Spatial Origin-Destination Flow Imputation Using Graph Convolutional Networks

Xin Yao, Yong Gao, Di Zhu, Ed Manley, Jiaoe Wang, and Yu Liu

***Abstract*— Due to the limitation of data collection techniques and privacy issues, the problem of missing spatial origin-destination flows frequently occurs. Data imputation provides great support for the acquisition of complete flow data, which enables us to better understand regional connections and mobility patterns. However, existing models or approaches neglect the network structure of spatial flows, thus resulting in inappropriate estimates and a low performance. The development of graph neural networks offers a powerful tool to deal with graph-structured data. In this article, we proposed a spatial interaction graph convolutional network model, which combines graph convolution and a mapping function to predict flow data from the perspective of network learning. This model utilizes geographical unit embedding in local spatial networks to improve prediction accuracy. A negative sampling technique is adopted to reduce misestimation. Experiments on Beijing taxi trip data verified the usefulness of our model in spatial flow prediction. We also demonstrated that a biased training sample had a negative impact on the model's performance. More attributes of geographical units, a more proper negative sampling rate and a larger training set can increase the prediction accuracy of flow data.**

***Index Terms*— Origin-destination flow, data imputation, spatial interaction network, graph embedding, graph convolution.**

I. Introduction

SPATIAL flow data, also known as origin-destination (OD) data or spatial interaction data, have been widely researched in urban planning, economics, tourism and many other fields [1]–[4]. Especially in intelligent transportation systems (ITSs), spatial flows are important for traffic dispatching optimization, transportation services planning and travel route management [5], [6], because they demonstrate regional connections and mobility patterns [7]–[9]. When aggregated into different regions, flows form a spatial interaction network, through which population, material and information transfer between regions. This provides vital insight into geographical phenomenon. In a spatially embedded network, the nodes represent geographical regions and the weights of edges are measured by flow volumes or intensities (e.g., the amount of movements). The nodes are spatially agglomerate, making it possible to discover spatial communities [10]–[12]. On the other hand, spatial interaction networks possess the characteristics of complex networks (e.g., scale free and small world), which helps to understand their structures as well as formation and evolution mechanisms [13], [14]. The distribution of spatial flows is tightly related to geospatial patterns [7], [15]. For example, a train station generates more flows than a subway station due to a higher human activity intensity. Flows to regions with favorable locations and traffic conditions are usually densely distributed and cover a large area. In urban studies, it is important to model the relationship between urban spatial structure and mobility flow distributions.

Limited by data collection techniques [16], [17] and privacy issues [18], it is common to encounter the problem of missing spatial flows. Data imputation is necessary to obtain relatively complete flow sets, which helps to grasp flow distributions and reveal their relations with spatial configurations. This also deepens the understanding of the geographical mobility patterns behind flow data. Existing models such as the gravity model and the radiation model are derived from physical laws or optimization theories [19]–[22]. They quantify flow volumes based on the heterogeneity and separation of geographical regions. However, neglecting the network structure of spatial flows may incur inaccuracy in spatial flow imputation. The estimation of flow intensity is also affected by the neighborhood structures of origin and destination regions. For instance, if two regions share more neighbors in a spatial interaction network, they are more likely to have close flow intensities to the same destination. Other factors pose a challenge to traditional models as well. For example, intra-urban movements largely depend on locational accessibility and land use types [15], and some regular trips are distance-insensitive. Besides, these models adopt simple solving methods such as linear regression and programming, which limits their capability of data imputation when dealing with complex flow distributions.

The emergence of artificial neural networks offers a black-box technique to depict complicated geospatial patterns [23], [24]. Since 1990s, many studies have applied neural networks to regional flow estimation [25]–[27]. Nevertheless,

Manuscript received July 8, 2019; revised April 19, 2020; accepted June 3, 2020. Date of publication July 1, 2020; date of current version November 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41971331, Grant 41830645, Grant 41625003, and Grant 41771425, in part by the National Key Research and Development Program of China under Grant 2017YFB0503602, and in part by the Smart Guangzhou Spatio-Temporal Information Cloud Platform Construction under Grant GZIT2016-A5-147. The Associate Editor for this article was M. Chowdhury. *(Corresponding author: Yong Gao.)*

Xin Yao, Yong Gao, Di Zhu, and Yu Liu are with the Institute of Remote Sensing and Geographical Information Systems, School of Earth and Space Sciences, Peking University, Beijing 100871, China (e-mail: yaoxin@pku.edu.cn; gaoyong@pku.edu.cn; patrick.zhu@pku.edu.cn; liuyu@urban.pku.edu.cn).

Ed Manley is with the School of Geography, University of Leeds, Leeds LS2 9JT, U.K. (e-mail: e.j.manley@leeds.ac.uk).

Jiaoe Wang is with the Institute of Geographic Sciences and Natural Resources Research, China Academy of Sciences, Beijing 100101, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: wangje@igsnrr.ac.cn).

Digital Object Identifier 10.1109/TITS.2020.3003310

they are in fact a gravity model because of the same inputs, and neural networks just improve their fitting ability. Recently, graph convolutional networks (GCNs) have earned their prevalence owing to the abilities of feature extraction and nonlinear fitting on arbitrarily graph-structured data [28], [29]. In geographical applications, relevant studies mainly focus on point-based forecasting [30], [31] and spatial pattern inference [32], [33] by constructing various types of spatial graphs. As an extension, relational graph convolutional networks (R-GCNs) [34] apply the GCN framework to knowledge bases and exhibit a good performance in link prediction. Such model presents great potential for spatial flow imputation by handling flows between geographical regions as a spatial interaction relation.

Inspired by R-GCNs, we proposed a spatial interaction graph convolutional network (SI-GCN) model to impute missing spatial flows in this research. The model comprises three main modules, including the spatial representation layer, the encoder and the decoder, and optimizes parameters with locational attributes and observed flows. Benefiting from graph convolution, our model can capture the network structure of flows and utilize geographical unit embedding to achieve a better performance than traditional models in flow imputation. The contributions are two-fold. First, we extended R-GCNs for spatial data imputation. Second, we introduced a network learning perspective to the prediction of spatial origin-destination flows.

II. Related Work

A. Origin-Destination Flow Prediction

Estimating spatial flow distributions has received much attention in the past few decades [19], [35], [36]. The gravity model, which is the most extensively used, suggests that flow intensity relies on both regional attributes and geographical separations [22], [35], [37]. Although the gravity model uses global parameters and ignores locational differences, it is concise to be understood. Gravity neural network models [25]–[27], [38], [39] combine the gravity model with three-layer artificial neural networks. Therefore, they use the same input variables as the gravity model. Neural networks increase the prediction accuracy to a certain extent. Simini *et al.* [20] proposed a radiation model according to the individual decision making of destinations. This model relates flow intensities to population distributions and is parameter-free. In addition, intervening opportunity model [40] and entropy-maximizing model [41] were also proposed. Spatial flows can be inferred from the change of point distributions as well. Zhu *et al.* [21] proposed a linear programming method to reproduce migration flows with the consecutive snapshots of population distributions, aiming to minimize transferring costs. The drawback is that real intensities cannot be obtained. Some studies estimate origin-destination flows with statistical methods. Models using Bayesian inference [42], [43] and Markov chains [44], [45] rely on defining a path-link incidence matrix and a transition probability matrix according to transport networks (e.g., road networks), respectively. These methods do not directly consider spatial interaction networks and overlook the positions of geographical units.

Other methods involve flow transformation between different geographical units. For instance, spatial interaction interpolation [46] reallocates flows between different zoning systems, but it is not suitable for sparse data.

B. Graph Convolutional Networks in Spatial Analysis

Different from raster data and remote sensing images, many spatial data are in graph structures such as road networks, spatial topological relations and mobility flows. Graph convolutional networks [28], [29], [34] provide a uniform framework to extract knowledge from such irregular datasets. In geographical analysis, GCNs are often used to model spatial dependency by building either distance or topological relations among geographical units, so that spatial prediction become possible. Yu *et al.* [30] and Zhang *et al.* [47] combined spatial graph convolution and temporal modeling on road graphs to forecast traffic data at monitor stations. Chai *et al.* [31] and Geng *et al.* [48] also adopted the GCN model to estimate shared-bike rent and return demands after constructing multi-graphs among bike stations or regions. It is demonstrated that the spatial structure captured by GCNs improves forecasting accuracy. Additionally, GCNs can be used to infer spatial patterns. Yan *et al.* [32] constructed graphs for building groups using the minimum spanning tree and the Delaunay triangulation, and distinguish whether buildings were in a regular arrangement. Zhu *et al*. [33] inferred the unknown properties of urban places in their connected contexts and evaluated the influence of different spatial interaction measures on their spatial predictability. These studies successfully applied GCNs to point-based spatial data, but it still remains a challenge to predict spatial flows since a flow involves two locations (i.e., an origin and a destination) and has a more complex structure.

III. Spatial Interaction Graph Convolutional Networks

A. Task and Assumption

Generally, a spatial interaction network is defined as a weighted directed graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{F})$, where $\boldsymbol{V} = \{v_i\}$ is a set of k geographical units (nodes) and $\boldsymbol{F} = \{f_{ij}\}$ is a set of spatial flows (edges), $i, j = 1, 2, \ldots, k$. Here v_i represents the i-th unit and can be grids [15], traffic analysis zones (TAZs) [49], streets [50] or any other units of analysis. f_{ij} is the flow from v_i to v_j.

The task of spatial flow data imputation is to predict missing flows based on $\boldsymbol{V}$ and an observed flow set $\boldsymbol{F}_{\mathrm{p}} \subseteq \boldsymbol{F}$. The flows in $\boldsymbol{F}_{\mathrm{p}}$ are called positive samples. As flow estimation can be regarded as quantitative spatial reasoning, R-GCNs developed for statistical relational learning provide a useful tool. In R-GCNs, every node updates its state according to its relationship with neighboring nodes. This process is repeated for several times and the final node states are used for various tasks, such as node classification and link prediction. However, two limitations cause a problem that R-GCNs cannot be directly applied to spatial flows. First, they have not yet employed node features [34]. The spatial and non-spatial attributes of geographical units are critical elements

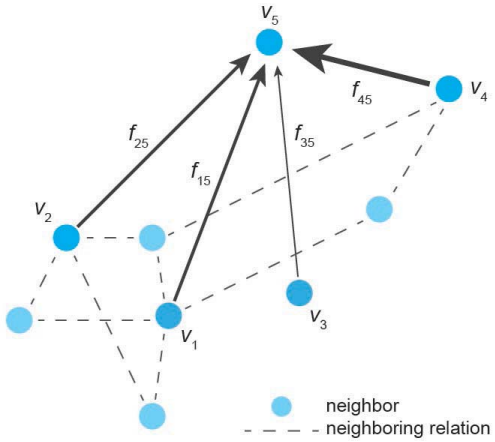


Fig. 1. Flows among different geographical units. For simplicity, V_5 is the same destination.

of flow prediction [19], [22]. Second, R-GCNs are designed to judge whether a link exists or not. It makes less sense for spatial flow prediction compared to calculating specific intensities. Therefore, we proposed a spatial interaction graph convolutional network (SI-GCN) model based on R-GCNs.

Our model is based on the following assumption: Given two flows f_{ij} and f_{mn}, if the origins v_i and v_m are similar and so do the destinations v_j and v_n, the two flow intensities should be approximately equivalent. Specifically, we considered two similarities:

- The **first-order similarity** measures the closeness of attributes between geographical units.
- The **second-order similarity** measures the neighborhood structural proximity between geographical units in a spatial interaction network.

Fig. 1 illustrates a simple case. Because v_1 and v_2 share many common neighbors (second-order similarity) and are in spatial proximity to each other (first-order similarity), f_{25} is similar to f_{15}. Therefore, the intensity of f_{15} can be inferred with f_{25}, or vice versa. On the contrary, v_3 shares no neighbors with v_1 and v_4 is far away from v_1, so f_{35} and f_{45} may differ from f_{15}. Graph convolution is effective to generate similar representation vectors for similar geographical units regarding both their attributes and network topology, then we can estimate flow intensities with these representation vectors.

B. Model Architecture

As Fig. 2(a) shows, the SI-GCN model consists of three main parts. The spatial representation layer is responsible for data organization and processing. The encoder generates latent representations for all geographical units with graph convolution. As a result, similar units get similar representations, which are then utilized by the decoder to predict missing flows. Compared with R-GCNs, the proposed model is feasible to use any type and number of attributes of geographical units for spatial flow prediction. For instance, by inputting positional information, the model indirectly introduces travel costs and the distance decay effect [51], [52] into the topological structure of spatial interaction networks. Additionally, the training objective is also changed. Rather than assign scores as large as possible on observed links to indicate the possibility of their existence, the SI-GCN model makes flow estimates approach real values exactly. The details of the model's implementation are described as follows.

The inputs of the model include geographical units $\boldsymbol{V}$ and observations $\boldsymbol{F}_{\mathrm{p}}$. The spatial representation layer works on three tasks. First, it organizes the attributes of geographical units as numeric node feature vectors, so that every unit becomes computable. Nominal and ordinal attributes will be transformed into numeric values using one-hot encoding. These input attributes determine the first-order similarity between nodes, but do not contain network structural information. Second, a spatial subgraph is constructed based on $\boldsymbol{F}_{\mathrm{p}}$. It enables the model to capture the observed structure of a spatial interaction network, and thus graph convolution can be implemented. Third, this layer also conducts negative sampling [53], [54] before every training process of the model to reduce overestimation. It samples a set of negative (fake) flows $\boldsymbol{F}_{\mathrm{n}}$ by randomly replacing the origin or destination of positive examples. For instance, given a real flow f_{ij} with the origin v_i and the destination v_j, we change either v_i or v_j to another random unit v_m and get the fake flow f_{mj} or f_{im}, respectively. A negative flow is rejected if it appears in $\boldsymbol{F}_{\mathrm{p}}$ by chance. The intensity of negative flows is set to zero, indicating that the random combination of an origin and a destination hardly produce flows. Training with both $\boldsymbol{F}_{\mathrm{p}}$ and $\boldsymbol{F}_{\mathrm{n}}$ forces the model to not only well predict real flows, but also distinguish them from fake ones, and thus improves the model's performance. Note that $\boldsymbol{F}_{\mathrm{n}}$ is not involved in the spatial subgraph construction.

The encoder is, in essence, a graph embedding [55], [56] with L graph convolutional layers. The convolution process in every layer contains two steps: For each geographical unit, the message passing step collects neighboring node representations (i.e., states), and the state updating step calculates its new representation vector, as Fig. 2(b) shows. Eq. (1) formulates the two steps uniformly:

$$\boldsymbol{z}_i^{l+1} = \sigma\left(k_i \sum_{j \in \boldsymbol{N}_i} \boldsymbol{W}^l \boldsymbol{z}_j^l\right) \tag{1}$$

where $\boldsymbol{z}_j^l$ is the hidden representation of v_j in the l-th layer, $l = 1, 2, \ldots, L$. $\boldsymbol{z}_j^0$ is the input feature vector in the first layer and $\boldsymbol{z}_j^L$ is the output of the encoder in the last layer and also called the embedding vector. $\boldsymbol{W}^l$ is a layer-specific weight matrix. k_i is a normalization constant. $\boldsymbol{N}_i$ is the index set of the neighbors of v_i (i.e., the geographical units interacting with v_i). We add i to $\boldsymbol{N}_i$ so that self-state is considered as well. Finally, the state of v_i is updated through an activation function such as ReLu: $\sigma(x) = \max(0, x)$. For every node, graph convolution integrates its first-order and second-order characteristics and maps them to a latent representation vector. Similar nodes are close to each other in the latent vector space. The SI-GCN model can deal with large-scale spatial interaction networks and extract useful patterns, because it transforms complex flow distributions into low-dimensional node vectors.

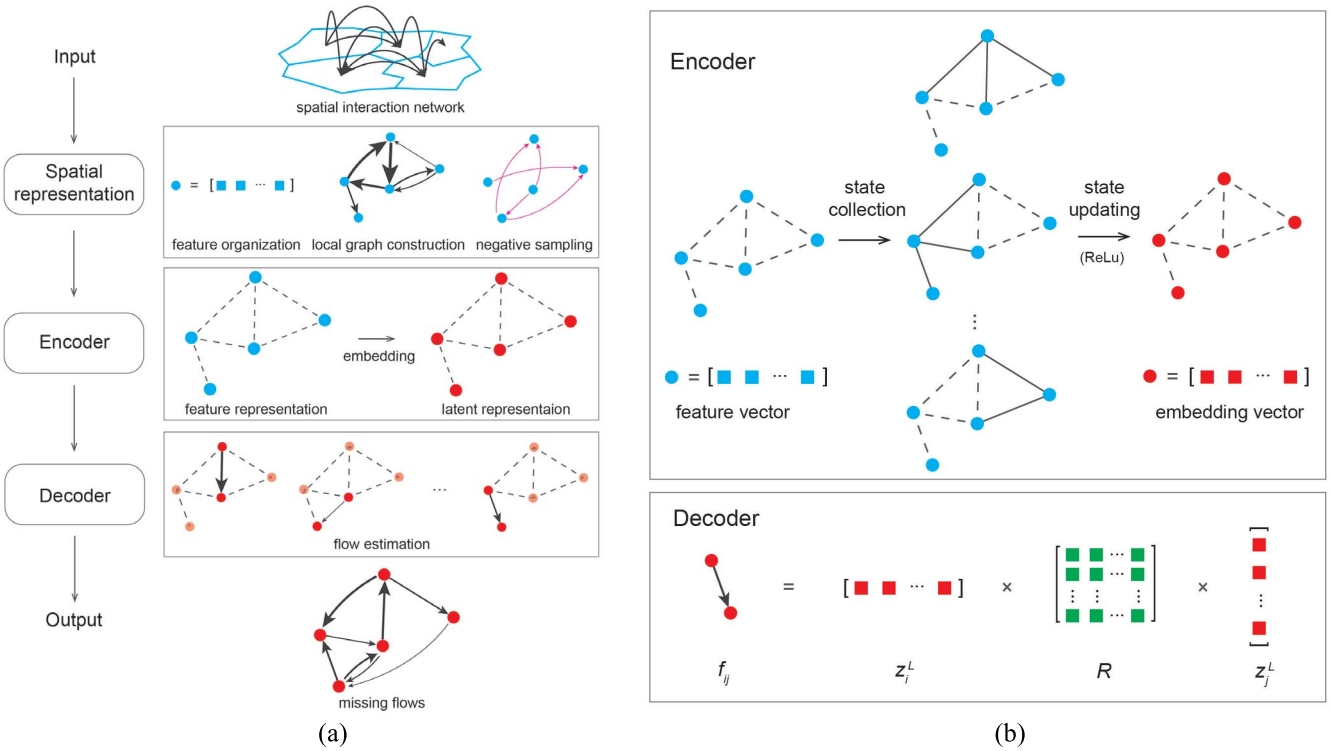


Fig. 2. The SI-GCN model. (a) Model architecture. (b) The encoder using one graph convolution layer and the decoder using a bilinear transformation. Circles represent geographical units and squares represent elements in a vector/matrix.

The decoder calculates flow intensity with the embedding vectors of geographical units. It uses a bilinear transformation [57], which associates the spatial interaction relation with a parameter matrix $\boldsymbol{R}$, to estimate flow intensity f_{ij} from v_i to v_j, as Eq. (2) shows:

$$f_{ij} = (z_i^L)^T \boldsymbol{R} z_j^L \tag{2}$$

The mark T denotes transpose. This operation maps two embedding vectors of geographical units to a numeric value, namely, the flow intensity. As shown in Fig. 2(b), each row of $\boldsymbol{R}$ relates to a dimension of the embedding vectors and determines the importance of the corresponding latent feature for flow prediction. Other mapping functions apply to the decoder as well [54].

In a training procedure, the full-batch gradient descent optimization is adopted to tune the model's parameters. It means all training flows are involved in the training procedure in every iteration. The objective is to make the estimates of positive flows approximate ground truth exactly and those of negative instances close to zero as much as possible. We use mean squared error (MSE) for loss computation, as denoted by Eq. (3). f_{ij}^* is the true values and N is the number of flows.

$$L = \frac{1}{N} \sum_{f_{ij} \in \boldsymbol{F}_\mathrm{p} \cup \boldsymbol{F}_\mathrm{n}} (f_{ij} - f_{ij}^*)^2 \tag{3}$$

IV. Evaluation

A. Dataset and Model Settings

We applied our model to taxi trip data at the extent of the Fifth Ring Road of Beijing, China. The dataset contains 1,115,132 trips during five workdays (May 13th to 17th, 2013) from 17,397 taxis, more than a quarter of all taxis in the city. Taxi trips from car-hailing apps/platforms were also included. Each trip records pick-up and drop-off locations. We partitioned the study area into grids with a side length of 1 km as geographical units (30 rows and 30 columns), which have been widely utilized in urban studies [58], [59]. Then the taxi trips were aggregated into these grids to obtain taxi flows at the grid level. Flow intensity was measured by the number of taxi trips originating and ending in two grids. For instance, if there are n trips from one grid to another during the five days, the corresponding flow intensity is n. Note that we ignored the temporal variation of flows because this research focuses on spatial data imputation. Although we collected taxi trips over several days, only one flow from a grid to another was generated during this period. Considering that a relatively low flow intensity indicated individual behaviors and was not a stable travel pattern, we removed the flows with taxi trips less than 30. The grids and the taxi flows are shown in Fig. 3.

Various attributes of geographical units have an impact on flow distributions, such as position, population and land-use type [15], [37], [46]. Usually, the more attributes are available, the more accurate flow prediction will be. For simplicity, we provided three attributes for the SI-GCN model, including position, propulsiveness and attractiveness, which are represented by the centroid coordinates, the number of pick-ups and the number of drop-offs of grids, respectively. These attributes are often used in traditional models [19], [51], [60].

To evaluate the proposed model, we randomly divided the taxi flow data into three parts as training (60%), validation

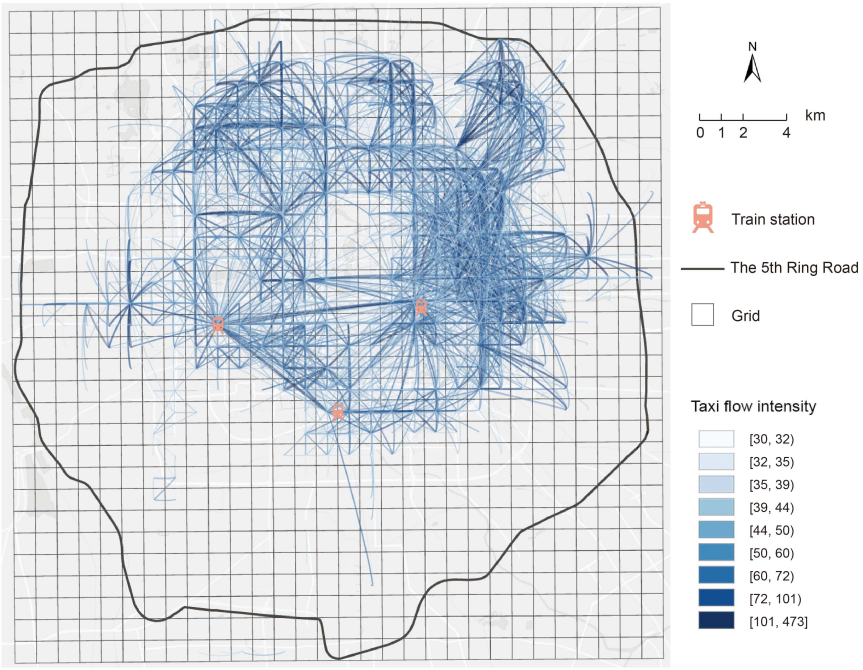


Fig. 3. The grids and taxi flow network in Beijing.

(20%) and test (20%) sets. The training and test sets represented observed flows (i.e., $\boldsymbol{F}_{\mathrm{p}}$) and missing flows, respectively. The validation set evaluated the generalization ability of the model in a training process. A flow can appear in only one of the three sets. The negative sample $\boldsymbol{F}_{\mathrm{n}}$ was 25% the size of the training set. The learning rate, the dimension of embedding vectors and the number of iterations were set to 0.005, 500 and 40,000, respectively. Since many graph convolutional layers would cause over-smoothing problems [61], we adopted one layer for geographical unit embedding.

B. Baselines and Metrics

Three well-known mobility models were selected for comparison, which all utilized the same inputs to predict missing flows.

Gravity model (GM) [22], [35] suggests that the flow intensity f_{ij} between two regions v_i and v_j is positively associated with v_i's propulsiveness p_i and v_j's attractiveness a_j, while the travel cost d_{ij} between them has an inverse effect. Eq. (4) shows the original form (**GM_O**) of the model:

$$f_{ij} = K\frac{p_i a_j}{d_{ij}^{\beta}} \tag{4}$$

where β is a distance decay coefficient and K is a scale constant. The travel cost between two grids is usually measured by the Euclidean distance between their centroids. As Eq. (5) shows, linear regression is used to estimate the parameters after taking a logarithmic transformation of Eq. (4). We obtained β and K with the training set and predicted flows in the test set.

$$\log\frac{f_{ij}}{p_i a_j} = \log K - \beta \log d_{ij} \tag{5}$$

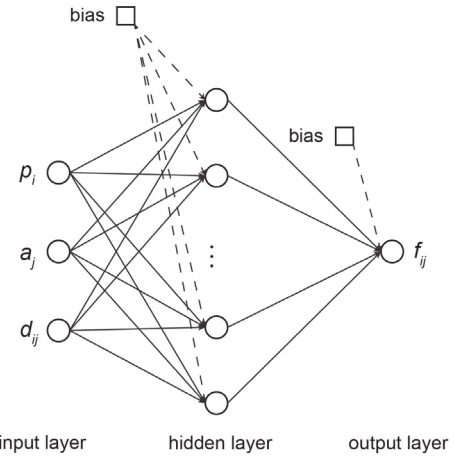


Fig. 4. The structure of the gravity neural network model.

In addition, an improved version of the gravity model was also considered, which replaced p_i and a_j with p_i^{α} and a_j^{γ}, respectively. The power exponents α and γ can be estimated using linear regression as well. We denoted this model as **GM_P**.

Gravity neural network (GNN) [26], [38] is a fully-connected feedforward network consisting of one input layer, one hidden layer and one output layer (Fig. 4). The input layer contains three neurons, corresponding to p_i, a_i and d_{ij} in the gravity model. The output layer produces the prediction f_{ij}. The hidden layer contains several neurons. We used three GNN models that had 10, 20 and 30 hidden neurons, denoted by **GNN_10**, **GNN_20** and **GNN_30**, respectively. They were trained independently with the error back propagating and

stochastic gradient descent techniques. They adopted the same loss function, learning rate and the number of iterations as SI-GCN.

Radiation model(RM) [20] assumes that the flow from v_i to v_j depends on their population s_i and s_j and the population s_{ij} of a circular region around v_i whose radius is the distance between v_i and v_j, as Eq. (6) shows:

$$f_{ij} = T_i \frac{s_i s_j}{(s_i + s_{ij})(s_i + s_j + s_{ij})} \tag{6}$$

T_i indicates the total flows originating at v_i. Considering that this model is parameter-free, we directly used it for flow prediction. To keep consistent with other models, the sum of pick-ups and drop-offs was adopted as the population of a grid.

We adopted four metrics as performance measurements. Root mean squared error (RMSE) and mean absolute percentage error (MAPE) describe how much flow estimates deviate from ground truth. Spearman correlation coefficient (SCC) reflects the change consistency between estimates and ground truth. The common part of commuters (CPC) [62] is a similarity measurement and is formulated as:

$$CPC = \frac{2\sum_{i,j} \min(f_{ij}, f^*_{ij})}{\sum_{i,j} f_{ij} + \sum_{i,j} f^*_{ij}} \tag{7}$$

It denotes the extent that estimates match ground truth on the whole. $CPC = 1$ indicates that they are equivalent.

C. Results

We repeated the experiment for ten times. Before each experiment we resampled the training, validation and test sets randomly and all models used the same sets. In each experiment we ran GNNs and SI-GCN for five times and reported their average performances, and ran GMs and RM for one time because they have stable analytical solutions.

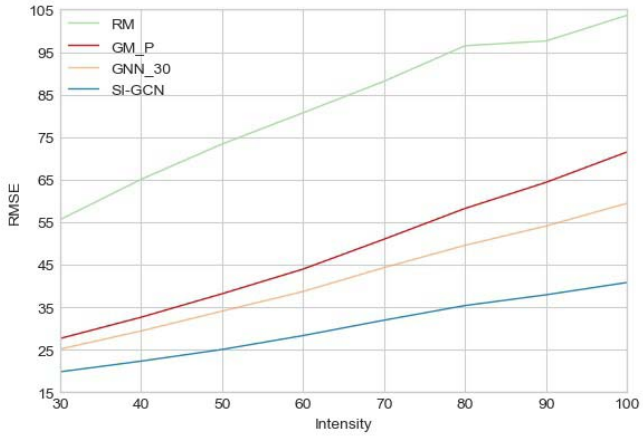
Table I shows the average performance of every model regarding all ten experiments. For GM_O, β and K are 1.38 and 7.41e-05, respectively. With respect to GM_P, we obtained the parameters $(K, \alpha, \gamma, \beta) = (4.19, 0.21, 0.2, 0.55)$. It can be discovered that the addition of exponent parameters does increase the goodness-of-fit of the gravity model. RM is situated between GM_O and GM_P in the performance of flow prediction. Compared with traditional models, neural network based methods can produce better estimates. For GNNs, the incremental hidden neurons do not significantly improve the accuracy. Our model achieves the best performance regarding all four metrics. To be specific, it outperforms GNN_30, the second-best model, by 21.3% and 12.3% in terms of RMSE and MAPE, respectively. SCC has increased 8.0% while CPC also rises slightly. These improvements validate the effectiveness of the proposed model.

We further analyzed the prediction of test flows at different intensity and distance levels. At first, we computed the RMSE of taxi flows in the test set with intensities larger than a given value. As Fig. 5(a) shows, all the curves display a rising trend, indicating that more errors exist in the prediction of large-intensity flows. Among these models, SI-GCN produces the least errors at any intensity. As for distance levels, we divided the test data into five classes using the Jenks natural breaks optimization [63], representing taxi flows in different distances, as Fig. 5(b) shows. In general, SI-GCN achieves the highest accuracy. For flows in the long-distance class, GNN_30 performs a little better

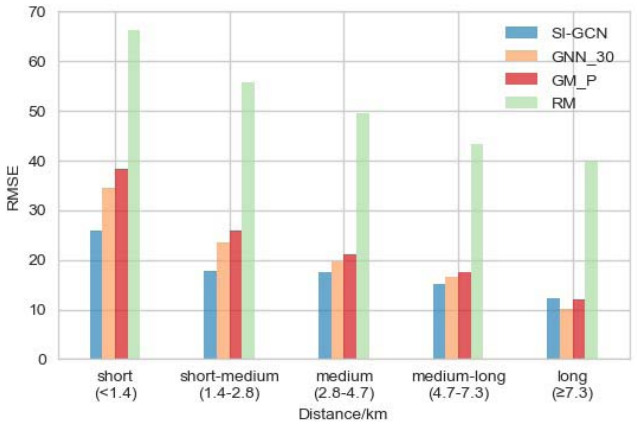
TABLE I
PREDICTION RESULTS OF TEST DATA USING DIFFERENT MODELS

Metrics / Models	RMSE	MAPE	SCC	CPC
GM_O	161.581	90.1%	0.568*	0.603
GM_P	27.592	27.7%	0.643*	0.848
RM	55.547	83.5%	0.624*	0.421
GNN_10	25.227	27.8%	0.648*	0.860
GNN_20	25.174	27.8%	0.649*	0.860
GNN_30	25.082	27.7%	0.651*	0.861
SI-GCN	**19.727**	**24.3%**	**0.703***	**0.885**

* p-value < 0.001



(a)



(b)

Fig. 5. The prediction errors of different flows. (a) Flows in various intensities. (b) Flows in various distances.

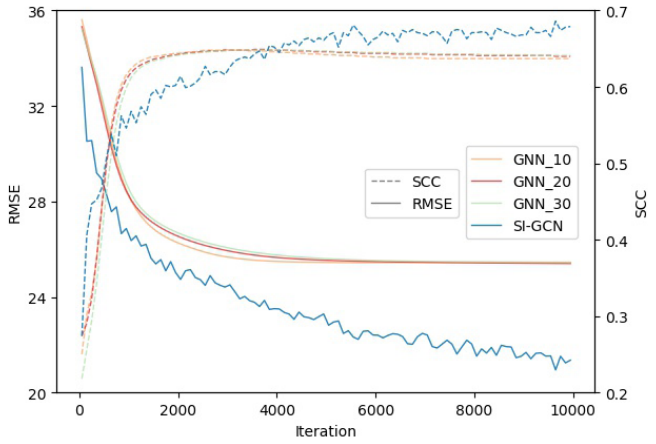


Fig. 6. The performances of neural models during their training processes.

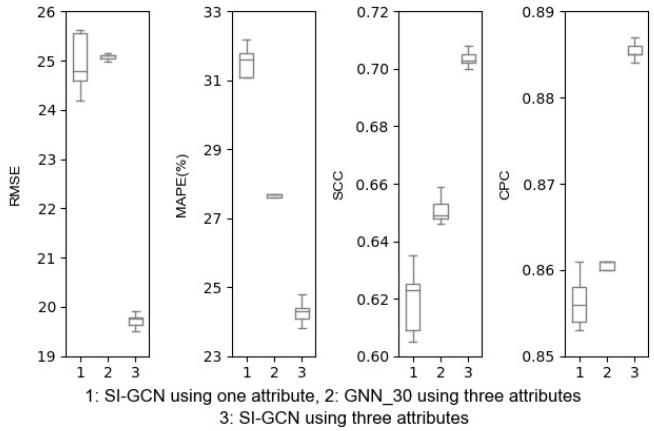


Fig. 7. Prediction results using different numbers of geographical unit attributes.

than SI-GCN, which may denote that the gravity model is suitable for large-scale systems [64]. Furthermore, due to the distance decay effect [51], [52], long-distance flows are generally in lower intensity and much less than shorter ones. Specifically, the average intensities of the five classes from short to long are 77.9, 60.4, 49.4, 41.9 and 39.7. Because the prediction deviations of low values are small (cf., Fig. 5(a)), all model produce the lowest prediction errors for long-distance flows.

Furthermore, we assessed the training performance of all neural models. Fig. 6 displays the RMSE and SCC change of test flows during the first 10,000 iterations of one experiment. In general, SI-GCN has the best learning efficiency than others because the corresponding RMSE curve descends faster than others, and the SCC curve exceeds those of GNNs after certain iterations. Moreover, since GNNs are in simple structure with fewer parameters and can continuously approach local optimum, their curves are much smoother.

V. Discussions

A. Flow Prediction With Limited Node Attributes

In some applications, it is difficult to obtain many attributes of geographical units. Positional information is often the most easily available in spatial analysis. To test SI-GCN's performance with limited node attributes, we repeated the Beijing taxi flow experiment in Section IV but only used centroid coordinates as the attribute inputs of grids. Fig. 7 shows that more node attributes, which help the model to better measure the similarity between geographical units, can improve the prediction accuracy. In addition, by comparing the boxplots we can find that a model using more input attributes can produce more stable estimates.

B. The Impact of Training Data

As a machine learning method, the SI-GCN model requires a training set to optimize its parameters iteratively. Therefore, the quality of training data determines the model's prediction ability. In this section, we investigated the impact of various training data on spatial flow prediction. Data and model settings remained the same as those in Section IV if not mentioned otherwise.

1) Taxi Trip Threshold: The experiment in Section IV kept taxi flows that had at least 30 trips between two grids, and they presented a distinctive taxi travel pattern. We also set different thresholds and used the corresponding data for model evaluation. Fig. 8(a) shows that SI-GCN outperforms other models given any flow distribution. Besides, RMSEs get larger with the increase of the threshold, as more low-intensity flows are removed.

2) Training Set Size: If training flows ($\boldsymbol{F}_{\mathrm{p}}$) form a totally disconnected graph, where any two flows do not share origin and destination locations (i.e., any geographical unit does not have neighbors), the model would fail in capturing network topological characteristics. However, such situation is unlikely to appear in practice unless flows are too sparse to train a model. We sampled the taxi flow data in proportions varying from 20% to 80% at an interval of 20% as different training sets to train SI-GCN. The corresponding test set contains half of the remaining flows. As Fig. 8(b) shows, the model predicts flows more accurately with a larger training set. At the same time, the shorter error bars indicate a more stable performance. This is because more flows form a relatively complete graph structure that contains more information about their spatial distribution.

3) Negative Sampling Rate: When we train SI-GCN with positive flows in a training set, the model tends to predict them accurately as much as possible. However, flow data are usually sparse (cf., Fig. 3), resulting in the problem of overfitting. Negative sampling promotes a better model because it provides more data in the training process and improves the generalization ability of SI-GCN. As shown in Fig. 8(c), we explored eleven negative sampling rates. 0 means no negative instance while 100% denotes that $\boldsymbol{F}_{\mathrm{n}}$ has the same size as $\boldsymbol{F}_{\mathrm{p}}$. The SCC curve moves up and then fluctuates after 0.7. The RMSE curve descends before the rate 30%, indicating that negative sampling improves the model's performance. It is interesting to find that errors become larger as the sampling rate continues increasing. The reason is that the model can only ensure that negative flows do not appear in $\boldsymbol{F}_{\mathrm{p}}$ but cannot

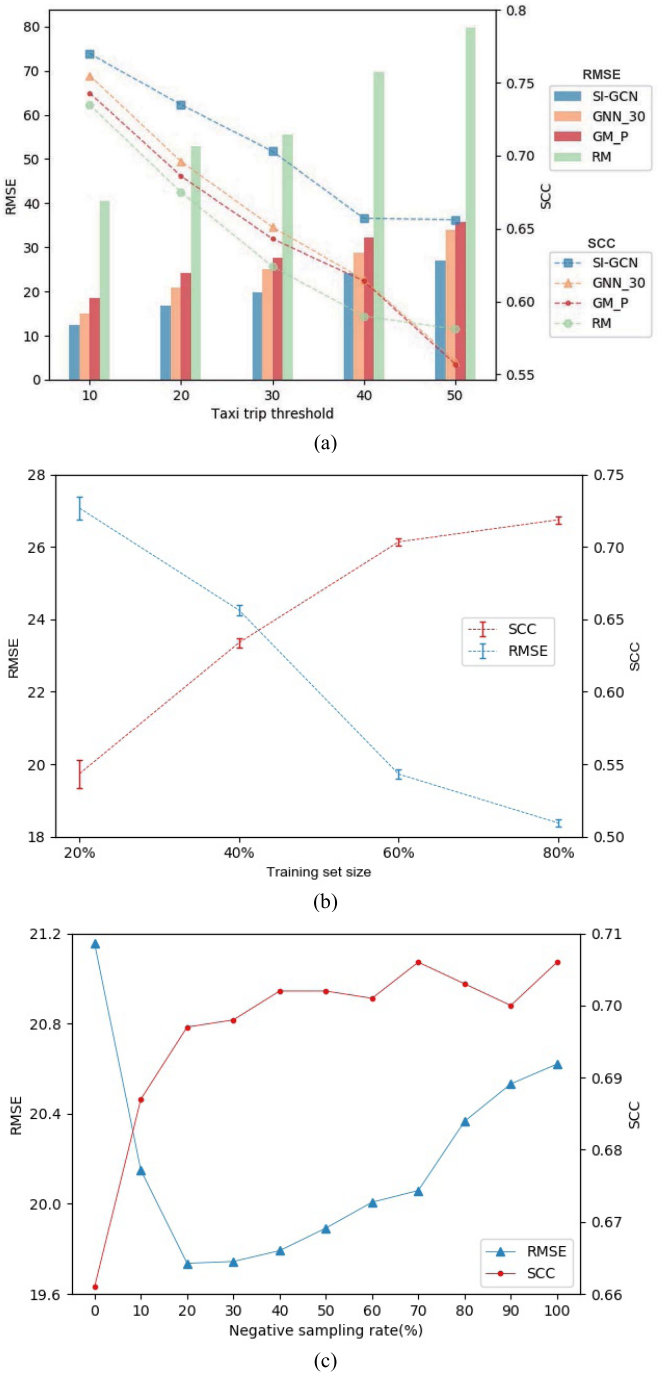


Fig. 8. The impacts of different training data on flow prediction. (a) Taxi trip threshold. (b) Training set size. (c) Negative sampling rate.

judge whether the test set contains them. The more negative flows it generates, the higher probability these flows appear in the test set. A test flow will be underestimated once treated as a negative flow whose intensity is zero. Therefore, it is unnecessary to set a large negative sampling rate.

4) Training Flow Intensity: Apart from random sampling, we carried out another two experiments that sampled training flows based on their intensity. In the first experiment, taxi flows with higher intensities were more likely to be selected as training data, and thus the corresponding test set would

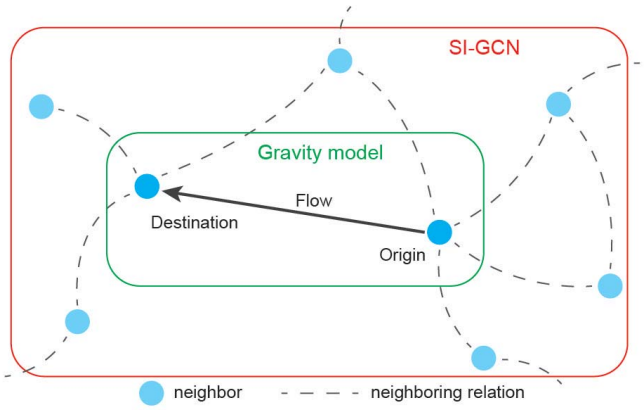


Fig. 9. The difference between the gravity model and the SI-GCN model.

TABLE II
THE IMPACT OF TRAINING FLOWS IN DIFFERENT INTENSITIES

Metrics / Training flows	RMSE	MAPE	SCC	CPC
High intensity	18.618	29.4%	0.516*	0.870
Low intensity	37.747	26.6%	0.705*	0.835

* p-value < 0.001

contain more low-intensity flows. The second experiment was to the contrary. Table II shows that compared with the random training sample (cf., the bottom row of Table I), biased samples generally lead to a worse performance of SI-GCN. Note that comparing the MAPEs rather than the RMSEs makes more sense, because the absolute prediction deviation of high-intensity flows is usually greater than that of low-intensity ones.

C. Predictability of Origin-Destination Flows

The distribution of flow data is complex and presents three patterns. The first is an overall trend relevant to spatial scales. Traditional models concentrate on explaining this mechanism using a general function, e.g., for the gravity model, the distance decay coefficient of intra-urban flows is larger than that of inter-urban flows [65], [66]. The second is regional heterogeneity. Some research [67], [68] has contributed to localizing flow models to exhibit regional difference. The last is unpredictable randomness related to individual behaviors. In Section IV, we removed the taxi flows containing less than 30 trips, i.e., six trips per day on average, to reduce the impact of randomness as much as possible. In practice, it is impossible to get exact predictions, and an upper bound exists in the accuracy of flow data imputation.

Regional flow intensity depends on many factors, which can be divided into two types. The first is endogenous factors that determine the basic intensity of regional flows, such as the position, population and land-use of geographical units. Their distributions are tightly related to geospatial patterns. The second is exogenous factors that cause the fluctuation of flow intensity, such as weather conditions, holidays and

special events. They can be used to forecast regional flows in a time series model. As with classical mobility models (e.g., the gravity model and the radiation model), this research concentrates on the endogenous factors using a cross-sectional dataset.

Essentially the gravity model and the SI-GCN model predict flows in the same way that both are functions of the attributes of geographical units. The former directly uses original attributes while the latter firstly generates node embeddings in a spatially interaction network configuration. Fig. 8 illustrates the difference between the two models. The gravity model is only concerned with origin and destination locations themselves. The limitation is that not all trips simply follow a power law against travel costs, especially at an urban scale. The SI-GCN model considers the neighborhood structure of geographical units. However, even with the state-of-the-art graph convolution technique, prediction errors still exist. As Table I shows, RMSE and MAPE cannot fall below 10 while SCC and CPC cannot reach 1. There are two ways to further increase flow prediction accuracy. One is to improve existing models or to develop new models, and the other is to provide extra locational attributes.

D. Applications of the SI-GCN Model

Based on modeling the relationship between flow distributions and geospatial patterns, the application of the SI-GCN model lies in three aspects. The first is to predict missing flows. For instance, due to navigation equipment malfunction or data management issues, we may loss flow data or get invalid data among some regions. Sometimes we can only get an incomplete flow sample. In these situations, the proposed model can be used for data imputation. The second application is to evaluate the spatial interaction "potentials" between two regions. For example, given an air transport network of a country, if we operated a new route between cities that did not have one before, an expected passenger volume could be calculated. The volume is difficult to verify but helps to decide whether the new route is worth starting, which attaches great significance for transportation planning. Under this circumstance, negative sampling is no longer needed because the model aims to predict flows that do not exist. The third application is relatively theoretical. By comparing prediction results and the ground truth, we can analyze some spatial effects, e.g., Fig. 5(b) shows the distance decay effect has an impact on prediction errors. Moreover, as a black-box model, SI-GCN can be regarded as a complement to traditional mobility models.

VI. Conclusions and Future Work

Spatial origin-destination flow imputation enables us to obtain missing data and helps to discover the underlying relationship between geographical configurations and flow patterns. Nevertheless, the complex data structure of a flow and many geographical and social factors make it difficult to learn and predict flow distributions. In this research, we applied graph convolution to spatial origin-destination flow imputation and proposed a spatial interaction graph convolutional network (SI-GCN) model. It assumes that similar geographical units are more likely to produce similar flow patterns.

The proposed model adopts three techniques to improve the prediction accuracy. The first is geographical unit embedding. Regions with close spatial and non-spatial attributes as well as network neighborhood structures have similar embedding vectors. Graph convolution endows the model with the ability to integrate the first-order and the second-order similarities between geographical units. The second is negative sampling. Eq. (3) sums the loss on both positive and negative samples, pushing the model to well estimate both real and fake flows. Third, the nonlinear fitting ability of graph convolution makes it more easily to fit complex spatial distributions. The experiments using Beijing taxi trip data validated the usefulness of our model. Furthermore, we discussed that training the model with sufficient attributes of geographical units and a large training set can significantly increase the prediction accuracy. However, a biased training sample results in more errors and a larger negative sampling rate is not necessary. Our model also has a limitation. Because of its complicated structure and numerous parameters, the running time far exceeds those of baselines. To improve the model's efficiency, we can reduce training batch size, the dimension of embedding vectors or the negative sampling rate. A more efficient GCN framework can be adopted as well [69].

Future work will focus on incorporating temporal information for the forecasting of origin-destination flows, which can be implemented by combining time series models.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments to improve this paper.

References

[1] M. Batty, "Fifty years of urban modeling: Macro-statics to micro-dynamics," in *The Dynamics of Complex Urban Systems*, S. Albeverio, D. Andrey, P. Giordano, A. Vancheri, eds. Mendrisio, Switzerland: Physica-Verlag HD, 2008, pp. 1–20.

[2] A. Esparza and A. J. Krmenec, "Business services in the space economy: A model of spatial interaction," *Papers Regional Sci.*, vol. 73, no. 1, pp. 55–72, Jan. 2005.

[3] E. Manley and A. Dennett, "New forms of data for understanding urban activity in developing countries," *Appl. Spatial Anal. Policy*, vol. 12, no. 1, pp. 45–70, Mar. 2019.

[4] E. Marrocu and R. Paci, "Different tourists to different destinations. Evidence from spatial interaction models," *Tour. Manage.*, vol. 39, pp. 71–83, Dec. 2013.

[5] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.

[6] X. Zheng *et al.*, "Big data for social transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 620–630, Mar. 2016.

[7] Y. Liu *et al.*, "Social sensing: A new approach to understanding our socioeconomic environments," *Ann. Assoc. Amer. Geographers*, vol. 105, no. 3, pp. 512–530, May 2015.

[8] W. Tobler, "Spatial interaction patterns," *J. Environ. Syst.*, vol. 6, no. 4, pp. 271–301, 1976.

[9] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering spatial patterns in origin-destination mobility data," *Trans. GIS*, vol. 16, no. 3, pp. 411–429, Jun. 2012.

[10] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, "Uncovering space-independent communities in spatial networks," *Proc. Nat. Acad. Sci. USA*, vol. 108, no. 19, pp. 7663–7668, May 2011.

[11] S. Gao, Y. Liu, Y. Wang, and X. Ma, "Discovering spatial interaction communities from mobile phone data," *Trans. GIS*, vol. 17, no. 3, pp. 463–481, Jun. 2013.

[12] Y. Chen, J. Xu, and M. Xu, "Finding community structure in spatially constrained complex networks," *Int. J. Geograph. Inf. Sci.*, vol. 29, no. 6, pp. 889–911, Jun. 2015.

[13] J. Lin, "Network analysis of China's aviation system, statistical and spatial structure," *J. Transp. Geography*, vol. 22, pp. 109–117, May 2012.

[14] J. Wang, H. Mo, and F. Wang, "Evolution of air transport network of China 1930–2012," *J. Transp. Geography*, vol. 40, pp. 145–158, Oct. 2014.

[15] X. Liu, C. Kang, L. Gong, and Y. Liu, "Incorporating spatial interaction patterns in classifying and understanding urban land use," *Int. J. Geogr. Inf. Sci.*, vol. 30, no. 2, pp. 334–350, 2016.

[16] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervas. Comput.*, vol. 10, no. 4, pp. 36–44, Apr. 2011.

[17] Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin, "Understanding the bias of call detail records in human mobility research," *Int. J. Geograph. Inf. Sci.*, vol. 30, no. 9, pp. 1738–1762, Sep. 2016.

[18] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, p. 1376, Dec. 2013.

[19] J. R. Roy and J.-C. Thill, "Spatial interaction modelling," *Papers Regional Sci.*, vol. 83, no. 1, pp. 339–361, 2004.

[20] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, Apr. 2012.

[21] D. Zhu, Z. Huang, L. Shi, L. Wu, and Y. Liu, "Inferring spatial interaction patterns from sequential snapshots of spatial distributions," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 4, pp. 783–805, Apr. 2018.

[22] K. E. Haynes, and A. S. Fotheringham, *Gravity and Spatial Interaction Models*. London, U.K.: Sage, 1984.

[23] J. Mennis and D. Guo, "Spatial data mining and geographic knowledge discovery—An introduction," *Comput., Environ. Urban Syst.*, vol. 33, no. 6, pp. 403–408, 2009.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] S. Openshaw, "Modelling spatial interaction using a neural net," in *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, M. M. Fischer, P. Nijkamp, eds. Berlin, Germany: Springer, 1993, pp. 147–164.

[26] M. M. Fischer and S. Gopal, "Artificial neural networks: A new approach to modeling interregional telecommunication flows," *J. Reg. Sci.*, vol. 34, no. 4, pp. 503–527, 1994.

[27] W. R. Black, "Spatial interaction modeling using artificial neural networks," *J. Transp. Geography*, vol. 3, no. 3, pp. 159–166, Sep. 1995.

[28] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Conf. Track 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[29] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. 30th Conf. Neural Inf. Proces. Syst. (NIPS)*, 2016, pp. 3844–3852.

[30] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.

[31] D. Chai, L. Wang, and Q. Yang, "Bike flow prediction with multi-graph convolutional networks," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2018, pp. 397–400.

[32] X. Yan, T. Ai, M. Yang, and H. Yin, "A graph convolutional neural network for classification of building patterns using spatial vector data," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 259–273, Apr. 2019.

[33] D. Zhu *et al.*, "Understanding place characteristics in geographic contexts through graph convolutional neural networks," *Ann. Amer. Assoc. Geographers*, vol. 110, no. 2, pp. 408–420, Mar. 2020.

[34] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. 15th Int. Conf. ESWC*, 2018, pp. 593–607.

[35] H. Barbosa *et al.*, "Human mobility: Models and applications," *Phys. Rep.*, vol. 734, pp. 1–74, Mar. 2018.

[36] G. J. Abel and N. Sander, "Quantifying global international migration flows," *Science*, vol. 343, no. 6178, pp. 1520–1522, Mar. 2014.

[37] T. Grosche, F. Rothlauf, and A. Heinzl, "Gravity models for airline passenger volume estimation," *J. Air Transp. Manage.*, vol. 13, no. 4, pp. 175–183, Jul. 2007.

[38] M. Mozolin, J.-C. Thill, and E. Lynn Usery, "Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation," *Transp. Res. B, Methodol.*, vol. 34, no. 1, pp. 53–73, Jan. 2000.

[39] H. Murat Celik, "Modeling freight distribution using artificial neural networks," *J. Transp. Geography*, vol. 12, no. 2, pp. 141–148, Jun. 2004.

[40] S. A. Stouffer, "Intervening opportunities: A theory relating mobility and distance," *Amer. Sociol. Rev.*, vol. 5, no. 6, pp. 845–867, Dec. 1940.

[41] A. G. Wilson, "A statistical theory of spatial distribution models," *Transp. Res.*, vol. 1, no. 3, pp. 253–269, Nov. 1967.

[42] M. L. Hazelton, "Inference for origin–destination matrices: Estimation, prediction and reconstruction," *Transp. Res. B, Methodol.*, vol. 35, no. 7, pp. 667–676, Aug. 2001.

[43] B. Li, "Bayesian inference for origin-destination matrices of transport networks using the EM algorithm," *Technometrics*, vol. 47, no. 4, pp. 399–408, Nov. 2005.

[44] A. Tesselkin and V. Khabarov, "Estimation of origin-destination matrices based on Markov chains," *Proc. Eng.*, vol. 178, pp. 107–116, Jan. 2017.

[45] B. Li, "Markov models for Bayesian analysis about transit route origin-destination matrices," *Transp. Res. B, Methodol.*, vol. 43, no. 3, pp. 301–310, Mar. 2009.

[46] X. Jang and X. Yao, "Interpolating spatial interaction data," *Trans. GIS*, vol. 15, no. 4, pp. 541–555, 2011.

[47] Y. Zhang, T. Cheng, Y. Ren, and K. Xie, "A novel residual graph convolution deep learning model for short-term network-based traffic forecasting," *Int. J. GeoGraph. Inf. Sci.*, vol. 34, no. 5, pp. 969–995, May 2020.

[48] X. Geng *et al.*, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 3656–3663.

[49] Y. Long and J.-C. Thill, "Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing," *Comput., Environ. Urban Syst.*, vol. 53, pp. 19–35, Sep. 2015.

[50] D. Zhu, N. Wang, L. Wu, and Y. Liu, "Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data," *Appl. Geography*, vol. 86, pp. 152–164, Sep. 2017.

[51] A. S. Fotheringham, "Spatial structure and distance-decay parameters," *Ann. Assoc. Amer. Geographers*, vol. 71, no. 3, pp. 425–436, 1981.

[52] P. J. Taylor, "Distance transformation and distance decay functions," *Geograph. Anal.*, vol. 3, no. 3, pp. 221–238, Sep. 2010.

[53] B. Kotnis and V. Nastase, "Analysis of the impact of negative sampling on link prediction in knowledge graphs," 2017, *arXiv:1708.06816*. [Online]. Available: http://arxiv.org/abs/1708.06816

[54] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 2071–2080.

[55] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[56] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1616–1637, Sep. 2018.

[57] B. Yang, W.-T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. Conf. Track 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–12.

[58] X. Kong, Y. Liu, Y. Wang, D. Tong, and J. Zhang, "Investigating public facility characteristics from a spatial interaction perspective: A case study of beijing hospitals using taxi data," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 2, p. 38, Feb. 2017.

[59] M. Mazzoli, A. Molas, A. Bassolas, M. Lenormand, P. Colet, and J. J. Ramasco, "Field theory for recurrent mobility," *Nature Commun.*, vol. 10, no. 1, p. 3895, Dec. 2019.

[60] J.-C. Thill and M. Mozolin, "Feedforward neural networks for spatial interaction: Are they trustworthy forecasting tools," in *Spatial Economic Science*, A. Reggiani, Ed. Berlin, Germany: Springer, 2000, pp. 355–381.

[61] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2018, pp. 3538–3545.

[62] M. Lenormand, S. Huet, F. Gargiulo, and G. Deffuant, "A universal model of commuting networks," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e45985.

[63] G. F. Jenks, "The data model concept in statistical mapping," *Int. Yearbook Cartography*, vol. 7, no. 1, pp. 186–190, 1967.

[64] Y. Chen, "The distance-decay function of geographical gravity model: Power law or exponential law?" *Chaos, Solitons Fractals*, vol. 77, pp. 174–189, Aug. 2015.

[65] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *J. Geograph. Syst.*, vol. 14, no. 4, pp. 463–483, Oct. 2012.
[66] Y. Xiao, F. Wang, Y. Liu, and J. Wang, "Reconstructing gravitational attractions of major cities in China from air passenger flow data, 2001–2008: A particle swarm optimization approach," *Prof. Geographer*, vol. 65, no. 2, pp. 265–282, May 2013.
[67] T. Nakaya, "Local spatial interaction modelling based on the geographically weighted regression approach," *GeoJournal*, vol. 53, no. 4, pp. 347–358, 2001.
[68] M. Kordi and A. S. Fotheringham, "Spatially weighted interaction models (SWIM)," *Ann. Amer. Assoc. Geographers*, vol. 106, no. 5, pp. 990–1012, Sep. 2016.
[69] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 257–266.

Xin Yao received the B.S. degree from Wuhan University in 2015. He is currently pursuing the Ph.D. degree in GIScience with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His primary research interests include spatial data mining and geographical information visualization.

Yong Gao received the B.S. degree from Beijing Normal University in 1997, and the M.S. and Ph.D. degrees from Peking University in 2000 and 2003, respectively. He is currently an Associate Professor of GIScience with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interests include spatial data mining and geographic information science.

Di Zhu received the B.S. degree in geographic information systems from Peking University in 2014 and the B.S. degree in economics from the National School of Development, Peking University, in 2014, where he is currently pursuing the Ph.D. degree in GIScience with the Institute of Remote Sensing and Geographical Information Systems. His research interests include geographical analysis and deep learning.

Ed Manley received the Ph.D. degree in engineering from University College London in 2013. He is currently a Professor of urban analytics with the School of Geography, University of Leeds. He is also a Turing Fellow with the Alan Turing Institute and an Honorary Professor with University College London. His research interests include urban data science, agent-based modeling, spatial cognition, data visualization, and travel and mobility.

Jiaoe Wang received the B.A. degree from Beijing Normal University in 2003 and the Ph.D. degree from the Institute of Geographic Sciences and Natural Resources Research (IGSNRR), China Academy of Sciences (CAS), in 2008. She is currently a Researcher with IGSNRR, CAS, and a Professor with the College of Resources and Environment, University of Chinese Academy of Sciences. Her main research interests include transport geography and regional development, big data and urban transportation, economic geography, and innovation and industry development.

Yu Liu received the B.S., M.S., and Ph.D. degrees from Peking University in 1994, 1997, and 2003, respectively. He is currently a Professor of GIScience with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interest mainly concentrates on the humanities and social sciences based on big geo-data.