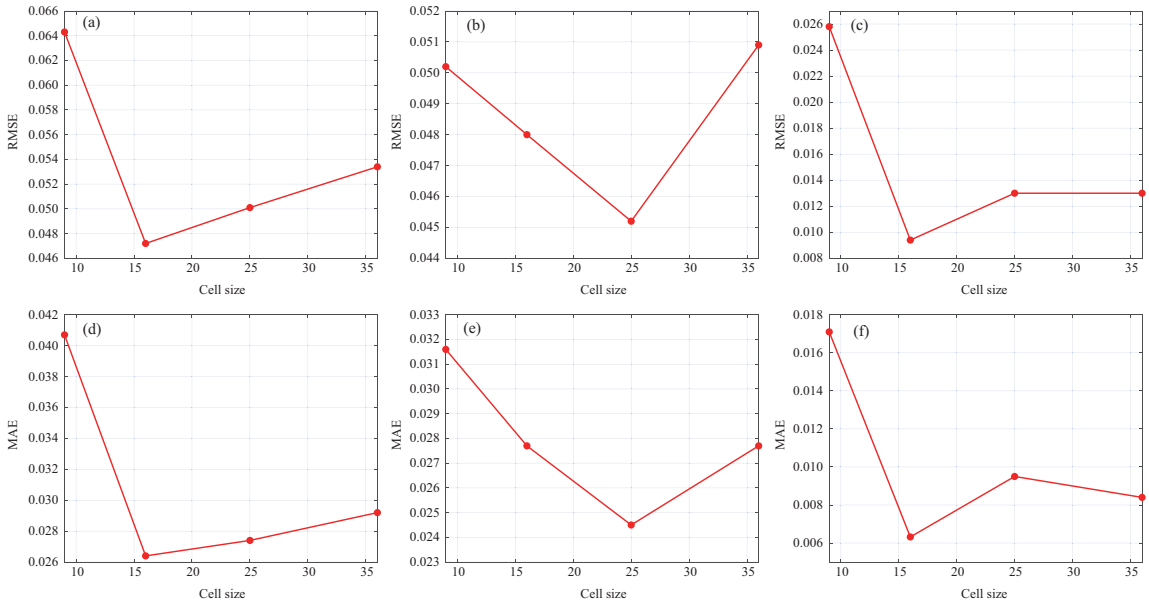

Figure 7 (Color online) Prediction performance of our method on the Milan dataset for different cell sizes. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different cell sizes. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different cell sizes.

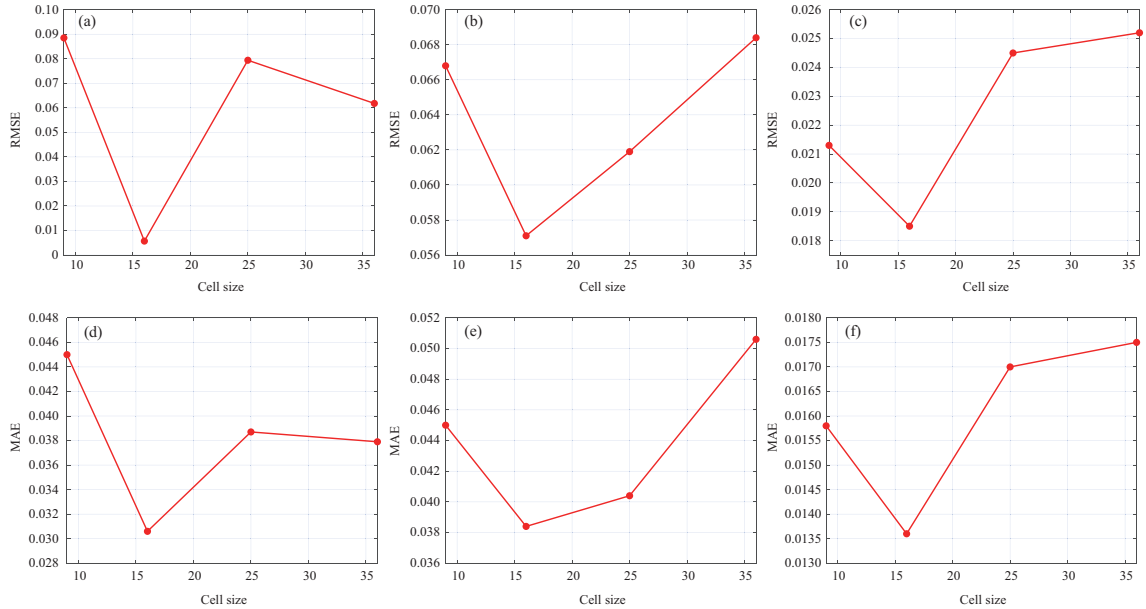

Figure 8 (Color online) Prediction performance of our method on the Trentino dataset for different cell sizes. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different cell sizes. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different cell sizes.

15.4701, 4.1189, and 19.3220, respectively. However, the optimal learning rate for the "Internet" service is 0.0005, with an RMSE of 98.6580 and 13.1588. The reason is that the causalities between "SMS" and "Call" services mainly focus on spatio-temporal features, while the "Internet" service contains potential causalities, such as the regional economy and population application preference. To mine these potential causalities, the model must be trained with a lower learning rate, allowing for optimal prediction results.

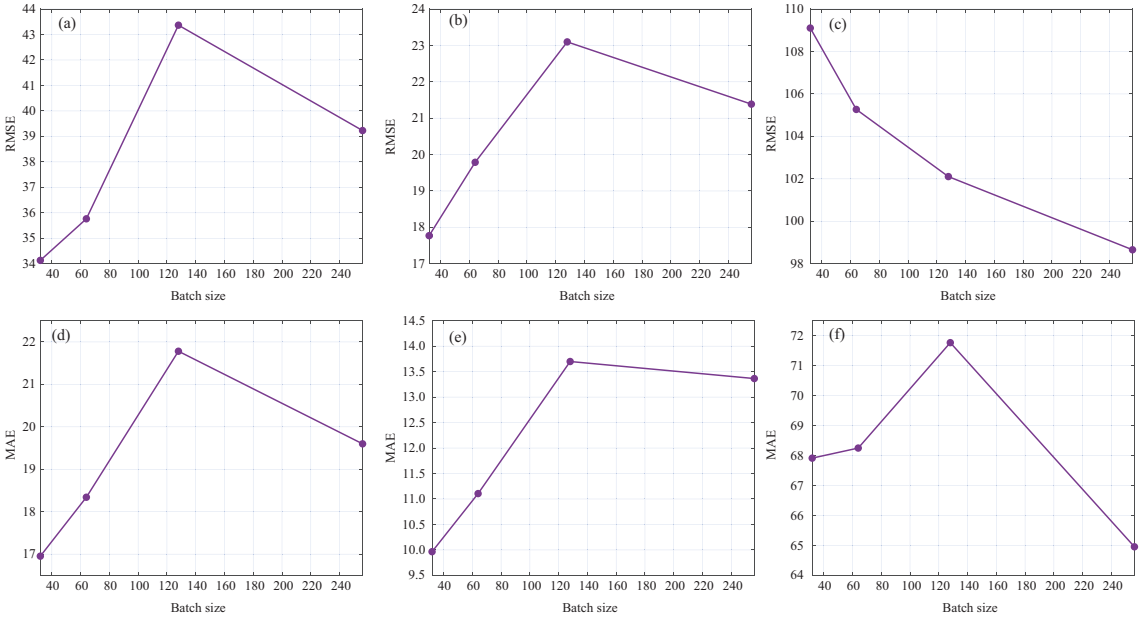

Figure 9 (Color online) Prediction performance of our method on the Milan dataset for different batch sizes. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different batch sizes. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different batch sizes.

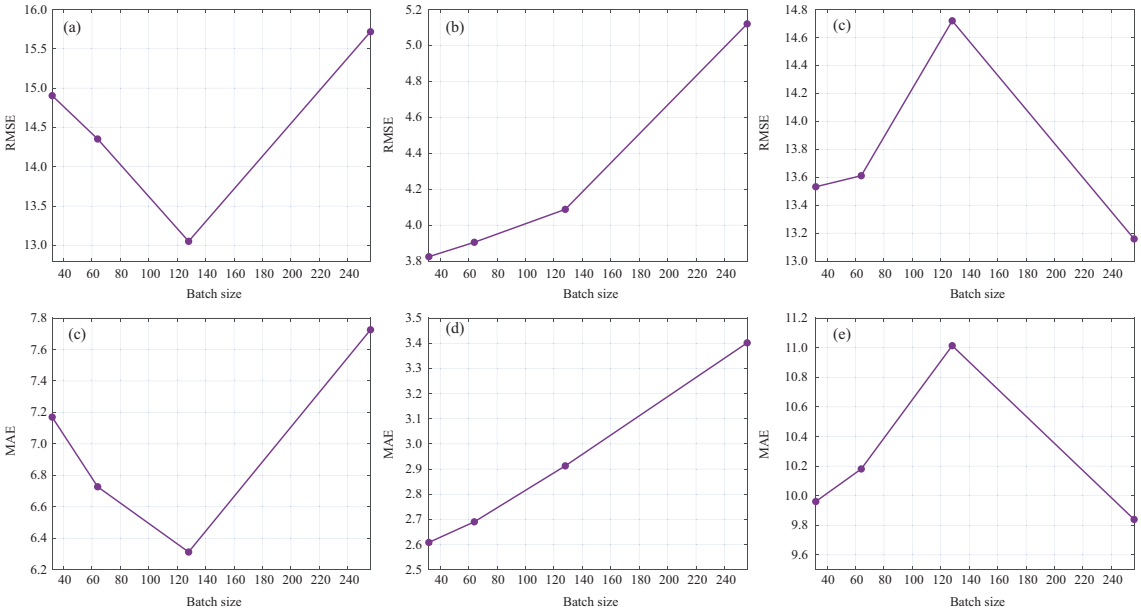

Figure 10 (Color online) Prediction performance of our method on the Trentino dataset for different batch sizes. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different batch sizes. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different batch sizes.

5 Conclusion

This paper proposed a traffic prediction method SSTCD based on causal discovery to solve the problem of end-to-end network traffic prediction. First, we adopt Granger causal discovery to mine the spatio-temporal causality on observed data to identify decisive prediction variables. Then, a GNN-based model is built to make causal predictions with spatio-temporal causality. To achieve the causal discovery without groundtruth causal graph in the observed data, we design a self-supervised manner to improve the accuracy of the traffic prediction task. Finally, we verify the effectiveness of the proposed method through

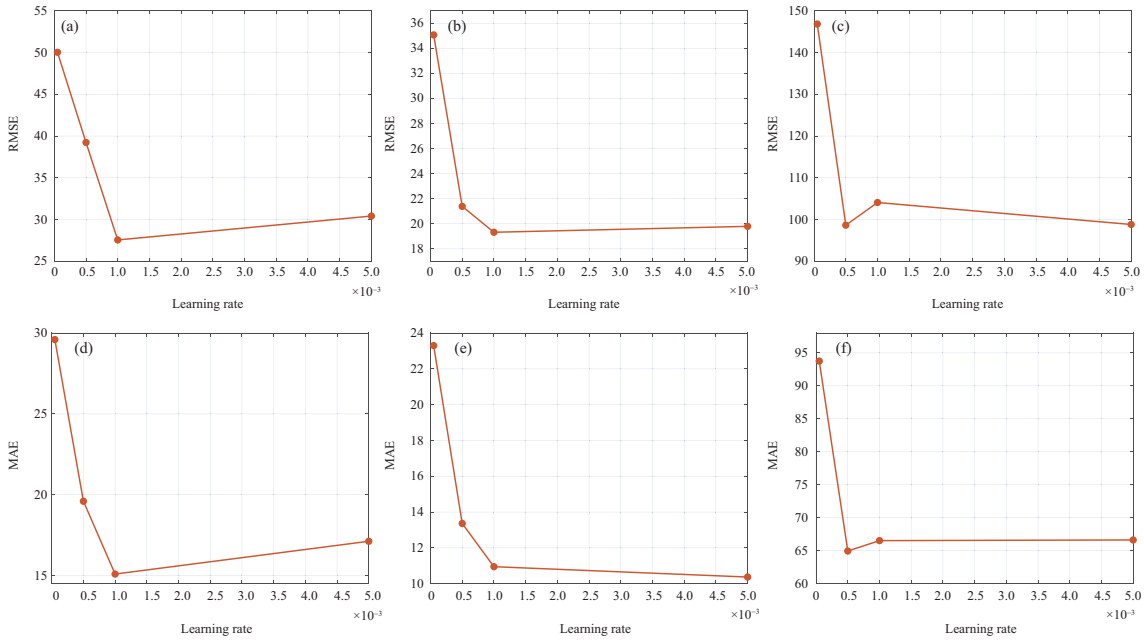

Figure 11 (Color online) Prediction performance of our method on the Milan dataset for different learning rates. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different learning rates. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different learning rates.

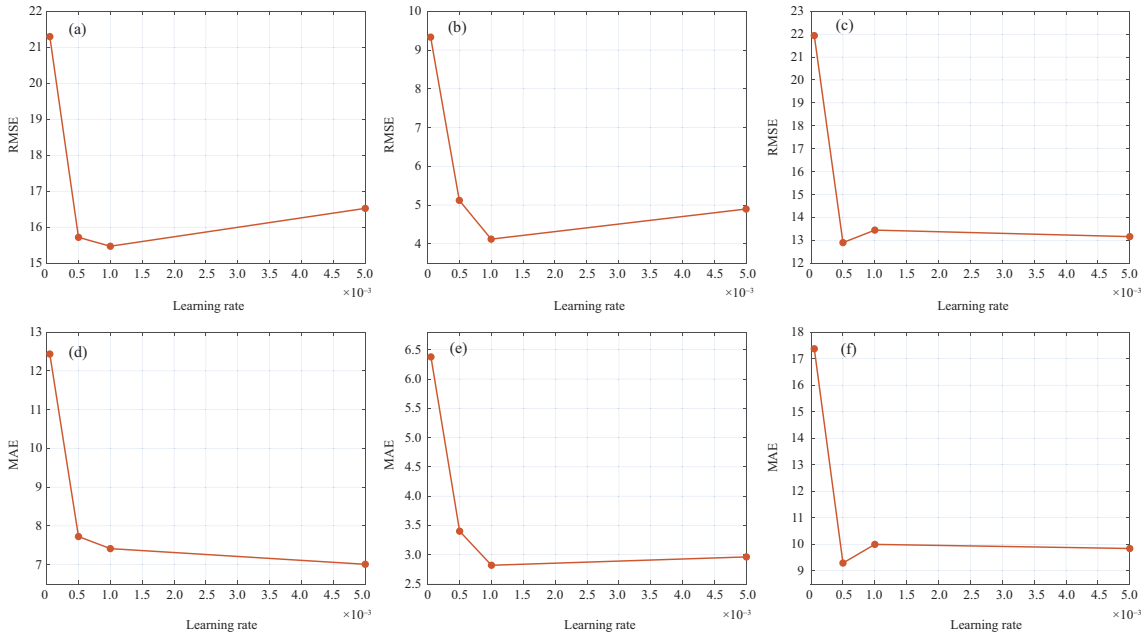

Figure 12 (Color online) Prediction performance of our method on the Trentino dataset for different learning rates. RMSE of our method on (a) "SMS", (b) "Call", and (c) "Internet" for different learning rates. MAE of our method on (d) "SMS", (e) "Call", and (f) "Internet" for different learning rates.

simulation results on two real datasets, which achieved 30.4% and 19.1% performance improvements on RMSE and MAE to other baseline methods, respectively.

Acknowledgements This work was supported by Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (Grant No. GML-KF-22-01), National Natural Science Foundation of China (Grant Nos. 62201419, 62372357), Key Research and Development Program of Shaanxi (Grant No. 2022ZDLGY05-08), and ISN State Key Laboratory.

References

1 Siriwardhana Y, Porambage P, Liyanage M, et al. A survey on mobile augmented reality with 5G mobile edge computing: architectures, applications, and technical aspects. IEEE Commun Surv Tut, 2021, 23: 1160–1192

2 Du J, Yu F R, Lu G, et al. MEC-assisted immersive VR video streaming over terahertz wireless networks: a deep reinforcement learning approach. IEEE Int Things J, 2020, 7: 9517–9529
3 Liberatore M J, Wagner W P. Virtual, mixed, and augmented reality: a systematic review for immersive systems research. Virtual Reality, 2021, 25: 773–799
4 Fang T, Yuan F, Ao L, et al. Joint task offloading, D2D pairing, and resource allocation in device-enhanced MEC: a potential game approach. IEEE Int Things J, 2022, 9: 3226–3237
5 Portilla-Figueras A, Llopis-Sánchez S, Jiménez-Fernández S, et al. Examining 5G technology-based applications for military communications. In: Proceedings of the European Symposium on Research in Computer Security, 2022. 449–465
6 Chu P, Zhang J A, Wang X X, et al. Semi-persistent V2X resource allocation with traffic prediction in two-tier cellular networks. In: Proceedings of the 89th Vehicular Technology Conference (VTC2019-Spring), 2019. 1–6
7 Zhou X, Zhang Y, Li Z, et al. Large-scale cellular traffic prediction based on graph convolutional networks with transfer learning. Neural Comput Appl, 2022, 34: 5549–5559
8 Zheng T H, Li B C. Poisoning attacks on deep learning based wireless traffic prediction. In: Proceedings of the IEEE Conference on Computer Communications, 2022. 660–669
9 Wang Y Q, Jiang D D, Huo L W, et al. A new traffic prediction algorithm to software defined networking. Mobile Netw Appl, 2021, 26: 716–725
10 Nie L S, Ning Z L, Obaidat M S, et al. A reinforcement learning-based network traffic prediction mechanism in intelligent internet of things. IEEE Trans Ind Inf, 2021, 17: 2169–2180
11 Li M, Wang Y W, Wang Z W, et al. A deep learning method based on an attention mechanism for wireless network traffic prediction. Ad Hoc Netw, 2020, 107: 102258
12 Xu H Y, Huang Y D, Duan Z H, et al. Multivariate time series forecasting based on causal inference with transfer entropy and graph neural network. 2020. ArXiv:2005.01185
13 Pearl J. Causal inference. In: Proceedings of Workshop on Causality: Objectives and Assessment at NIPS, 2010. 39–58
14 Nogueira A R, Pugnana A, Ruggieri S, et al. Methods and tools for causal discovery and causal inference. WIREs Data Min Knowl, 2022, 12:
15 Granger C W. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 1969, 37: 424–438
16 Box G E, Jenkins G M, Reinsel G C, et al. Time Series Analysis: Forecasting and Control. Hoboken: John Wiley & Sons, 2015
17 Zivot E, Wang J H. Vector autoregressive models for multivariate time series. In: Modeling Financial Time Series with S-PLUS®. Berlin: Springer, 2006. 385–429
18 Wang J, Tang J, Xu Z Y, et al. Spatiotemporal modeling and prediction in cellular networks: a big data enabled deep learning approach. In: Proceedings of the IEEE Conference on Computer Communications, 2017. 1–9
19 Li Z Y, Fu Y C, Zhao P C, et al. A dynamic spatiotemporal prediction method for urban network traffic. In: Proceedings of the 96th Vehicular Technology Conference (VTC2022-Fall), 2022. 1–5
20 Wang K, Li K L, Zhou L Q, et al. Multiple convolutional neural networks for multivariate time series prediction. Neurocomputing, 2019, 360: 107–119
21 Zhang C T, Zhang H X, Qiao J P, et al. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. IEEE J Sel Areas Commun, 2019, 37: 1389–1401
22 Zhang C T, Dang S P, Shihada B, et al. Dual attention-based federated learning for wireless traffic prediction. In: Proceedings of the IEEE Conference on Computer Communications, 2021. 1–10
23 Cheng L, Guo R C, Moraffah R, et al. Evaluation methods and measures for causal learning algorithms. IEEE Trans Artif Intell, 2022, 3: 924–943
24 Luo Y N, Peng J, Ma J Z. When causal inference meets deep learning. Nat Mach Intell, 2020, 2: 426–427
25 Nauta M, Bucur D, Seifert C. Causal discovery with attention-based convolutional neural networks. MAKE, 2019, 1: 312–340
26 Löwe S, Madras D, Zemel R, et al. Amortized causal discovery: learning to infer causal graphs from time-series data. In: Proceedings of the Conference on Causal Learning and Reasoning, 2022. 509–525
27 Wang Y X, Cao F Y, Yu K, et al. Local causal discovery in multiple manipulated datasets. IEEE Trans Neural Netw Learn Syst, 2023, 34: 7235–7247
28 Monti R P, Zhang K, Hyvärinen A. Causal discovery with general non-linear relationships using non-linear ICA. In: Proceedings of the Uncertainty in Artificial Intelligence, 2020. 186–195
29 Yu L X, Li M, Jin W Q, et al. STEP: a spatio-temporal fine-granular user traffic prediction system for cellular networks. IEEE Trans Mobile Comput, 2021, 20: 3453–3466
30 Tank A, Covert I, Foti N, et al. Neural Granger causality. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 4267–279
31 Barlacchi G, De Nadai M, Larcher R, et al. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. Sci Data, 2015, 2: 150055
32 Sun H Y, Liu H X, Xiao H, et al. Short term traffic forecasting using the local linear regression model. SN Appl Sci, 2002, 2: 1159
33 Sapankevych N, Sankar R. Time series prediction using support vector machines: a survey. IEEE Comput Intell Mag, 2009, 4: 24–38
34 Qiu C, Zhang Y Y, Feng Z Y, et al. Spatio-temporal wireless traffic prediction with recurrent neural network. IEEE Wireless Commun Lett, 2018, 7: 554–557
35 Wang X, Zhao J, Zhu L, et al. Adaptive multi-receptive field spatial-temporal graph convolutional network for traffic forecasting. In: Proceedings of the IEEE Global Communications Conference (GLOBECOM), 2021. 1–7
36 Yao Y, Gu B, Su Z, et al. MVSTGN: a multi-view spatial-temporal graph network for cellular traffic prediction. IEEE Trans Mobile Comput, 2023, 22: 2837–2849