Graph Deep Learning Model for Network-based Predictive Hotspot Mapping of Sparse Spatio-Temporal Events

Yang Zhang*, Tao Cheng

SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental & Geomatic Engineering, University College London, London WC1E 6BT, UK

Abstract: The predictive hotspot mapping of sparse spatio-temporal events (e.g., crime and traffic accidents) aims to forecast areas or locations with higher average risk of event occurrence, which is important to offer insight for preventative strategies. Although a network-based structure can better capture the micro-level variation of spatio-temporal events, existing deep learning methods of sparse events forecasting are either based on area or grid units due to the data sparsity in both space and time, and the complex network topology. To overcome these challenges, this paper develops the first deep learning (DL) model for network-based predictive mapping of sparse spatio-temporal events. Leveraging a graph-based representation of the network-structured data, a gated localised diffusion network (GLDNet) is introduced, which integrating a gated network to model the temporal propagation and a novel localised diffusion network to model the spatial propagation confined by the network topology. To deal with the sparsity issue, we reformulate the research problem as an imbalance regression task and employ a weighted loss function to train the DL model. The framework is validated on a crime forecasting case of South Chicago, USA, which outperforms the state-of-the-art benchmark by 12% and 25% in terms of the mean hit rate at 10% and 20% coverage level, respectively.

Key words: deep learning; sparse spatio-temporal data; predictive hotspot mapping; graph; network

1 Introduction

Hotspot mapping highlights areas/locations that have higher than average incidence of events, sparsely distributing in space and time. Early efforts on hotspot mapping are primarily retrospective, aiming to measure and detect space-time clusters of historical sparse spatio-temporal data, including public safety (Miaou et al., 2003), earthquake occurrences (Vere-Jones, 2009), crimes (Davies and Marchione, 2015, Ratcliffe, 2010), epidemiology (Ugarte et al., 2014), and environmental science (Apte et al., 2017), among others. Recently, there has been an increasing interest in using historical data to produce hotspot maps for predictive purpose owing to its prospective benefits. For instance, accurate crime forecasting can help police enforcements to prevent criminal behaviours, and traffic accident prediction is useful for road safety interventions and traffic reengineering.

* Corresponding author: email: yang.zhang.16@ucl.ac.uk (Y. Zhang)

Predictive hotspot mapping aims to generate a prospective risk surface, which is a graphical representation of the estimated probabilities of an event, or series of events, occurring at each of the geographic units within an area of interest (Bowers et al., 2004). According to the spatial scales of the geographic units, existing predictive hotspot mapping algorithms are categorised into two groups: area-level methods and network-level methods.

1.1 Area-Level Predictive Hotspot Mapping and Limitations

Most previous efforts have been devoted to the statistical predictive mapping of hotspots based on area units, including grids (Mohler et al., 2011, Chainey et al., 2008) and administrative districts (Kleinman et al., 2005, Santini and Valentini, 2011). They usually divided an area of interest into regular grids, and then aggregated historical crime data within each grid. Commonly used statistical approaches include prospective hot spotting (Bowers et al., 2004), prospective space-time scan statistics (Takahashi et al., 2008), the self-exciting point process (SEPP) (Mohler et al., 2011), and temporally weighted kernel density estimate (KDE) (Porter and Reich, 2012). These methods aim to generate an expected probability surface to indicate where events are likely to occur in a particular period (e.g., the next day). For instance, series of studies (Mohler et al., 2011, Mohler, 2014) have investigated to use the SEPP for predictive crime hotspot mapping and they have shown that SEPP outperforms many other statistical approaches. However, statistical models require large efforts on parameter inference and empirical studies have reported several non-functional scenarios of these approaches (Adepeju, 2017, Rosser and Cheng, 2016).

Alternatively, deep learning (DL), a cutting-edge machine learning method, has been used for spatio-temporal events modelling and prediction using the grid-based data representation. The success of DL models, such as convolution neural network (CNN) and recurrent neural network (RNN), is owing to their powerful capability of spatial or temporal feature representation without prior knowledge. For example, a spatio-temporal residual CNN model was employed to predict the crimes in cities (Wang et al. (2017b). Lin et al. (2018) proposed a deep neural network by incorporating the criminal environmental information with spatio-temporal crime features for grid-based crime forecasting. Similarly, Ke et al. (2017) developed an ‘end-to-end’ CNN-based DL model for on-demand ride services (car-hailing events) prediction. Similar works include (Cheng et al., 2018, Liu et al., 2018, Liao et al., 2018, Ren et al., 2019a).

All abovementioned approaches, however, are area-level predictive hotspots mapping methods, which have several limitations. First, the size of the area units is often empirically and arbitrarily determined by the planners. The level of prediction precision changes in response to different spatial scales, which is referred to as the modifiable areal unit problem (Levin and Peres, 2017, Cheng and Adepeju, 2014). Second, the level of spatial correlation between adjacent area units may vary across the region of interest due to the underlying physical structures, which is referred to as the spatial heterogeneity (Cheng et al., 2014). For example, the concentration pattern of crimes in cities might be affected by natural or artificial geographic features (e.g., lakes, parks and street network), but the area-level data representation omits the geographical information, which is likely to decrease the prediction accuracy (Rosser and Cheng, 2016).

1.2 Network-Level Predictive Hotspot Mapping and Motivations

In the real world, many events naturally reside in a network structure, such as many urban crimes (Rosser et al., 2017), road traffic accidents (Yannis et al., 2017, Xie and Yan, 2008), and taxis' pick-up events (Tang et al., 2016). The network structure might fundamentally influence the distribution patterns of hotspots. Hence, network-based data representation is necessary to better capture the micro-level variation of events over space and time (Sullivan and Feinn, 2012, Zhu et al., 2017). In addition, the use of network-based predictive hotspot mapping is also well-motivated considering its practical utility. For instance, street-network-based crime prediction could effectively guide the police patrolling in cities (Chen et al., 2018, Chen et al., 2017). Traffic demand (e.g., taxi-hailing) forecasting alongside the road network contributes to optimising the urban transportation resources allocation and improving traffic efficiency (Tang et al., 2016).

Recent attention has been shifting from macroscale area-level predictive mapping to microscale network-level hotspot forecasting. Several works pioneered to extend some conventional statistical models to a network-based version. For examples, network-based space-time scan statistics has been proposed for retrospective crime hotspot mapping (Shiode and Shiode, 2014) and traffic congestion detection (Anbaroglu et al., 2014). Similarly, Tang et al. (2016) developed a network KDE method to analyse the distribution characteristics of taxi's pick-up events over the street network space. For the purpose of predictive mapping, Rosser et al. (2017) presented the first attempt of a prospective network-time KDE (NTKDE) model for crime forecasting, which is the state-of-the-art network-based predictive hotspot mapping method. However, the NTKDE draws various parameter assumptions about the latent dynamics governing the generation of the observed hotspot patterns. In addition, it also suffered from time-consuming parameter inference issue.

Lately, standard DL models have been extended to deal with network-structured (i.e., graph-structured) data (Defferrard et al., 2016, Henaff et al., 2015), and latest studies have focused on developing DL models for graph-structured spatio-temporal data modelling and prediction. For instance, Li et al. (2017) introduced a diffusion convolutional recurrent neural network to predict road network traffic flow. The model used graph convolution with weight sharing strategy to capture spatial dependency and leveraged recurrent neural networks to model temporal dependency. (Yu et al., 2017) proposed a spatio-temporal graph convolution networks to predict the traffic flow at turning junction networks. Similar works can be referred to (Zhang et al., 2019, Ren et al., 2019b). However, these studies focused on dense space-time process with flow data (e.g., traffic flow or population flow), which differs from the sparse events data (e.g., crimes). Therefore, DL methods for predictive hotspot mapping of sparse spatio-temporal events at a network level are still a research gap.

1.3 Challenges and Aims

The key challenge of DL for predictive hotspots mapping lies in how to model the complex spatio-temporal dependencies of sparse events along the network. The sparsity means that counting the events over space and time results in many zero counts of some segments/links of the network. The difficulties are in three aspects.

First, in temporal domain, the sequence of event counts over time is far from a continuous function that can be approximated in conventional DL models (Mhaskar and Poggio, 2016). In addition, predictive mapping requires sufficient events being included in the input, so that the input length is usually much longer than that in classical time series predictions. In this scenario, RNN-based DL model for temporal modelling suffers from time-consuming iterations and slow response to dynamic changes.

Second, in spatial domain, commonly-used graph-based DL models learn spatial dependencies using weight sharing strategy. Directly applying such DL models to spatially sparse event data will lead to all zero weights (Wang et al., 2017a), and then fail to generate a predictive mapping.

Third, predictive mapping is not a trivial regression task. At a chosen prediction time, most spatial units have zero event counts, which leads to the creation of an imbalanced regression learning scenario (Krawczyk, 2016). In this scenario, the direct use of the standard regression loss function for parameter learning will make the DL model prone to overfitting or lead to all zero predictions.

In this paper, we aim to develop a novel and effective graph-based DL framework, named gated localised diffusion network (GLDNet), to generate predictive hotspot mapping of spatio-temporal events in network space. As many events occurred due to the human activities (e.g. crime and traffic accidents), here, we consider the street-network-based events without loss of generality. Here the street network is represented as a weighted and undirected graph, where event counts are defined as the values on the vertex set of the graph. In GLDNet, the temporal propagation of historical events is modelled by a gated network and the associated spatial propagation is captured via a localised diffusion network in terms of the network distance and topology, which overcomes the spatial heterogeneity. In the model training process, a weighted regression loss function is employed to solve the issue of many zero observations. The proposed model is evaluated using crime data from the City of Chicago, Illinois, USA to prove its feasibility and effectiveness. To our best knowledge, this is the first attempt to develop graph-based DL approaches for predictive hotspot mapping of sparse spatio-temporal data on networks.

The remainder of this paper is organised as follows. Section 2 formulises the research problem of the network-based predictive hotspot mapping. Section 3 presents the proposed novel GLDNet approach. Sections 4 describes the crime forecasting case study to demonstrate the effectiveness of the proposed model. Finally, concluding remarks are given in Section 5.

2 Problem Formulation

The goal of this is to generate a probability distribution on the street network to indicate where events (e.g. crimes) are likely to occur at a particular period (e.g., the next day). This section first discusses the network-based data representation as a graph, and then the network-based predictive hotspot mapping problem is formulised.

2.1 Network-Structured Data Representation

In real world, many events occur and present a spatial distribution on the road network, such as urban crime, traffic accidents and taxis' pick-up on street segments. To perform spatio-temporal modelling on the network, it is essential to represent the network-based data in formal mathematical terms. This can be done using terminology from graph theory and graph signal processing (Shuman et al., 2013).

First, a dual approach is adopted to express a street network structure as a graph $G = (V, E, W)$, where V, E, and W are the graph nodes set, edge set and weight matrix, respectively. The dual approach represents the street segment as a set of graph nodes V, while each intersection between any pair of adjacent street segments i and j is turned into one edge $e_{ij} \in E$. The weight w_{ij} associated with each graph edge e_{ij} often indicates the similarity between the two vertices it connects. According to the First Law of Geography (Tobler, 1970), the spatial dependency of adjacent geographic units decays with the increase of distance. As this law is also applicable for sparse spatio-temporal events (Leitner et al., 2018, Weisburd et al., 2009), this paper defines the edge weight to be inversely proportional to the network distance between the midpoints of two adjacent street segments. A common way to calculate the edge weight is via a Gaussian kernel weighting function:

$$w_{ij} = \begin{cases} \exp\left(-\frac{[dist(i,j)]^2}{2\alpha^2}\right) & \text{if } e_{ij} \text{ is an edge} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

for some parameter α. The $dist(i,j)$ represents the distance between the two midpoints of street segments i and j.

Second, events are matched to the nearest street segments and they are then aggregated at a chosen time scale (e.g., one day interval), as the values at the associated graph nodes. Supposing a graph has N nodes, the values at the N nodes at a certain time t are collectively referred to as a graph signal $\boldsymbol{x}_t \in \mathbb{R}^N$. Figure 1 presents a graphical example to illustrate the representation approach. This representation allows for a network-oriented computational analysis regarding the network topology.

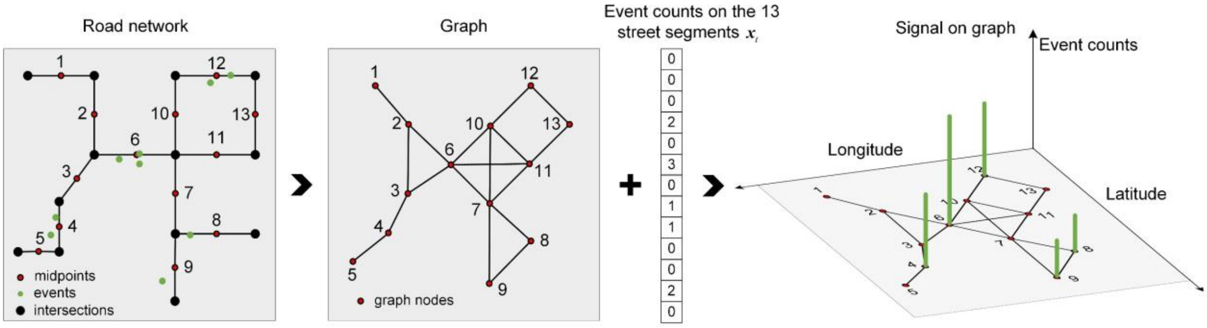

Figure 1. Represent the network-structured data as a graph signal.

2.2 *Network-based Predictive Hotspot Mapping Problem*

Network-based predictive hotspot mapping aims to learn a mapping function *F*, which uses historical observations in previous *M* time steps (referred to as the time window) to forecast where events are likely to occur in the next time step given the topology of a graph G. Thus, this problem can be formulated as:

$$\boldsymbol{y}_t = F(\boldsymbol{x}_{t-1}, \cdots, \boldsymbol{x}_{t-M} | G) \tag{2}$$

where $\boldsymbol{x}_i \in \mathbb{R}^N$ is a signal on a graph with *N* nodes at time step *i*. Each entry of $\boldsymbol{x}_i$ is the historical event counts at a graph node. $\boldsymbol{y}_t \in \mathbb{R}^N$ is the estimated graph signal, indicating the predicted occurrence probability at time step *t*.

3 Graph-based Deep Learning Method

The aim of the predictive hotspot mapping is to develop methods to model the spatio-temporal propagation of the events. This section presents the methodological details of the proposed gated localised diffusion network (GLDNet) model, which enables to carry out predictive mapping of sparse events in the space-time dimension whose spatial propagation is confined by a network structure.

3.1 *Modelling the Temporal Spread of Events*

Many types of events exhibit temporal dependency which suggests these events are not independent but are instead related in predictable ways over time. Among DL models, RNN and its variants (e.g., long short-term memory), which utilise a gating mechanism and a recurrent structure, become widespread in temporal dependency learning. However, event sequences are much temporally sparser than traditional time series. This raises two issues. First, DL models poorly approximate the discrete point process. Second, DL models require sufficient historical observations to learn unseen patterns for forecasting. Thus, the input sequence is usually much longer for predictive hotspot mapping than that for time series forecasting. In this circumstance, RNN-based models suffer from time-consuming iterations and a slow response to dynamic changes.

To solve these two issues, we consider data smoothing to boost the performance of a gated network for temporal dependency modelling of sparse data. The smooth technique augments the data in temporal domain and the proposed gated network enables an effective long-range temporal dependency modelling.

3.1.1 Data smoothing

This paper proposes to use the simple exponential smoothing (SES) technique to construct an augmented data sequence serving as the input features. SES is a rule of thumb technique for smoothing time-tagged data using the exponential window function, which assigns exponentially decreasing weights over time to the past observations. Supposing the raw data sequence is denoted as $\{x_t\}$ beginning at time $t = 0$, a simple form of SE is given by the formulas:

$$\begin{aligned} s_0 &= x_0 \\ s_t &= \alpha_s x_t + (1 - \alpha_s) s_{t-1}, t > 0 \end{aligned} \tag{3}$$

where s_t is the output of the SES algorithm, α_s is the smoothing factor, and $0 < \alpha_s < 1$. Figure 2 displays an example of the SES with $\alpha_s = 0.1$. A smaller smoother factor leads to a smoother augumented data sequence. The data smoothing scheme allows the DL model to learn the latent temporal patterns from temporally sparse data.

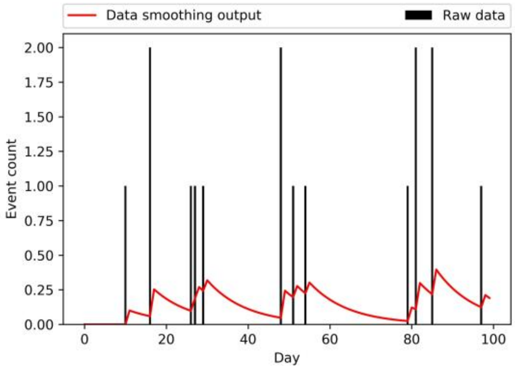

Figure 2. A simple exponential smoothing example with $\alpha_s = 0.1$

3.1.2 Gated Network

An effective way to achieve a temporal dependency modelling of a very long sequence is to use the gating mechanism (Dauphin and Grangier, 2015) without recurrent structure. The gating mechanism allows the network to control what temporal influence can be propagated through a gate structure.

This article utilises a gated network (GNet) to model the temporal propagation of events. Comparing with RNN or LSTM, the gated network has only one gate, which greatly speed up the model training process. The GNet consists of $h_0,\ldots,h_L$ hidden layers, which is formulated as:

$$\boldsymbol{X}^{l+1} = h_l(\boldsymbol{X}^l) = ReLU(\boldsymbol{X}^l \boldsymbol{W}^l + \boldsymbol{b}^l) \odot \sigma(\boldsymbol{X}^l \boldsymbol{V}^l + \boldsymbol{c}^l) + \boldsymbol{d}^l \tag{4}$$

Let N represents the number of graph nodes, and n^l denote the dimension of the node features of the l-th hidden layer (referred to as the hidden units). Then, $\boldsymbol{X}^l \in \mathbb{R}^{N\times n^{l-1}}$ is the input of the layer h_l, $\boldsymbol{X}^{l+1} \in \mathbb{R}^{N\times n^l}$ is the output. $\boldsymbol{W}^l$, $\boldsymbol{V}^l \in \mathbb{R}^{n^{l-1}\times n^l}$, $\boldsymbol{b}^l, \boldsymbol{c}^l \in \mathbb{R}^{n^l}$, $\boldsymbol{d}^l \in \mathbb{R}^N$ are learnable parameters. $ReLU(x) = \max(0, x)$ is the Rectified Linear Unit (ReLU) function, which allows the gradient to easily propagate through a deep structure (Britz, 2015). $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function and $\odot$ is the element-wise product. In this paper, the hidden layer is termed as the gated ReLU layer (GRL).

More specifically, $\sigma(\boldsymbol{X}^l\boldsymbol{V}^l + \boldsymbol{c}^l)$ is termed as a gate, which modulates the non-linear projection $ReLU(\boldsymbol{X}^l\boldsymbol{W}^l + \boldsymbol{b}^l)$ to control the information passed on in this layer and model the accumulative influence of the past events on the current status. In addition, $\boldsymbol{d}^l$ gives a base probability level for the occurrence of the future event at each graph node. The base probability is analogous to the background intensity of event occurrence in the self-exciting point process model for predictive hotspot mapping (Mohler et al., 2011). The proposed gated network is then constructed by stacking multiple GRLs, written as:

$$GNet(\boldsymbol{X}) = h_L\left(h_{L-1}(\cdots h_2(h_1(\boldsymbol{X}))\cdots)\right) \tag{5}$$

3.2 *Modelling the Spatial Spread of Events*

Empirical studies have confirmed the tendency for some events to concentrate on a range of street segments (Rosser et al., 2017, Braga et al., 2011). However, modelling the spatial propagation of network-structured events is difficult because the spread is constrained by the associated graph topology. Additionally, the event counts are extremely spatially sparse. Traditional DL models adopt either a 'fully-connected structure without weight sharing' or a 'global parameter-sharing scheme'. The former easily overfits while the latter leads to all zero parameters to fit the extreme sparse data.

To solve these issues, this paper develops a localised diffusion network (LDN). The network-based spatial propagation is modelled as an information diffusion process on a graph, which adopts a novel localised parameter sharing scheme as a trade-off between the 'no weight sharing scheme' and the 'global weight sharing scheme'.

3.2.1 *Information Diffusion Process*

We consider the spread of sparse events as an information diffusion process (Zhou and Schölkopf, 2004). A basic information diffusion process is characterised by random walks on a graph G with a transition matrix $\boldsymbol{P}$, defined as:

$$\boldsymbol{P} = D^{-1}W \tag{6}$$

where $W \in \mathbb{R}^{N\times N}$ is the weight matrix of the graph, $D \in \mathbb{R}^{N\times N}$ is the diagonal matrix with $D_{ii} = \sum_j w_{ij}$, and the element p_{ij} in $\boldsymbol{P} \in \mathbb{R}^{N\times N}$ denotes the probability of the event influence propagating from the i-th node to a successive j-th node. Suppose a graph signal $\boldsymbol{x} \in \mathbb{R}^N$ is a column vector where the *i*-th entry $\boldsymbol{x}(i)$ indicates the value at the *i*-th graph node. The values at all graph nodes after one-step random walk $\boldsymbol{x}'$ is written as:

$$\boldsymbol{x}' = \boldsymbol{P}\boldsymbol{x} \tag{7}$$

Specifically, the *i*-th row of $\boldsymbol{x}'$ in Eq. (7) satisfies:

$$\boldsymbol{x}'(i) = (\boldsymbol{P}\boldsymbol{x})(i) = \frac{\sum_{j\in\varkappa_i} w_{ij}\boldsymbol{x}(i)}{\sum_{j\in\varkappa_i} w_{ij}} \tag{8}$$

where the neighbourhood $\varkappa_i$ is the set of vertices connected to the *i*-th vertex by a weighted edge. According to Eq. (8), each node can receive the information from its first-order adjacent neighbours after one step random walk. Therefore, a one-step random walk essentially serves the purpose of measuring to what extent a graph node is affected by its nearby vertices. The information diffusion process can be regarded as a sequence of random walks on the graph and the states of graph nodes will achieve a stationary distribution after infinite random walks.

3.2.2 Localised Diffusion Network

In practice, a stational distribution of a diffusion process can be approximated as a weighted combination of finite *K*-step random walks on a graph (Teng, 2016). This paper proposes a novel LDN to model a *K*-step truncated diffusion process for spatial modelling. The LDN consists of *K* LDN layers, denoted as g_0,…,g_K. Let $\boldsymbol{X}^k \in \mathbb{R}^{N\times m^{k-1}}$ be the input of the *k*-th layer g_k, where *N* is the number of graph nodes, and m^{k-1} is the dimension of the output features of the (*k*-1)-th layer (referred to as the hidden units). The *k*-th layer is formulated as:

$$\boldsymbol{X}^{k+1} = g_k(\boldsymbol{X}^k) = ReLU(\boldsymbol{X}^k * \boldsymbol{\theta}^k + (\boldsymbol{P}\boldsymbol{X}^k) * \boldsymbol{\eta}^k) \tag{9}$$

where $\boldsymbol{\theta}^k$, $\boldsymbol{\eta}^k \in \mathbb{R}^{N\times m^{k-1}\times m^k}$ are learnable parameters in the *k*-th layer, $\boldsymbol{X}^{k+1} \in \mathbb{R}^{N\times m^k}$ is the output of the *k*-th layer, ReLU is the activation function, and $*$ is multidimensional array operator. For any arrays $\boldsymbol{A} \in \mathbb{R}^{D_1\times D_2}$, $\boldsymbol{B} \in \mathbb{R}^{D_1\times D_2\times D_3}$, the operation $\boldsymbol{C} = \boldsymbol{A} * \boldsymbol{B} \in \mathbb{R}^{D_1\times D_3}$ satisfies that

$$c_{ij} = \sum_{k=1}^{D_2} a_{ik}\, b_{ikj}, 0 \le i < D_1 \text{ and } 0 \le j < D_3 \tag{10}$$

In Eq. (9), the first term $\boldsymbol{X}^k * \boldsymbol{\theta}^k \in \mathbb{R}^{N\times m^k}$ models the dependency of each node itself, because each node is its own zero$^{\text{th}}$-order neighbour. In the second component, $\boldsymbol{P}\boldsymbol{X}^k$ models the one-step random walk and the transition matrix $\boldsymbol{P}$ captures the static spatial similarity considering the network distance between nodes. However, the influence propagation over space might be heterogeneous. Thus, the learnable parameter $\boldsymbol{\eta}^k$ is employed to adaptively learn the spatial dependency between each node and its first-order adjacent neighbours.

Traditionally, a classical deep neural network has no weight-sharing scheme (Figure 3a), which means the connections between two successive layers have totally different learnable parameters (Ren et al., 2019b). For a large graph, the number of learnable parameters in one layer equals the number of graph edges. The over-parameterised structure substantially slows down the model training procedure and it would be easy to get overfitting (Louizos et al., 2017). A way to address both issues is using the global parameter-sharing scheme (Figure 3b). It means all nodes in the same layer use the same learnable parameter. It relies on the assumption of the homogeneity of the spatial dependency over the entire space, such as an image. However, this assumption is not suitable for the heterogeneous sparse network. For example, in Figure 3a, most of the segments have no observed events, but they are adjacent to several segments with non-zero event counts. The global weight sharing scheme will make the learnable parameters to be all zeros.

Here we propose to adopt a localised parameter-sharing scheme. In this scheme, the learnable parameters are only shared within the local neighbourhood of each node. For example, in Figure 3d, the first-order adjacent neighbours (nodes 2, 3, 7, 10, and 11) of node 6 in the LDN layer share the parameter, visualised using the blue lines with the same darkness. The localised parameter-sharing scheme allows the model to account for spatial heterogeneity and it is suitable for spatially sparse data.

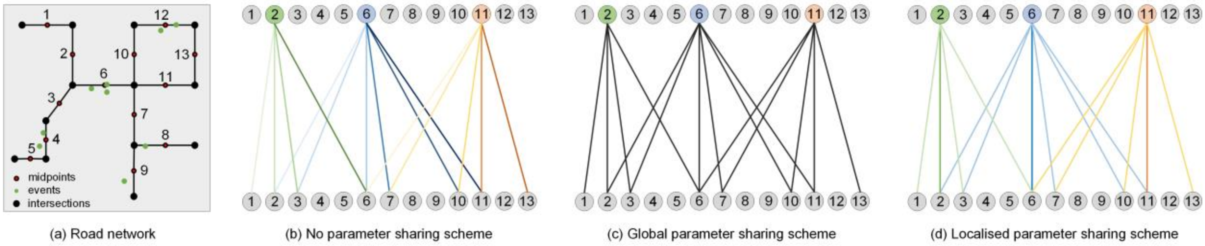
1
12
2
10
13
6
11
3
7
5
4
8
9
midpoints
events
intersections
(a) Road network
1 2 3 4 5 6 7 8 9 10 11 12 13
(b) No parameter sharing scheme
(c) Global parameter sharing scheme
(d) Localised parameter sharing scheme

Figure 3. Inner structure of parameter sharing schemes. Connections with the same colour share the same learnable parameter. (a) Road network (b) 'no parameter sharing scheme' (c) 'global parameter sharing scheme' and (d) 'localised parameter sharing scheme'.

LDN is constructed by stacking multiple LDN layers. As any high-order neighbours can be derived from the first-order neighbours, the proposed LDN can capture the spatial dependencies between a node and its, at most, K-th-order adjacent neighbours by successively stacking K hidden layers. The proposed LDN structure consisting of K hidden layers with an initial input $\boldsymbol{X}$ is formulated as:

$$LDN(\boldsymbol{X}) = g_K\left(g_{K-1}(\cdots g_2(g_1(\boldsymbol{X}))\cdots)\right) \tag{11}$$

Figure 4 demonstrates a two-layer LDN model. Taking the No.6 road segment as an example, the spatial dependencies between the zero$^{\text{th}}$- (segment 6) and the first-order (segments 2, 3, 7, 10, and 11) neighbours are captured via the first LDN layer. Then, the second-order (segments 1, 4, 8, 9, 12, and 13) spatial dependency is modelled via the second LDN layer.

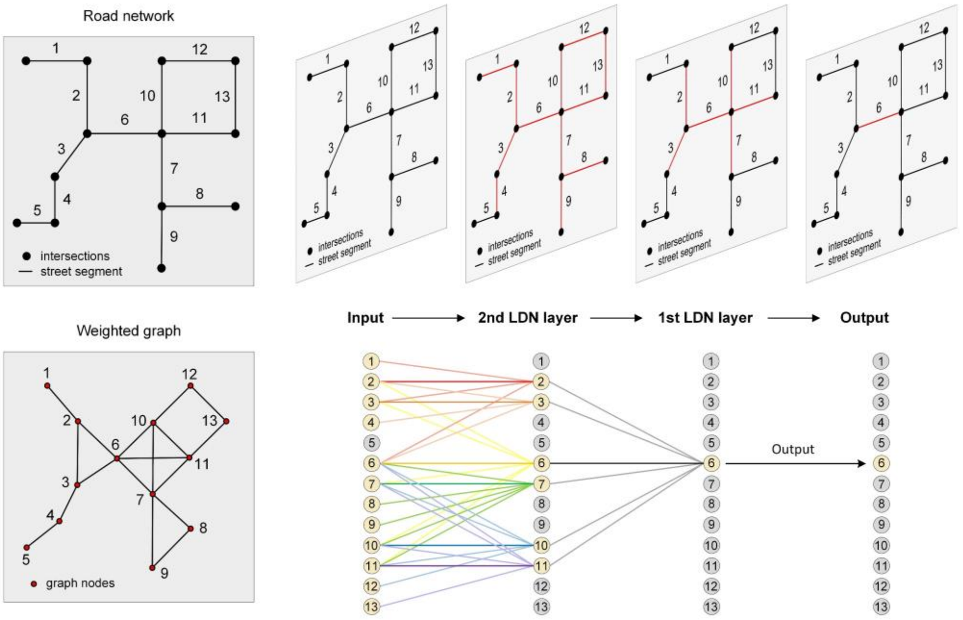

Figure 4. A two-layer LDN the spatial dependency between each node and its zero-, first- and second-order adjacent neighbours. The inner connections in each LDN layer can preserve the graph topology. The connections with the same colour in the lower right plot indicate that they share the same learnable parameters.

3.3 Gated Localised Diffusion Network

The influence of each historical event is spread through time and, as well, distributed across its immediate vicinity in space. To generate the network-based predictive mapping of spatio-temporal sparse events, a GLDNet model is constructed by integrating the gated network, the LDN and a fully connected layer, as displayed in Figure 5. The GLDNet can be formulated as:

$$GLDNet(\boldsymbol{X}) = g_K\big(\cdots g_1\big(h_L(\cdots h_1(\boldsymbol{X})\cdots)\big)\cdots\big)\cdot \boldsymbol{\omega}_{fc} + \ b_{fc} \tag{12}$$

where the output of $g_K(\cdot)$ is of dimension $\mathbb{R}^{N\times m^K}$, $\boldsymbol{\omega}_{fc} \in \mathbb{R}^{m^K}$ and b_{fc} are the learnable parameters in the fully-connected layer.

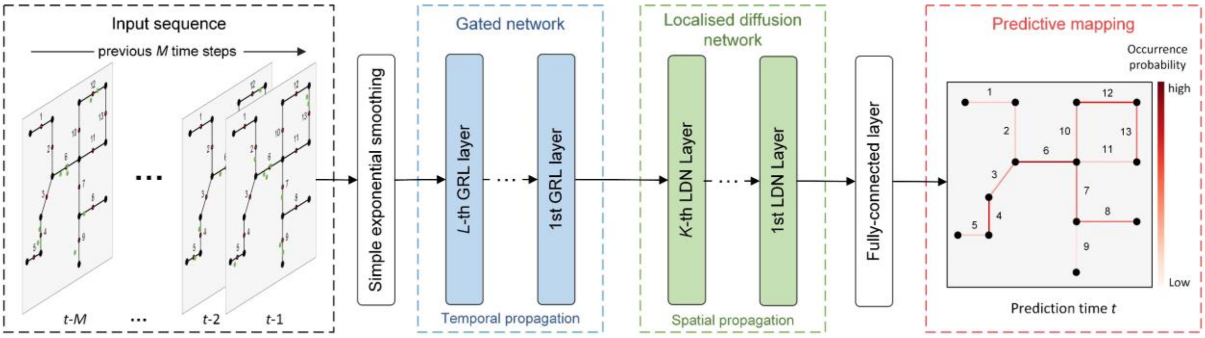

Figure 5. Framework of GLDNet

In this DL framework, the input sequence is first smoothed by SES and then fed into the GNet, which consists of L GRL layers for temporal propagation modelling. Afterwards, the output of the GNet is passed to a K-layer LDN for street-network-based spatial propagation modelling. The final output after the fully-connected layer is a predictive mapping indicating the probability of event occurrence on each street segment at the next time step.

3.4 Parameter Learning

All learnable parameters in the GLDNet can be adaptively adjusted during model training by minimising a training loss function. Traditionally, the most frequently used criterion is the mean-square error (MSE) for regression. However, the predictive mapping task is not a trivial regression task, because the observed events are typically rare. MSE measure is insufficient to measure error appropriately for sparse data. Thus, DL models simply using MSE as the loss function will tend to be biased to the most frequent and uninteresting cases (locations without event occurrence).

To tackle this issue, we treat the predictive mapping of sparse events as an imbalanced regression task. A novel weighted loss function is employed to make the learning algorithm focuses on these rare events. This strategy operates by assigning high misprediction cost to rare cases. The weighted loss function is formulated as:

$$loss = \frac{1}{N}\sum_{i=0}^{N} \omega_i(\hat{x}_i - x_i)^2 \tag{13}$$

where N is the number of graph node, the x_i, $\hat{x}_i$ are, respectively, the observed and predicted values at the i-th graph node. ω_i is the weight assigned to each square error $(\hat{x}_i - x_i)^2$, defined as

$$\omega_i = \begin{cases} x_i & \text{if } x_i > 0 \\ \rho & \text{otherwise,} \end{cases} \tag{14}$$

where $\rho \in [0, 1)$ is a predefined coefficient. That is, if $x_i > 0$ (there are events occurring at the i-th node), the misprediction cost is $x_i(\hat{x}_i - x_i)^2$, otherwise, it is $\rho(\hat{x}_i - x_i)^2$. With a small coefficient ρ, the misprediction cost at graph node with non-zero value is higher than the misprediction cost at node with zero value. Consequently, the weighted loss function enables the GLDNet model to focus on the most important and rare event occurrence. Leveraging the weighted loss function, all learnable parameters can be trained via backpropagation algorithm (LeCun et al., 2015).

4 Case Study

4.1 Experiment data and pre-processing

We validate the proposed GLDNet using a crime forecasting case. The crime data and the corresponding street network are downloaded from the City of Chicago open data portal. We consider three different crime types, burglary (3995 crimes), assault (5707 crimes), and theft (13033 crimes) in the south side of Chicago, covering a two-year period (731 days) starting on 1st January 2016. Chicago's sides are defined as a collection of multiple community areas, which are used for urban planning purposes. Figure 6 (a) illustrates the street network (5910 segments) of the South Chicago and Figure 6 (b)-(d) show the distribution of the total assaults, burglaries and thefts during the study period. The three subplots also highlight several geographical features (parks and cemetery) that have no crimes being observed, just like ‘holes’ in the spatial pattern, which suggests that crimes are preferentially aligned with streets.

Crimes are aggregated daily by street segment and predictions are made one day ahead. The one-day timescale is required by the patrolling police officers, who are assigned patrolling tasks on a daily basis. According to Figure 6a, the daily crimes are very sparse, and the spatial pattern varies across crime types. To construct the weight matrix of the graph representation of the street network, we set the parameter α^2 in Eq. (1) to be 5.

To construct the input-output pairs, we fix the time window M to be 100 in the three tasks. Thus, the length of the input-output pairs is 631. In this experiment, we use the last 100-day crime data to test both models. Regarding GLDNet training, we divide the rest of the data into a training set (431-day) and a validation set (100-day). The validation data are not used to train the model but to monitor the change of the predictive performance of GLDNet during the training process to prevent overfitting. The smoothing factor α_s in in Eq. (3) is set to be 0.5.

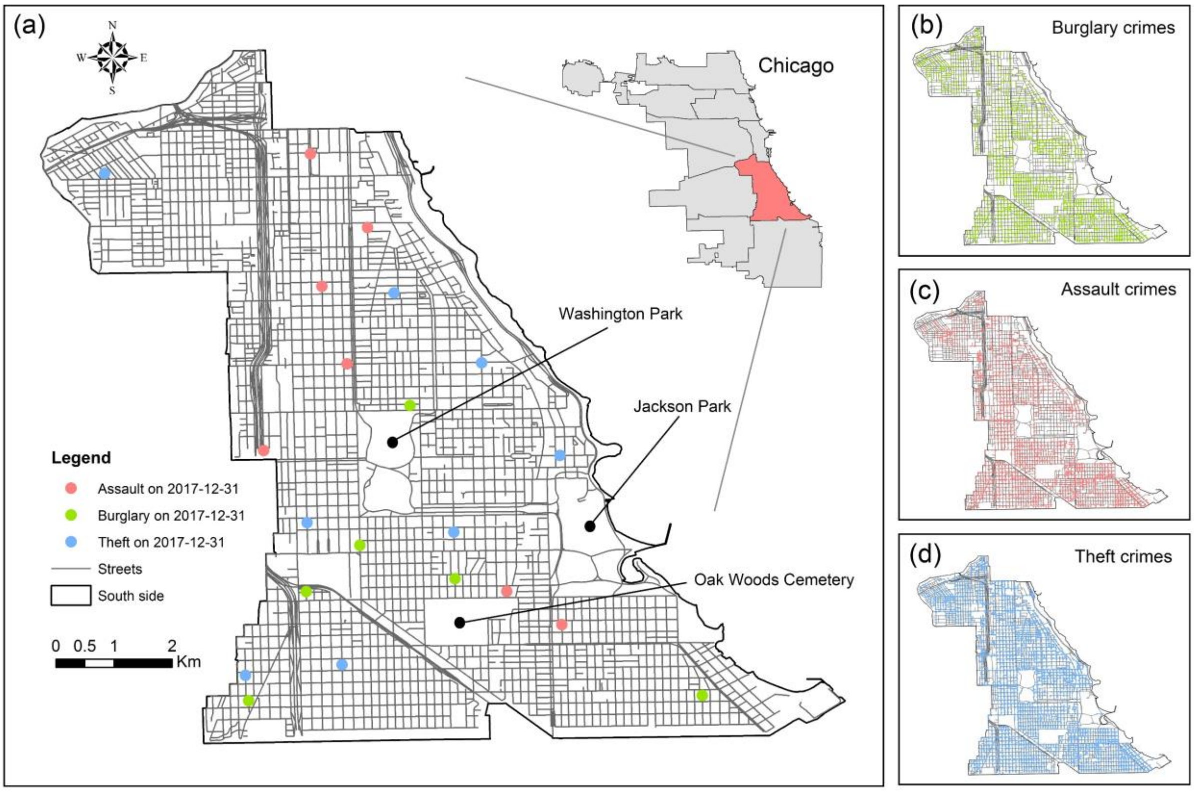

Figure 6. (a) The street network in South Chicago. The map also shows one-day crime records of the three different crime types on 31st December 2017. (b) (c) and (d) The South Chicago with assault, burglary, and theft crimes in two-year period overlaid, respectively.

4.2 *Performance measurement*

Due to the sparsity of events, we evaluate the mapping effectiveness using the mean hit rate with statistical quantification. ‘Hit rate’ is a widely-used metric to measure the performance of predictive hotspot mapping (Adepeju et al., 2016, Bowers et al., 2004, Rosser and Cheng, 2016). Hit rate is defined as the proportion of events accurately captured by the predicted locations. To calculate the hit rate, we first compute the prediction on each street segment, and then sort all segments by the predicted values in descending order. The street segments are selected in the sorted order and the proportion of the events falling on those street segments is tallied. In this paper, the hit rate is computed at maximum 20% street length coverage and the mean of the hit rates over all consecutive testing days is used as the metric for predictive mapping assessment.

The mean hit rate can directly reflect the comparison results of any two models, but the statistical significance of such results is unknown. To provide a robust evaluation of the improvement of one method over the other, we also present a statistical quantification. It means we treat the hit rates at a given coverage percentage of two different methods over all testing days as a pair of time series. Wilcoxon Signed-Rank (WSR) test is employed to assess whether one method is significantly better than the other. The statistic of the WSR test is given by

$$S_{wsr} = \sum_{i=1}^{D} \left(sgn\left(y_i^1 - y_i^2\right) \cdot R_i \right) \tag{15}$$

where D is the number of testing days, sgn is the sign function, y_i^1 and y_i^2 denote the hit rate on the i-th testing day from model 1 and model 2 at a chosen length coverage, respectively. R_i is the rank of the i-th difference $y_i^2 - y_i^1$. After obtaining the statistic S_{wsr}, we can get the corresponding p-value using a single tailed lookup and determine whether the difference is statistically significant or not.

4.3 GLDNet implementation

In the GLDNet model, the hyper-parameters primarily include the number of GRLs L, the number of hidden units n^l in each GRL, the number of LDN layers K, and the number of hidden units m^k in each LDN layer. For model training, we also need to predefine the weighted loss function parameter ρ (Eq. (14)). This paper employs the grid search approach to determine the optimal hyper-parameters of GLDNet by changing one of the hyper-parameters while keep the others unchanged. Grid search is the simplest possible way to get good hyperparameters of deep learning model and it has been widely used in many deep learning models (Zhang et al., 2019, Huang et al., 2014). The pro is that it can be easily parallelised. In the grid search, the hyper-parameter L is chosen from a range of [1, 4]; K is selected from one to seven; n^l and m^k are chosen from [2, 4, 8, 16]. The parameter ρ in the loss function is selected between 0.005 to 0.035 with the step of 0.005. For simplicity, the hidden unit number n^l in each GRL layer is set to be the same, so is the hidden unit number m^k in each LDN layer.

The GLDNet was implemented using the GPU-version TensorFlow 1.2 (Abadi et al., 2016). In model training, we set the batch size to be 50 and run 30 epochs (enough for convergence) with the Adam optimizer (Kingma and Ba, 2014). The learning rate is initialised to be 0.03 and then exponentially decays per 20 batches with a decay rate of 0.9. After the grid search, we adopt the cascade of two GRL layers, each of which has eight units, and five LDN layers, each of which has four hidden units. The GLDNet for the three crime forecasting tasks has the same configurations, but the model is trained separately for different crime types. The loss function parameter ρ is 0.005, 0.010 and 0.025, respectively, for burglary, assault and theft forecasting. More discussions on the hyper-parameter calibration are provided in the Section 4.5.

4.4 Results comparison

The proposed GLDNet model is compared with network-time kernel density estimation (NTKDE) proposed by (Rosser et al., 2017), which is the state-of-the-art statistical method for network-based predictive mapping. And a software using NTKDE has been deployed in Metropolitan Police Service for street network-based crime predictive mapping in London, UK (Cheng et al., 2016). The approach to determine the parameters of NTKDE can be referred to (Rosser et al., 2017). Due to the high sparsity of the data, existing spatio-temporal DL models, such as those for network-based flow prediction (Zhang et al., 2019, Yu et al., 2017, Ren et al., 2019b), cannot function in this case. Thus, the comparison experiment does not include other DL models.

Figure 7 presents the mean hit rate of GLDNet and NTKDE, averaged over the 100 testing days. Results show the GLDNet performs better than the NTKDE in all cases. For burglaries, assaults, and thefts, the GLDNet outperforms the NTKDE by 12% or more at 10% length coverage, and by around 25% at 20% length coverage. Comparing the three cases, the mean hit rate of burglaries is slightly lower than the other two crimes, which is probably because burglaries are sparser than assaults and thefts, leading to higher stochasticity. Overall, except for the minor decrease in the predictive performance of burglaries, the mean hit rate of GLDNet across different crime types is quite similar (around 40% and 60% at 10% and 20% coverage, respectively), which illustrates its consistent performance. In Figure 7, the background shading areas indicate the coverage levels where the daily hit rate of GLDNet is significantly higher than the NTKDE approach at the 5% significance level, evaluated using the WRS test. We can observe significant differences between the two results at the majority of coverage levels for three crime types.

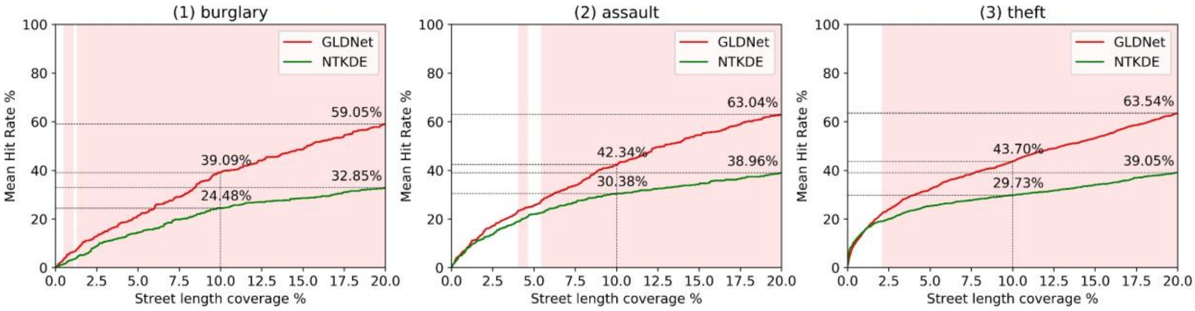

Figure 7. Plot of mean hit rate against street length coverage level for burglary, assault, and theft crimes in South Chicago for GLDNet and NTKED. Shaded regions indicate the coverage levels for which there is a

significant difference between the GLDNet and NTKDE results at the 5% significance level, evaluated using the WRS test.

The above analysis is based on mean hit rate. As we find that the sparsity of crimes may affect the performance of the approaches, we also examine the performance difference in hit rate corresponding to the level of daily crime count. We split 100-day data into equal-sized groups by ranking daily crime count. Figure 8 presents the performance of GLDNet and NTKDE in terms of hit rate by level of daily crime count. Consistent improvements in hit rate are observed in all level of daily crime counts for burglaries, assaults, and thefts. The improvements of hit rate at the length coverage of 10% or more are statistically significant in all cases, with the exception of the case shown in Figure 8d. Additionally, it is worth noting that the hit rate of NTKDE in the sparsest case (Figure 8a) notably decreases by around 50%. Instead, the proposed GLDNet can still yield satisfying results. This highlights the advantage of the proposed approach.

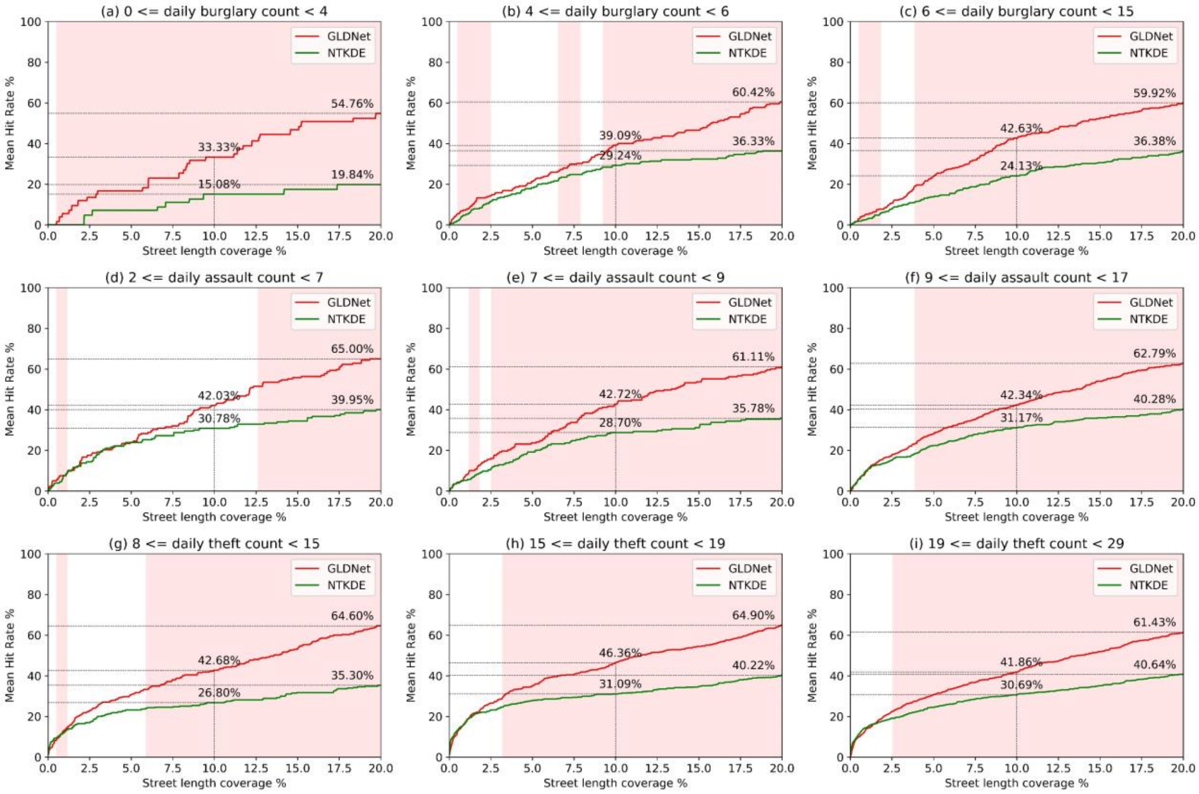

Figure 8. The plot of mean hit rate against street length coverage level for burglary, assault, and theft crimes in the South Chicago for GLDNet and NTKED at different daily crime count level (illustrated in each title). Shaded regions indicate the coverage levels for which there is a significant difference between the GLDNet and NTKDE results at 5% significance level, evaluated using WRS test.

In addition, to gain greater insight into the relative performance difference between the two methods, we also employ the relative daily difference in hit rate over the 100 days of test data at a chosen coverage level to compare the two methods. The relative improvement is defined as the absolute difference in the daily hit rates at a given coverage level between the two approaches, divided by the mean hit rate of NTKDE over all testing days. The positive values indicate the GLDNet outperform NTKDE on a given day, and vice versa. The results are illustrated in Figure 9 at three different coverage levels (5%, 10% and 20%). Overall, the mean values of the relative improvement at the three coverage levels are, respectively, 0.48, 0.60, and 0.80 for burglaries, 0.17, 0.40, and 0.62 for assaults, 0.28, 0.46, and 0.63 for thefts. A consistent increase of the relative improvement of predictive accuracy can be observed with an increase in the coverage level during the testing days.

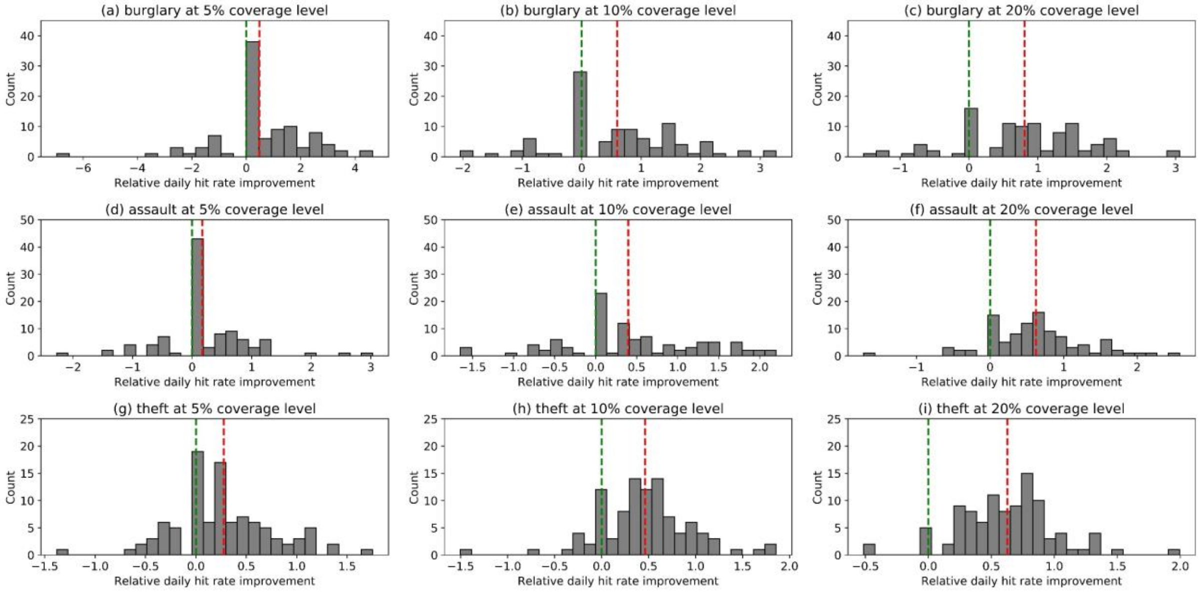

Figure 9. Histograms showing the relative improvement of predictive accuracy of GLDNet over NTKDE at three different coverage levels (5%, 10%, and 20%). The relative improvement is given by the difference in daily hit rate between the two approaches divided by the mean NTKDE hit rate. Positive values indicate that the proposed GLDNet performs better. The red dashed line indicates the mean value of the difference in each case. The green dashed line is the zero-line using as a reference.

For a more direct and concrete comparison, the predictive mapping results of the three crime types for the two approaches on the last testing day (31st December 2017) are selected as an example and visualised in Figure 10. Comparing the left and right panels, Figure 10 demonstrates several important differences between GLDNet and NTKDE. First, more crimes can be captured by the risk map at 10% coverage level of GLDNet than that of NTKDE. Second, the spatial distribution of the top-ranked street segments in GLDNet's heatmap is more concentrated than NTKDE's. It means GLDNet can generate the prediction results of higher compactness (connectedness) than NTKDE. Although compactness is not the primary objective of the prediction methods considered here, it is a useful characteristic in practical applications (Adepeju et al., 2016). This is because more compact and connected high risk locations are easier to patrol with limited police resourcing. Additionally, comparing the panels across rows in Figure 10, it illustrates that although the spatial patterns and the daily count of the three crime types are different, our approach performs well in all cases. It shows GLDNet can function on a wide variety of scenarios, irrespective of the specific spatio-temporal patterns within the data.

In summary, the proposed GLDNet performs better than NTKDE, which is probably owing to the deep structure and its strong ability of spatio-temporal dependency modelling without the assumptions about the specific statistical distribution form of the temporal and spatial propagations.

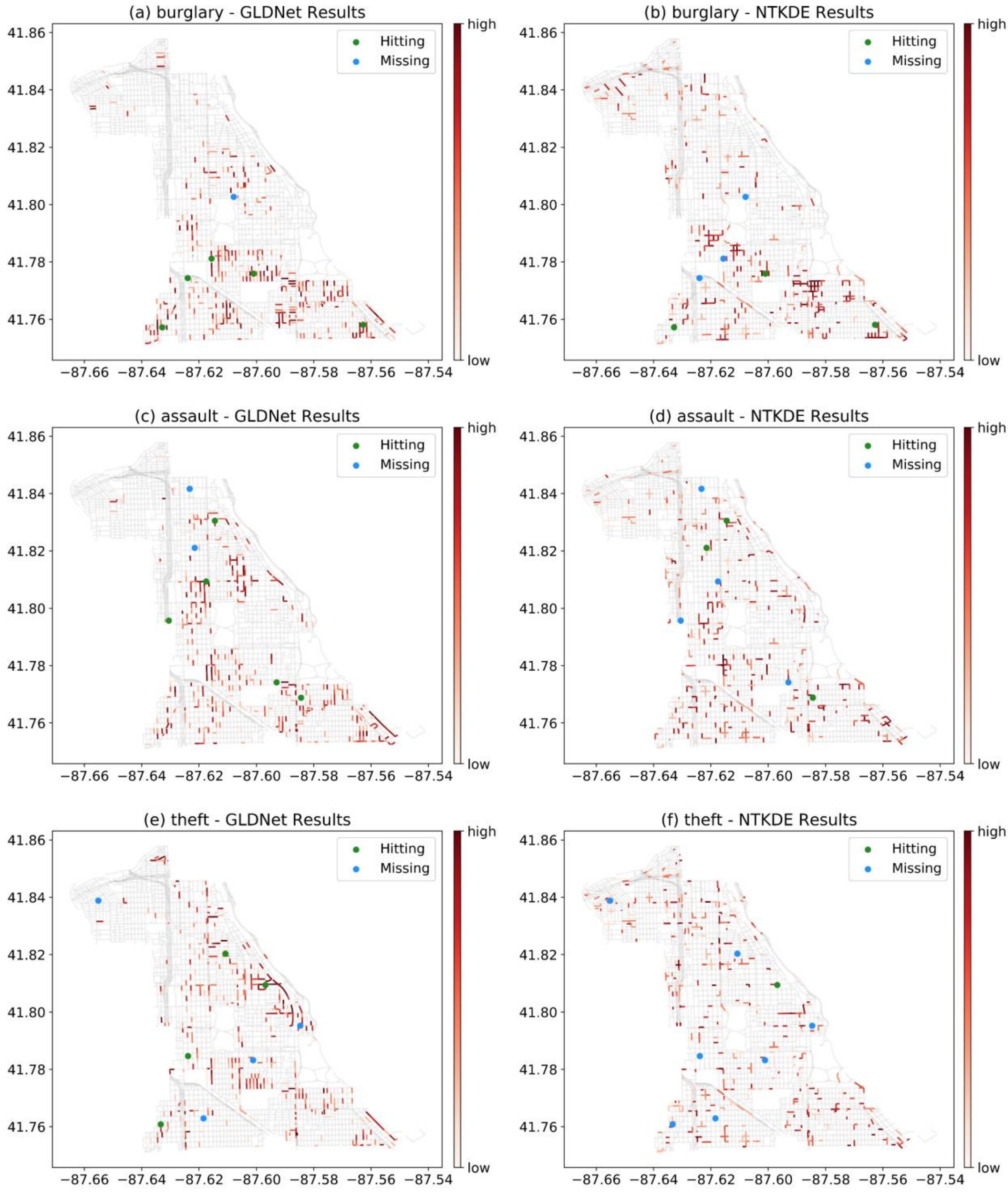
(a) burglary - GLDNet Results
(b) burglary - NTKDE Results
(c) assault - GLDNet Results
(d) assault - NTKDE Results
(e) theft - GLDNet Results
(f) theft - NTKDE Results
Hitting
Missing
high
low
41.86
41.84
41.82
41.80
41.78
41.76
−87.66
−87.64
−87.62
−87.60
−87.58
−87.56
−87.54

Figure 10. Network-based predictive mapping of three types of crimes, which shows the predicted risk on the last testing day (31st December 2017). In each subfigure, we just highlight street segments in red at 10% length coverage level. The darkness of red is proportional to the predicted probability of crime occurrence. The green dots indicate the crimes captured by the street segments in the top 10% length coverage and the blue dots are the non-captured crimes.

4.5 Sensitivity analysis

In this section, we conduct the sensitivity analysis on GLDNet, where four kinds of parameters are investigated, including the length of time window M, the number of GRL layers L, the number of LDN layers K and the parameter ρ in the loss function. Figure 11 presents the change of the mean hit rate at 10% and 20% coverage level by crime type and the parameter of interest.

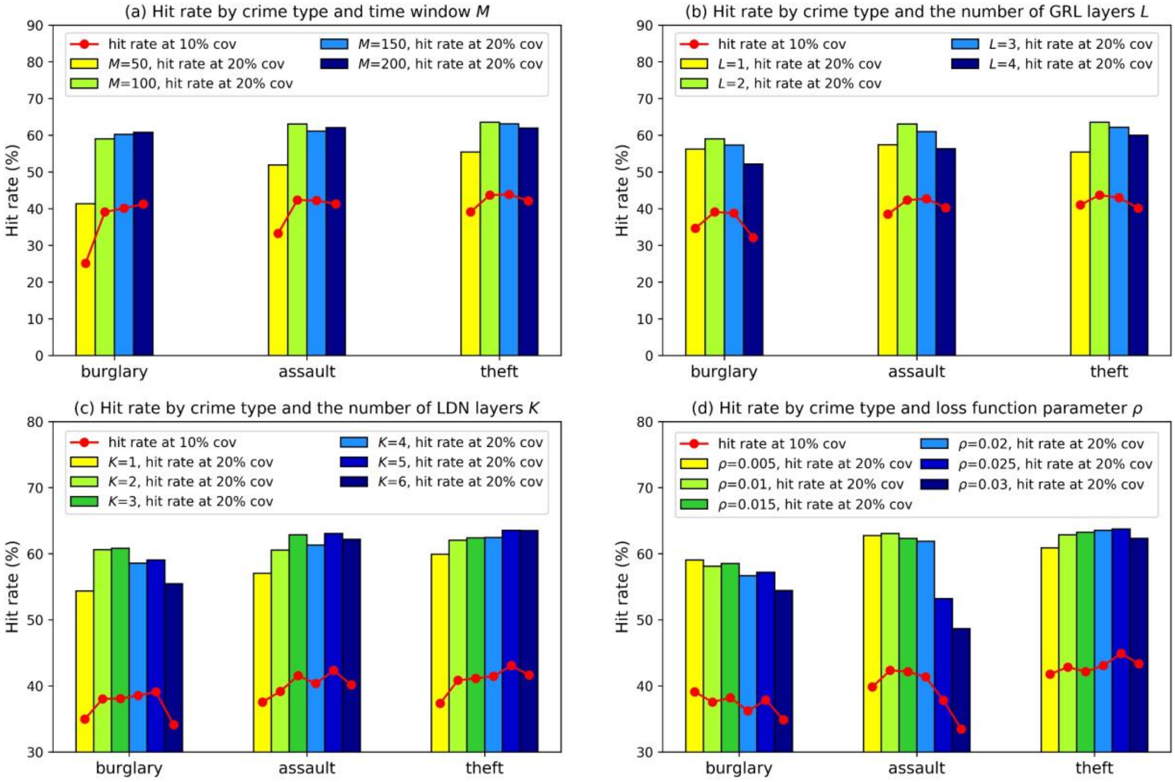

Figure 11. Sensitivity analysis of GLDNet. In each panel, the red line indicates the mean hit rate at 10% coverage level changing with the parameter of interest for each crime type while the bar plot shows the mean hit rate at 20% coverage level by crime type and the parameter of interest.

In terms of the time window length, Figure 11a shows that for burglary forecasting, the hit rate increases sharply when increasing the window length from 50 to 100, and it then keeps increasing but the improvement is slight. For assault and theft forecasting cases, the highest hit rates can be achieved at a time window of 100. The results show that burglary predictive mapping requires more temporal information than assault or theft forecasting, which is probably due to the higher sparsity of burglaries. Regarding the number of GRL layers L, Figure 11b demonstrates the change of the mean hit rate at the both coverage levels with this parameter is similar across crime type. The mean hit rate first rises and then slightly drops when the number of GRL layers is more than two. This is because increasing the parameter L may lead to the overfitting issues. From Figure 11c, it is found that the mean hit rates at 10% coverage level of the three crime types are the highest when the K is set to be five, which mean the spatial propagation modelling takes into account the dependency between each street and its, at most, fifth-order adjacent neighbours. However, the mean hit rate at 20% coverage level of the burglary and assault predictions achieve the highest value at $K = 3$ while that of the theft mapping peaks at $K = 5$. This difference is potentially due to the variation of the spatial pattern across crime type. As we care more about the hit rate at a lower coverage level, the number of LDN layers is set to be five. Figure 11d displays the sensitivity analysis on the loss function parameter ρ. It shows the optimal value of ρ for burglary, assault and theft forecasting is 0.005, 0.010 and 0.025, respectively. It is interesting to find that the optimal value of ρ is lower when the sparsity of the crime is higher. For example, in the assault case, the mean hit rate sharply drops when enlarging the ρ from 0.02 to 0.025. It demonstrates the imbalance regression cannot be overlooked in the sparse scenario.

5 Conclusions and future work

In this paper, we propose a novel DL framework for predictive mapping of sparse spatio-temporal data on urban street networks. In this framework, the network-based sparse spatio-temporal data are represented as graph signals on a weighted and undirected graph. Based on this representation, a GLDNet is proposed. The model integrates a gated network for temporal propagation modelling and a localised diffusion network for spatial propagation modelling considering the topology of the street network. Due to the sparsity of data, we treat this predictive mapping problem as an imbalance regression task. A weighted loss function is employed in the model training process to solve the imbalance issue. The model is validated on three crime forecasting cases using the real-world crime data provided by the City of Chicago open data portal. The proposed model outperforms the state-of-the-art benchmark NTKDE by 12% in the mean hit rate at 10% coverage level and by 25% at 20% coverage level for burglary, assault and theft forecasting. And the predictive hotspot map produced by GLDNet is more compact than NTKDE.

In the future, we intend to test our model with other sparse spatio-temporal data in other cities and countries, to show its potential application in broader fields, such as traffic accidents prediction and other emergence event forecasting. In addition, the street network is represented as a static graph corresponding to its physical topology. A dynamic graph representation could be incorporated in our framework to better model the spatial mutual influence considering the road reconstruction or urban form change. We also intend to incorporate external factors, e.g., weather and holidays, in the deep learning model to further improve its robustness.

References

Abadi, M., Barham, P., Chen, J., et al. TensorFlow: A System for Large-Scale Machine Learning. OSDI, 2016. 265-283.

Adepeju, M. 2017. *Modelling of sparse spatio-temporal point process (STPP)-An application in predictive policing*. UCL (University College London).

Adepeju, M., Rosser, G. & Cheng, T. 2016. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions-a crime case study. *International Journal of Geographical Information Science,* 30, 2133-2154.

Anbaroglu, B., Heydecker, B. & Cheng, T. 2014. Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies,* 48, 47-65.

Apte, J. S., Messier, K. P., Gani, S., et al. 2017. High-resolution air pollution mapping with Google street view cars: exploiting big data. *Environmental Science & Technology,* 51, 6999-7008.

Bowers, K. J., Johnson, S. D. & Pease, K. 2004. Prospective hot-spotting: The future of crime mapping? *British Journal of Criminology,* 44, 641-658.

Braga, A. A., Hureau, D. M. & Papachristos, A. V. 2011. The relevance of micro places to citywide robbery trends: a longitudinal analysis of robbery incidents at street corners and block faces in Boston. *Journal of Research in Crime and Delinquency,* 48, 7-32.

Britz, D. 2015. Recurrent Neural Networks Tutorial, Part 3–Backpropagation Through Time and Vanishing Gradients.

Chainey, S., Tompson, L. & Uhlig, S. 2008. The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal,* 21, 4-28.

Chen, H., Cheng, T. & Wise, S. 2017. Developing an online cooperative police patrol routing strategy. *Computers, Environment and Urban Systems,* 62, 19-29.

Chen, H., Cheng, T. & Ye, X. 2018. Designing efficient and balanced police patrol districts on an urban street network. *International Journal of Geographical Information Science*, 1-22.

Cheng, S., Lu, F., Peng, P., et al. 2018. Short-term traffic forecasting: An adaptive ST-KNN model that considers spatial heterogeneity. *Computers, Environment and Urban Systems,* 71, 186-198.

Cheng, T. & Adepeju, M. 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PloS one,* 9, e100465.

Cheng, T., Bowers, K. J., Longley, P. A., et al. 2016. CPC: Crime, Policing and Citizenship–Intelligent policing and big data. *London: UCL SpaceTime Lab.*

Cheng, T., Wang, J., Haworth, J., et al. 2014. A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling. *Geographical Analysis,* 46, 75-97.

Dauphin, Y. N. & Grangier, D. 2015. Predicting distributions with linearizing belief networks. *arXiv preprint arXiv:1511.05622.*

Davies, T. & Marchione, E. 2015. Event networks and the identification of crime pattern motifs. *PloS one,* 10, e0143638.

Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in Neural Information Processing Systems, 2016. 3844-3852.

Henaff, M., Bruna, J. & Lecun, Y. 2015. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163.*

Huang, W., Song, G., Hong, H., et al. 2014. Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning. *IEEE Transactions on Intelligent Transportation Systems,* 15**,** 2191-2201.

Ke, J., Zheng, H., Yang, H., et al. 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies,* 85**,** 591-608.

Kingma, D. P. & Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Kleinman, K., Abrams, A., Kulldorff, M., et al. 2005. A model-adjusted space–time scan statistic with an application to syndromic surveillance. *Epidemiology & Infection,* 133**,** 409-419.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence,* 5**,** 221-232.

Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *nature,* 521**,** 436.

Leitner, M., Glasner, P. & Kounadi, O. 2018. Laws of Geography. Oxford University Press.

Levin, D. A. & Peres, Y. 2017. *Markov chains and mixing times*, American Mathematical Soc.

Li, Y., Yu, R., Shahabi, C., et al. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926.*

Liao, S., Zhou, L., Di, X., et al. Large-scale short-term urban taxi demand forecasting using deep learning. Proceedings of the 23rd Asia and South Pacific Design Automation Conference, 2018. IEEE Press, 428-433.

Lin, Y.-L., Yen, M.-F. & Yu, L.-C. 2018. Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information,* 7**,** 298.

Liu, Q., Wang, B. & Zhu, Y. 2018. Short-Term Traffic Speed Forecasting Based on Attention Convolutional Neural Network for Arterials. *Computer-Aided Civil and Infrastructure Engineering.*

Louizos, C., Welling, M. & Kingma, D. P. 2017. Learning Sparse Neural Networks through $ L_0 $ Regularization. *arXiv preprint arXiv:1712.01312.*

Mhaskar, H. N. & Poggio, T. 2016. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications,* 14**,** 829-848.

Miaou, S.-P., Song, J. J. & Mallick, B. K. 2003. Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics,* 6.

Mohler, G. 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting,* 30, 491-497.

Mohler, G., Short, M. B., Brantingham, P. J., et al. 2011. Self-exciting point process modeling of crime. *Journal of the American Statistical Association,* 106, 100-108.

Porter, M. D. & Reich, B. J. 2012. Evaluating temporally weighted kernel density methods for predicting the next event location in a series. *Annals of GIS,* 18, 225-240.

Ratcliffe, J. 2010. Crime mapping: spatial and temporal challenges. *Handbook of quantitative criminology.* Springer.

Ren, Y., Chen, H., Han, Y., et al. 2019a. A hybrid integrated deep learning model for the prediction of citywide spatio-temporal flow volumes. *International Journal of Geographical Information Science*, 1-22.

Ren, Y., Cheng, T. & Zhang, Y. 2019b. Deep spatio-temporal residual neural networks for road-network-based data modeling. *International Journal of Geographical Information Science,* 33, 1894-1912.

Rosser, G. & Cheng, T. 2016. Improving the Robustness and Accuracy of Crime Prediction with the Self-Exciting Point Process Through Isotropic Triggering. *Applied Spatial Analysis and Policy,* 10.1007/s12061-016-9198-y.

Rosser, G., Davies, T., Bowers, K. J., et al. 2017. Predictive Crime Mapping: Arbitrary Grids or Street Networks? *Journal of Quantitative Criminology,* 33, 569-594.

Santini, M. & Valentini, R. 2011. Predicting hot-spots of land use changes in Italy by ensemble forecasting. *Regional Environmental Change,* 11, 483-502.

Shiode, S. & Shiode, N. 2014. Microscale Prediction of Near-Future Crime Concentrations with Street-Level Geosurveillance. *Geographical Analysis,* 46, 435-455.

Shuman, D. I., Narang, S. K., Frossard, P., et al. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine,* 30, 83-98.

Sullivan, G. M. & Feinn, R. 2012. Using Effect Size-or Why the P Value Is Not Enough. *Journal of graduate medical education,* 4, 279-282.

Takahashi, K., Kulldorff, M., Tango, T., et al. 2008. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics,* 7, 14.

Tang, L., Kan, Z., Zhang, X., et al. 2016. A network Kernel Density Estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science,* 30, 1717-1737.

Teng, S.-H. 2016. Scalable Algorithms for Data and Network Analysis. *Foundations and Trends® in Theoretical Computer Science,* 12, 1-274.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic geography,* 46, 234-240.

Ugarte, M. D., Adin, A., Goicoa, T., et al. 2014. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical methods in medical research,* 23, 507-530.

Vere-Jones, D. 2009. Some models and procedures for space-time point processes. *Environmental and Ecological Statistics,* 16, 173-195.

Wang, B., Yin, P., Bertozzi, A. L., et al. 2017a. Deep Learning for Real-Time Crime Forecasting and its Ternarization. *arXiv preprint arXiv:1711.08833.*

Wang, B., Zhang, D., Zhang, D., et al. 2017b. Deep learning for real time crime forecasting. *arXiv preprint arXiv:1707.03340.*

Weisburd, D., Bernasco, W. & Bruinsma, G. 2009. Putting crime in its place: units of analysis in spatial crime research. New York: Springer.

Xie, Z. & Yan, J. 2008. Kernel density estimation of traffic accidents in a network space. *Computers, environment and urban systems,* 32, 396-406.

Yannis, G., Dragomanovits, A., Laiou, A., et al. Road traffic accident prediction modelling: a literature review. Proceedings of the Institution of Civil Engineers-Transport, 2017. Thomas Telford Ltd, 245-254.

Yu, B., Yin, H. & Zhu, Z. 2017. Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting. *arXiv preprint arXiv:1709.04875.*

Zhang, Y., Cheng, T. & Ren, Y. 2019. A graph deep learning methods for short-term traffic forecasting on large road newtorks. *Computer-Aided Civil and Infrastructure Engineering,* 10.1111/mice.12450, 1-20.

Zhou, D. & Schölkopf, B. A regularization framework for learning from graph data. ICML workshop on statistical relational learning and Its connections to other fields, 2004. 67-8.

Zhu, D., Wang, N., Wu, L., et al. 2017. Street as a big geo-data assembly and analysis unit in urban studies: A case study using Beijing taxi data. *Applied Geography,* 86, 152-164.