Accident Analysis and Prevention 122 (2019) 239–254

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data

Jie Bao[a,b,c], Pan Liu[a,b], Satish V. Ukkusuri[c,*]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing, 210096, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing, 210096, China
[c] Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, 47906 IN, United States

ARTICLE INFO

Keywords:
Multi-source data
Spatiotemporal
Crash risk prediction
Deep learning

ABSTRACT

The primary objective of this study is to investigate how the deep learning approach contributes to citywide short-term crash risk prediction by leveraging multi-source datasets. This study uses data collected from Manhattan in New York City to illustrate the procedure. The following multiple datasets are collected: crash data, large-scale taxi GPS data, road network attributes, land use features, population data and weather data. A spatiotemporal convolutional long short-term memory network (STCL-Net) is proposed for predicting the citywide short-term crash risk. A total of nine prediction tasks are conducted and compared, including weekly, daily and hourly models with 8×3, 15×5 and 30×10 grids, respectively. The results suggest that the prediction performance of the proposed model decreases as the spatiotemporal resolution of prediction task increases. Moreover, four commonly-used econometric models, and four state-of-the-art machine-learning models are selected as benchmark methods to compare with the proposed STCL-Net for all the crash risk prediction tasks. The comparative analyses suggest that in general the proposed STCL-Net outperforms the benchmark methods for different crash risk prediction tasks in terms of higher prediction accuracy rate and lower false alarm rate. The results verify that the proposed spatiotemporal deep learning approach performs better at capturing the spatiotemporal characteristics for the citywide short-term crash risk prediction. In addition, the comparative analyses also reveal that econometric models perform better than machine-learning models in weekly crash risk prediction tasks, while they exhibit worse results than machine-learning models in daily crash risk prediction tasks. The results can potentially guide transportation safety engineers to select appropriate methods for different crash risk prediction tasks.

1. Introduction

During the past decades, considerable efforts have been devoted to analyzing crash data at various spatially aggregated levels, such as states (Noland, 2003), counties (Huang et al., 2010; Li et al., 2013), traffic analysis zones (TAZ) (Rhee et al., 2016; Bao et al., 2017), and grid-based spatial units (Xie et al., 2017). The spatial analysis of crashes has become more and more prevalent because researchers have come to believe that traffic safety is an essential component of urban transportation planning (FHWA, 2005; NCHRP, 2010). In addition, with a better understanding of the spatial pattern of crashes, transportation authorities can identify the regions with higher crash risks, and apply proactive countermeasures to these regions to enhance safety more efficiently.

In the past, the spatial analysis of crashes were usually conducted within a long-term period, such as one year or more. Thus, the analysis results cannot reveal the dynamic change of crash risk in each region within a short-term period. In this study, the short-term crash risk includes weekly, daily and hourly crash risk, respectively. With effective short-term prediction of crash risk in a certain region, transportation authorities can proactively allocate police forces, and provide a dynamic crash risk map to drivers for selecting a safer route. So far, many previous studies of short-term crash risk prediction were conducted on freeways and expressways at corridor-level (Abdel-Aty et al., 2012; Shi and Abdel-Aty, 2015), while very few studies focused on citywide short-term crash risk prediction at spatially-aggregated level. The reasons are mainly twofold: (a) the lack of effective and sufficient sensors for short-term traffic data collection within the entire urban areas; (b) when predicting the short-term crash risk in each region of urban areas, the spatial and temporal dependencies among explanatory variables are usually complex, non-linear and hierarchical. In this condition,

* Corresponding author.
E-mail addresses: baojie@seu.edu.cn (J. Bao), pan_liu@hotmail.com (P. Liu), sukkusur@purdue.edu (S.V. Ukkusuri).

https://doi.org/10.1016/j.aap.2018.10.015
Received 31 July 2018; Received in revised form 1 October 2018; Accepted 21 October 2018
Available online 01 November 2018
0001-4575/

traditional econometric models may fail to capture the spatiotemporal features of crash risk in each region, leading to lower prediction performance.

Recently, the availability of a large datasets associated with human activities in urban areas, resulting in a surge of studies on human mobility (Bao et al., 2017; González et al., 2008; Hasan et al., 2013). Transportation systems can greatly benefit from multi-sourced big data in the areas such as traffic flow prediction and travel demand estimation. Theoretically, big data also has potential to be incorporated in traffic safety studies to help transportation professionals better understand the mechanism and contributing factors of crashes. In recent studies, the authors of the paper investigated how to incorporate big data into spatially aggregated crash models (Bao et al., 2017, 2018).The results suggested that the human activity and mobility information extracted from social media data and large-scale taxi GPS data significantly affected the crash counts in each region. The results also confirmed the fact that the human activity and mobility information extracted from big data can serve as effective surrogate measures for traffic exposure and can greatly improve the performance of spatially aggregated crash models.

Moreover, deep learning technology have grown rapidly during the past few years in the research fields of computer vision, natural language processing, artificial intelligence, and pattern recognition (Krizhevsky et al., 2012; Deng, 2016). Some recent studies have also started applying deep learning technology into transportation fields, such as short-term traffic flow prediction (Lv et al., 2015), network travel time estimation (Hou and Edara, 2018), and travel mode inference (Dabiri and Heaslip, 2018). Compared with traditional statistical models and other learning architectures, deep learning can model complex non-linear relationships using distributed and hierarchical feature representation (Ma et al., 2015a,b), which has exhibited its superiority in short-term traffic flow and traffic speed prediction. Therefore, the emergence of big data and deep learning technology has the potential to advance the understanding of traditional transportation problems and enable the citywide short-term crash risk prediction.

The primary objective of this study is to develop a citywide short-term crash risk prediction model with multi-source data using deep learning theory. A large-scale taxi GPS data are used to depict the short-term human mobility in each region. Previous studies have suggested that taxi GPS dataset has larger sample size, and covers more age groups of travelers when compared to social media data (Chen et al., 2014). In addition, unlike cell phone data taxi GPS data are publicly available in many cities, providing researchers with an opportunity to test these research questions. To the best knowledge of the authors, this paper is one of the first attempts to employ spatiotemporal deep learning approach in citywide short-term crash risk prediction. The contributions of this paper can be summarized as follows:

(a) In the citywide short-term crash risk prediction, the crash risk in each region usually propagates temporally and spatially simultaneously (Chen et al., 2016). The proposed spatiotemporal deep learning architecture can explore both the spatial and temporal dependencies in the high dimensional set of explanatory variables through distributed and hierarchical feature extraction.
(b) Previous studies of short-term traffic flow or traffic speed prediction only consider historical traffic flow or speed as explanatory variables (Lv et al., 2015; Ma et al., 2015a,b). In contrast, crash risk prediction is a complex and nonlinear problem influenced by many fixed and stochastic factors in each region, such as weather, traffic exposure, land use and demographic characteristic. The proposed spatiotemporal deep learning architecture can integrate all these explanatory variables in an end-to-end deep learning architecture.
(c) Validated by the multi-source data collected from Manhattan, the proposed spatiotemporal deep learning approach outperforms the selected benchmark methods, including traditional econometric models and several state-of-the-art machine-learning models.
(d) This study compares the performance of econometric models and machine-learning models in different crash risk prediction tasks. In general, machine-learning models have better performance than econometric models with increased spatiotemporal resolution, which can guide transportation safety engineers to select appropriate methods for different crash risk prediction tasks.

The rest of the paper is organized as follows. Section 2 discusses the procedures for gathering various types of data from multiple data sources. Section 3 describes the structure of the proposed spatiotemporal convolutional long short-term memory network (STCL-Net), and the associated methodology for each component. Section 4 presents the results of data analysis and compares the predictive performance between the proposed approach and the benchmark models. Finally, conclusions are drawn and future research directions are indicated in section 5.

2. Data sources

This study uses the data collected from Manhattan borough of the New York City to illustrate the procedure for citywide short-term crash risk prediction. The study period is from January 1 st to December 31th, 2015. In this study, the Manhattan area is uniformly divided into different grid cells and the grid cell is considered as the basic spatial unit of crash analysis. Moreover, three different grids are studied and compared in this study: 8×3 grid size, 15×5 grid size, and 30×10 grid size (See Fig. 1). For the 15×5 and 30×10 grids, the sizes of each grid cell are close to the average size of zip code tabulation areas (ZCTA) and census tracts in Manhattan respectively, which are two commonly used spatial analysis units in previous studies (Bao et al., 2018; Ukkusuri et al., 2011).

The following six types of data are collected: crash data, taxi trip data, road network attributes, land use features, population data and weather data. The data are collected from multiple data sources. More specifically, the crash data is collected from the New York City Police Department (NYPD). The information obtained from the crash data includes the date, time, severity, collision type, and geo-location of each crash. A total of 53,354 crashes, including 45,636 property damage only (PDO) crashes, 7685 injury crashes and 33 fatal crashes, were reported during the selected time period in the study area. Considering that the crash risk was evaluated by both frequency and severity in many previous studies of hotspot identification (Xie et al., 2017; Chen et al., 2016; Ren et al., 2017), we define the crash risk of a grid cell as the sum of the severity level of all crashes occurred in that grid cell. For example, in this study the severity levels of PDO, injury and fatal crashes are defined as 1, 2 and 3, respectively. Accordingly, the crash risk of a grid cell is 5 if one fatal and two PDO crashes occurred in that grid cell.

The taxi GPS data is collected from the New York City Taxi & Limousine Commission (NYCTLC). The taxicabs of New York have two varieties: yellow and green. The taxis painted yellow can pick up passengers anywhere in the New York City, while the taxis painted green are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. To ensure that the taxi GPS data fully cover the whole study area, we collect the GPS data for both yellow and green taxis.

For each taxi trip the following information are extracted from the taxi GPS dataset: pick-up timestamp, pick-up geo-location, drop-off timestamp, drop-off geo-location, trip distance, and the payment information. More specifically, we follow the following three steps to extract trip information from the taxi GPS dataset. First, the taxi trips with pick-up and drop-off points within the study area are selected. Second, unreasonable trips are removed. Note that a trip is considered unreasonable if: (a) the travel distance is zero; (b) the fare is less than the starting price (2.5 dollars); (c) the duration is less than one-minute,

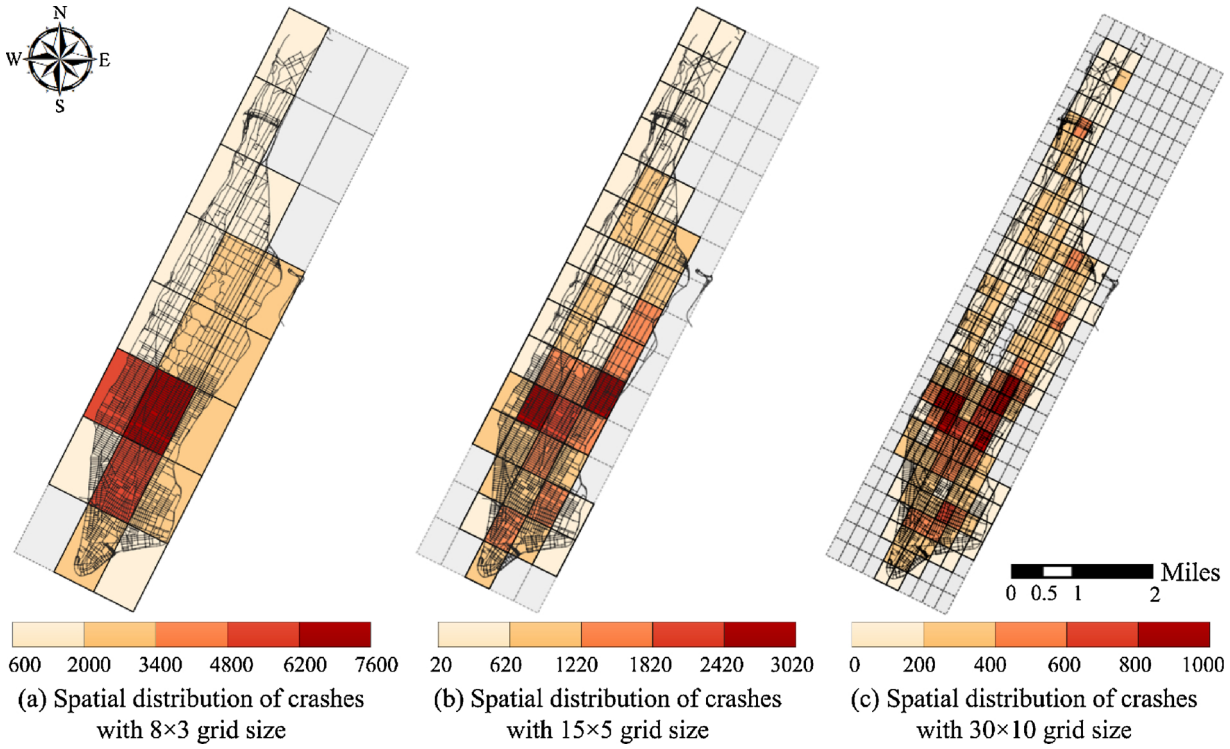


Fig. 1. Spatial distribution of crashes in Manhattan area with different grid sizes.

Table 1
Description of considered variables.

Category	Name	Description
Type I	DVKT	The daily vehicle kilometers traveled in each grid cell (10^6 vehicle. km)
	Commercial area	The ratio of the area allocated for commercial use in each grid cell
	Residential area	The ratio of the area allocated for residential use in each grid cell
	Population	The number of people in each grid cell (10^4)
	Road density	Road length / Area of each grid cell (km/km^2)
	Freeway percentage	Freeway length / total road length in each grid cell
	Arterial percentage	Arterial length / total road length in each grid cell
	Local road percentage	Local road length / total road length in each grid cell
	Intersections	The number of signalized intersections in each grid cell
Type II	Temperature	The average temperature for each week/day/hour (°F)
	Precipitation	The average precipitation for each week/day/hour (in)
	Snowfall	The average snowfall for each week/day/hour (in)
	Pressure	The average pressure for each week/day/hour (in. Hg)
	Wind speed	The average wind speed for each week/day/hour (MPH)
Type III	Crash risk	The crash risk in each grid cell for each week/day/hour
	Taxi trips	The number of taxi pick-ups and drop-offs in each grid cell for each week/day/hour (10^4)

or (d) the average speed is more than 80 miles per hour. Finally, only the trip records with both pick-up and drop-off timestamps are considered for further analyses because the timestamp information is critical for further spatial and temporal aggregation. The final dataset consists of 123,308,721 taxi trips recorded during the selected time period in the study area.

The road-network-attribute data are collected from the New York City Department of Transportation (NYCDOT) and the TIGER files of U.S. Census Bureau. An ArcGIS shape file depicting the road network attributes are obtained, and the information provided by the ArcGIS shape files include the length, road type and the intersections. Traffic volume data is collected from the New York State Department of Transportation (NYSDOT). Note that the NYSDOT only provides the average annual daily traffic (AADT) on freeways and major arterials. The daily vehicle kilometers traveled (DVKT) on freeways and major arterials is then computed for each grid cell on the basis of the AADT and the road network attributes. More specifically, we split the freeways and major arterials by the boundaries of the selected grid cells with the spatial tools provided by ArcGIS and calculate the length of each segment of the freeways and major arterials in each grid cell. The DVKT is then calculated by summarizing the products of road lengths and the AADT for different road segments.

The land use data which includes the ratio of commercial area and residential area are collected from the New York City Department of City Planning (NYCDCP). The population data is obtained from the U.S. Census Bureau. Note that the land use data and the population data are both at the census-tract level. We distribute the census-tract based data to each grid cell, weighting by the share of each census-tract's area within the grid cell's area, which is consistent with many previous studies (Noland et al., 2016; Faghih-Imani et al., 2014).

The weather data is collected from the National Climate Data Center (NCDC) website which provides daily and hourly aggregated weather information from weather stations across the United States. The obtained weather information include the average temperature, average

Table 2
Descriptive statistics of variables.

Variables	8 × 3 Grid size			15 × 5 Grid size			30 × 10 Grid size		
Type I	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
DVKT	0.032	1.028	0.641	0	0.499	0.205	0	0.201	0.051
Commercial area	0	0.842	0.329	0	1	0.346	0	1	0.307
Residential area	0.157	1	0.671	0	0.98	0.627	0	1	0.519
Population	0.268	10.964	7.093	0	7.842	2.267	0	2.76	0.555
Road density	0.503	31.561	18.835	0	34.001	18.835	0	38.265	18.835
Freeway percentage	0	0.543	0.052	0	1	0.045	0	1	0.047
Arterial percentage	0	0.328	0.088	0	0.768	0.112	0	1	0.106
Local road percentage	0.234	1	0.86	0	1	0.83	0	1	0.767
Intersections	29	2798	1510	0	1002	483	0	366	121
	Weekly time level			Daily time level			Hourly time level		
Type II	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Temperature	19.714	81.286	56.846	10	88	56.81	2	96	56.465
Precipitation	0	1.02	0.11	0	1.1	0.112	0	1.28	0.031
Snowfall	0	3.314	0.228	0	16.9	0.227	0	16.9	0.227
Pressure	29.604	30.217	29.913	29.41	30.6	29.913	29.14	30.67	29.891
Wind speed	9.286	13.571	5.305	11	16	5.299	0	21	5.241
	8 × 3 Grid size			15 × 5 Grid size			30 × 10 Grid size		
Type III	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Crash risk (Weekly)	0	217	49.054	0	100	15.695	0	35	3.924
Taxi trips (Weekly)	0.062	136.85	19.777	0	42.505	6.329	0	18.78	1.582
Crash risk (Daily)	0	48	6.975	0	27	2.232	0	14	0.558
Taxi trips (Daily)	0.004	21.898	2.815	0	6.881	0.901	0	2.941	0.225

(*continued on next page*)

precipitation, average snowfall, average pressure and the average wind speed. The considered variables in the present study are described in Table 1.

Finally, the crash risk, the taxi trip data, the DVKT, the land use data, the population data, the road network related data and the weather data are aggregated into corresponding grid cells with different spatiotemporal resolution by the software PostgreSQL and PostGIS. The descriptive statistics of the variables are summarized in Table 2.

3. Methodology

In this section, we construct a spatiotemporal convolutional long short-term memory network (STCL-Net) for predicting the citywide short-term crash risk. The proposed STCL-Net includes three components: the convolutional neural network (CNN), the long short-term memory (LSTM) neural network, and the convolutional long short-term memory neural network (ConvLSTM). The methods used in each component are briefly discussed in this section.

3.1. The structure of STCL-Net

As shown in Table 1, three different types of variables are incorporated in the short-term crash risk prediction model for each grid cell in urban areas. More specifically,

(a) Type I variables: this type of variables are *only spatially varied but temporally static* during the study period, mainly due to the relatively long time period of data collection. For this type of variables, the local spatial dependencies among grid cells should be addressed in the crash risk prediction model. For example, the DVKT of nearby grid cells will have greater influences on the crash risk of a specific location than distant grid cells.
(b) Type II variables: this type of variables are *only temporally varied but spatially static* during the study period, which refer to weather variables in this study. For this type of variables, the strong periodicity and temporal dependencies should be captured in the crash risk prediction model.
(c) Type III variables: this type of variables are *both spatially and temporally varied* during the study period. For example, the crash risk and the number of taxi trips in each grid cell change dynamically. For this type of variables, there exists both local spatial dependencies and temporal dependencies, which should be considered simultaneously in the crash risk prediction model.

Previous researchers have proposed numerous deep learning architectures to deal with the spatial and temporal dependencies among variables. For example, Zhu et al. employed the convolutional neural network (CNN) to capture the spatial dependencies and nonlinear traffic dynamics for network-level traffic incident detection (Zhu et al., 2018). Ma et al. applied the long short-term memory (LSTM) neural network to capture the temporal dependency for short-term traffic speed prediction using remote microwave sensor data (Ma et al., 2015a,b). Shi et al. proposed a convolutional LSTM neural network by innovatively integrating CNN and LSTM in one deep learning architecture (Shi et al., 2015). The proposed model can address the issue of spatial and temporal dependence simultaneously, and achieve remarkable prediction performance in spatiotemporal sequence prediction problems. In this study, to integrate the spatial and temporal dependencies of the three types of variables, we propose a spatiotemporal convolutional long short-term memory network (STCL-Net) for citywide short-term crash risk prediction. Fig. 2 illustrates the structure of proposed STCL-Net. More specifically, stacked CNN layers are developed to capture the local spatial dependencies and extract the spatial features among type I variables; stacked LSTM layers are developed to capture the temporal dependencies and extract the temporal features among type II variables; and stacked ConvLSTM layers are developed to capture the spatiotemporal features among type III variables. All the extracted high-level features from the three components are further merged together and input into multiple fully-connected layers to generate the final predicted crash risk map. The methods used in each

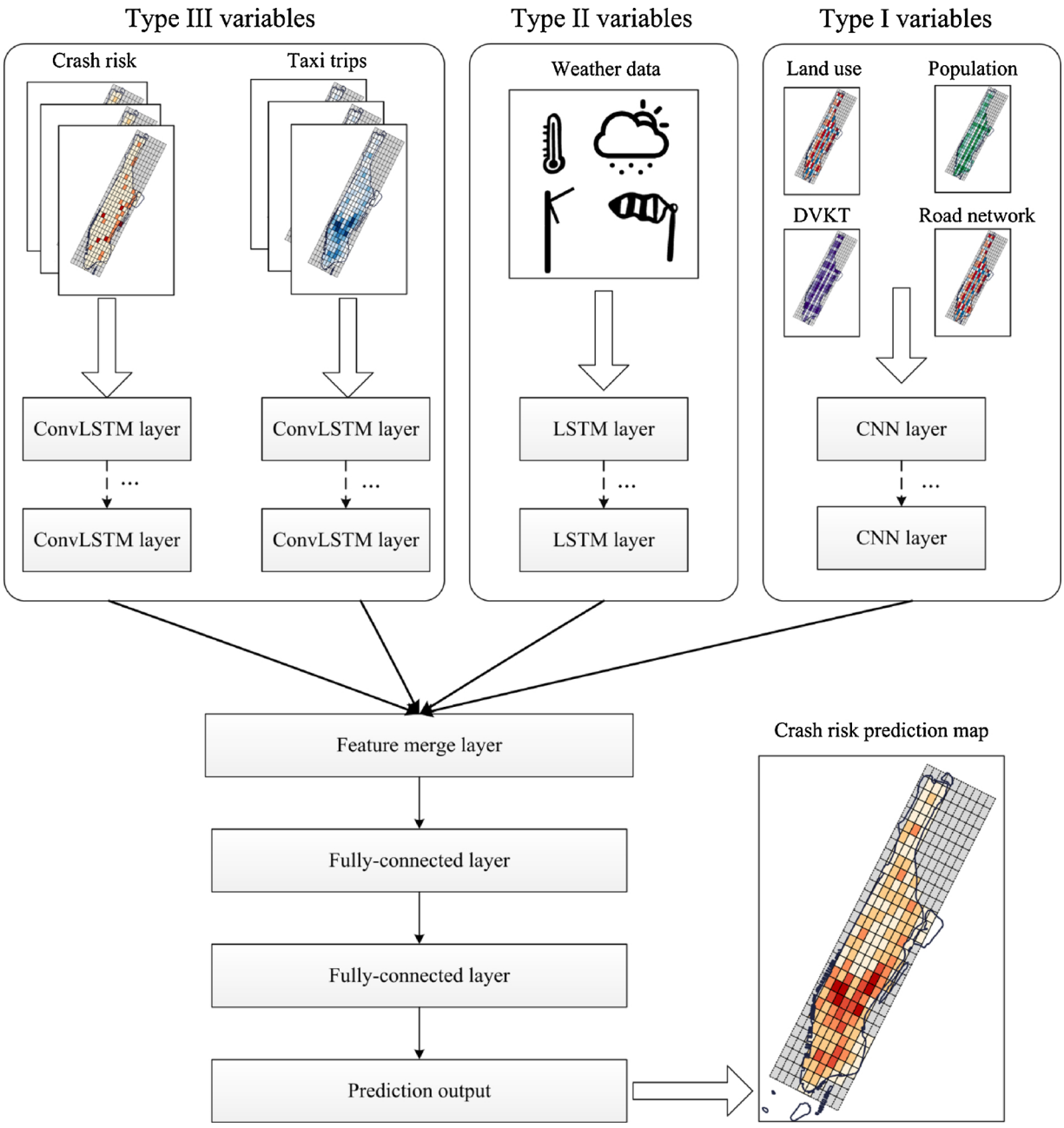


Fig. 2. The structure of proposed STCL-Net.

component are briefly explained as follows.

3.2. Spatial features extracted from CNN

In this study, nine CNNs are developed to extract the spatial features for type I variables respectively such as population, DVKT, land use, and road network variables (See Fig. 3). For each CNN, the input is a feature map of grid cells with value of the associated variable. The spatial dependencies extraction are performed mainly by the convolutional layer and pooling layer, which are the two most important layers in CNN.

In the population CNN, the convolutional layer consists of a set of learnable filters, and is connected to a small patch of the input population feature map locally. Then, the filter slides across the width and height of the input population feature map, and computes the dot products between the weights in the filter and the associated patch locally. Let X_{ij} represents the population value in grid cell (i, j) of the associated patch, then the k-th output feature map can be calculated by:

$$O^k = f\left(\sum_k W_k * X_{ij} + b_k\right) \tag{1}$$

Where W_k and b_k (k = 1, 2, 3,…K) indicates the weights and biases, K indicates the depth of the filter. * indicates the convolution operation, and f (·) indicates the nonlinear active function. In this study, the *Rectified Linear Units* (ReLU) function was used. Compared with other active functions like *tanh* (·) and *sigmoid* function, the learning rate of CNN with ReLU is much faster (Krizhevsky et al., 2012).

The pooling layer is then connected to the convolutional layer for reducing the spatial size of input population feature map, avoiding the problem of overfitting during model training (Scherer et al., 2010). Max pooling is the most commonly used pooling method, which computes the maximum value of neighbouring groups of grid cells in the same patch. The spatial dependencies in the population feature map are then extracted through a series of convolution layers and pooling layers.

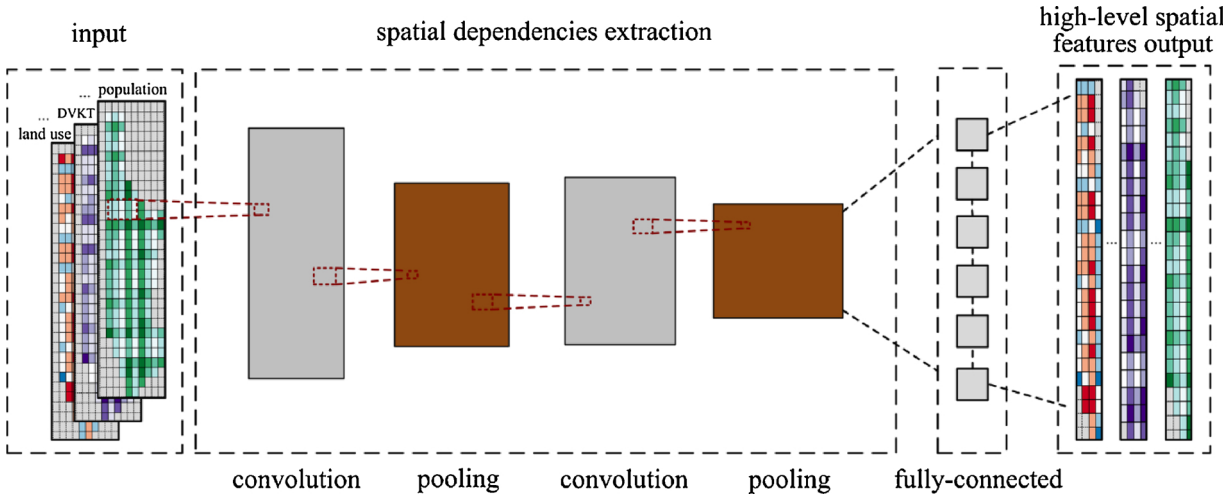


Fig. 3. Spatial features captured by CNN.

Finally, the captured spatial features of population variables are concatenated into a dense vector as the output. Similar process are conducted for land use, DVKT and road network variables, respectively.

3.3. Temporal features extracted from LSTM

For data with time-series characteristics, CNN usually fails to capture the temporal dependencies. In this context, the recurrent neural network (RNN) was proposed, where the connection between units is structured by timestamps (Van Lint et al., 2002). Traditional RNNs have exhibited a superior capability of handling nonlinear time series problems. However, when the time steps become large, traditional RNNs usually suffer from problems of gradient vanishing or gradient exploding (Ma et al., 2015a,b). To overcome these drawbacks, LSTM neural network was developed by introducing three kinds of gates into traditional RNNs (Hochreiter and Schmidhuber, 1997). Compared with traditional RNNs, LSTM neural network is capable to learn the time series with long time spans and automatically determine the optimal time lags for prediction.

The structure of LSTM neural network is shown in Fig. 4. Each LSTM cell maps the input vector sequence $\boldsymbol{x}$ to a hidden vector sequence $\boldsymbol{h}$ by T iterations. In the context of short-term crash risk prediction, $\boldsymbol{x}$ are considered as input weather data (e.g. temperature, precipitation, snowfall, pressure and wind speed), and $\boldsymbol{h}$ is the estimated value. The LSTM contains three gates: input gate, forget gate, and output gate, which are used to decide whether to add or remove historical weather information to a cell state. For each time step, the three gates will be iteratively calculated by Eqs (2)–(6).

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ci} \odot \boldsymbol{c}_{t-1} + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cf} \odot \boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{co} \odot \boldsymbol{c}_t + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t) \tag{6}$$

Where $\boldsymbol{W}_{xi}$ indicates the weights matrix between the input weather variables and the output of the input gate, and $\boldsymbol{b}_i$ indicates the bias of input gate. Similarly, $\boldsymbol{W}_{hi}$, $\boldsymbol{W}_{ci}$, $\boldsymbol{W}_{xf}$, $\boldsymbol{W}_{hf}$, $\boldsymbol{W}_{cf}$, $\boldsymbol{W}_{xc}$, $\boldsymbol{W}_{hc}$, $\boldsymbol{W}_{xo}$, $\boldsymbol{W}_{ho}$, $\boldsymbol{W}_{co}$ indicate the weights matrices which conduct a linear transformation from the vector of the first subscript to the second subscript, while $\boldsymbol{b}_f$, $\boldsymbol{b}_c$, $\boldsymbol{b}_o$ indicate the associated biases. ⊙ indicates the Hadamard product, which calculates the element-wise products of two vectors, matrices, or tensors with the same dimensions. σ (·) and *tanh* (·) are the two commonly-used nonlinear active functions. In addition, multiple LSTM layers are stacked as a deeper and more complicated neural network for

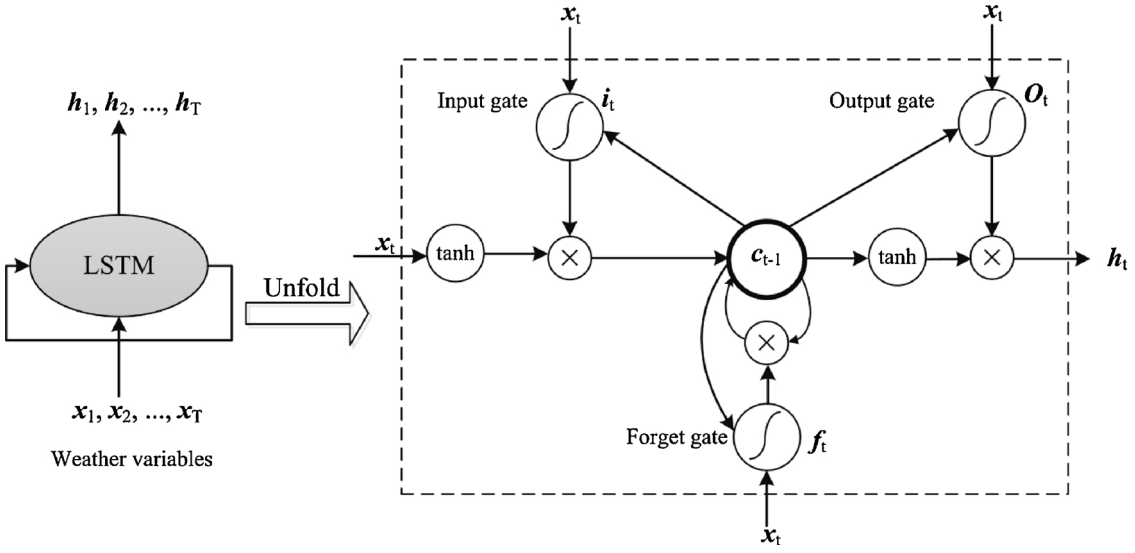


Fig. 4. Temporal features captured by LSTM.

Table 3
The confusion matrix.

	Predicted condition: Non-high crash risk	**Predicted condition:** High crash risk
True condition: Non-high crash risk	True Negative (TN)	False Positive (FP)
True condition: High crash risk	False Negative (FN)	True Positive (TP)

better exploring the temporal dependencies among weather data in this study.

3.4. Spatiotemporal features extracted from ConvLSTM

In the short-term crash risk prediction, the crash risk and taxi trip variables change both spatially and temporally. In this condition, traditional LSTM neural networks can only extract the temporal dependencies while fail to extract the spatial dependencies among those variables simultaneously. To address this issue, the convolutional LSTM neural network (ConvLSTM) was proposed by extending traditional LSTM to have convolutional structures in both input-to-state and state-to-state transitions (Shi et al., 2015). The core idea of ConvLSTM is transforming all the inputs, hidden states, cell outputs, and various gates of traditional LSTM to 3D tensors, the calculation of each gate can be found in detail in Shi et al (2015). Similar to traditional LSTM, multiple ConvLSTM layers are stacked to better capture spatiotemporal features among crash risk or taxi trips data in this study.

3.5. Feature merge layer and fully-connected layer

The spatial features extracted from CNN, the temporal features extracted from LSTM neural network, and the spatiotemporal features extracted from ConvLSTM neural network are concatenated into one dense vector in the feature merge layer, which represents the most high-level features of the input transportation network. Finally, the vector is transformed into model outputs through a couple of fully-connected layers. Accordingly, the prediction output can be calculated by:

$$\widehat{y_t} = W_{cnn}X_t^{cnn} + W_{lstm}X_t^{lstm} + W_{convlstm}X_t^{convlstm} + b_t \tag{7}$$

Where X_t^{cnn}, X_t^{lstm}, $X_t^{convlstm}$ indicate the extracted features by CNN, LSTM and ConvLSTM layers at t time step, respectively. W_{cnn}, W_{lstm}, $W_{convlstm}$ and b_t indicate the associated weights and bias.

3.6. Objective function

During the training process of STCL-Net, the objective is to minimize the mean squared error between the estimated and real crash risk in each grid cell. To avoid the overfitting issues, the L2-norm regularization is applied and the objective function is given as:

$$\min_{w,b} \| y_t - \widehat{y_t} \|_2^2 + \alpha \| W \|_2^2 \tag{8}$$

Where W indicates all the weights in Eq. (7), and α indicates the regularization parameter which balances the bias-variance tradeoff.

3.7. Evaluation metrics

To comprehensively evaluate the performance of the proposed short-term crash risk prediction

model, three measures are used: mean squared error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The three measures can be calculated as follows:

$$MSE = \frac{1}{n_p}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{(i,j)} - \widehat{y}_{(i,j)})^2 \tag{9}$$

$$MAE = \frac{1}{n_p}\sum_{i=1}^{n}\sum_{j=1}^{m}|y_{(i,j)} - \widehat{y}_{(i,j)}| \tag{10}$$

$$MAPE = \frac{1}{n_p}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|y_{(i,j)} - \widehat{y}_{(i,j)}|}{y_{(i,j)}} \tag{11}$$

Where $y_{(i,j)}$, $\hat{y}_{(i,j)}$ indicate the ground truth and estimated value of the crash risk in gird cell (i, j), respectively. n, m indicate the number of rows and columns for the used grids in this study, and $n_p = n \times m$.

In this study, MSE and MAE are used to measure the total predictive performance in the whole test dataset, while MAPE only covers the top 5% highest crash risk samples in the test dataset (denoted as MAPE^{*H}), and is used to measure the model's predictive performance in grid cells and time periods of high crash risks. The predicted highly crash risky regions and time periods are particularly valuable for transportation authorities to allocate police forces and apply proactive countermeasures. Thus, the concern of zero ground truth crashes in Eq. (11) does not exist.

In addition, in order to further investigate the robustness and applicability of the proposed models, the prediction accuracy rate (PAR) and false positive rate (FPR) are used to evaluate the classification performance and false alarm rate in identifying the high crash risk status, respectively. Given the confusion matrix in Table 3, the prediction accuracy rate and false positive rate can be calculated as follows:

$$PAR = \frac{TN + TP}{TN + FP + FN + TP} \tag{12}$$

$$FPR = \frac{FP}{TN + FP} \tag{13}$$

4. Results of data analysis

4.1. Spatiotemporal pattern of taxi trips and crashes

The drop-off locations of taxi trips and reported crashes during the selected time period in the study area are plotted in Fig. 5(a) and (b) to illustrate the spatial distributions of the taxi trips and crashes. The density of the taxi trips and the crashes are depicted with different colors in which red color generally indicates high density, while blue generally indicates low density. By inspecting Fig. 5 visually we find that the spatial and temporal distribution of the taxi trips exhibit somewhat similar patterns with those of the crash counts. However, in certain time periods and spatial units they exhibit different patterns. For example, by inspecting Fig. 5(c) visually, we find that there is a significant positive peak bias between the distribution curve of taxi trips and the curve of crashes at daytime while in the nighttime the situation is opposite. This finding seems to provide some insightful suggestions for the transportation management that at daytime the increased taxi trips may serve as a potential indicator for crashes. However, during the evening rush hours (e.g. 4 pm-5 pm), the crash counts reach another peak but the taxi trips reach the peak around 7 pm. This distinction suggests that in the nighttime due to the shift of taxi drivers the taxi trips may not better indicate the crash occurrence. Moreover, the taxi trips remain relatively high in nighttime and are mostly associated with recreational activities, but seem not to be correlated with a large number of crashes. The findings that can be obtained from Fig. 5 are twofold. First, there is a strong relationship between taxi trips and crashes. The large-scale taxi GPS data can serve as a potential data source for citywide short-term crash risk prediction. Second, the relationship between taxi trips and crashes is very complex and may also be interacted with other external factors. Thus, other data

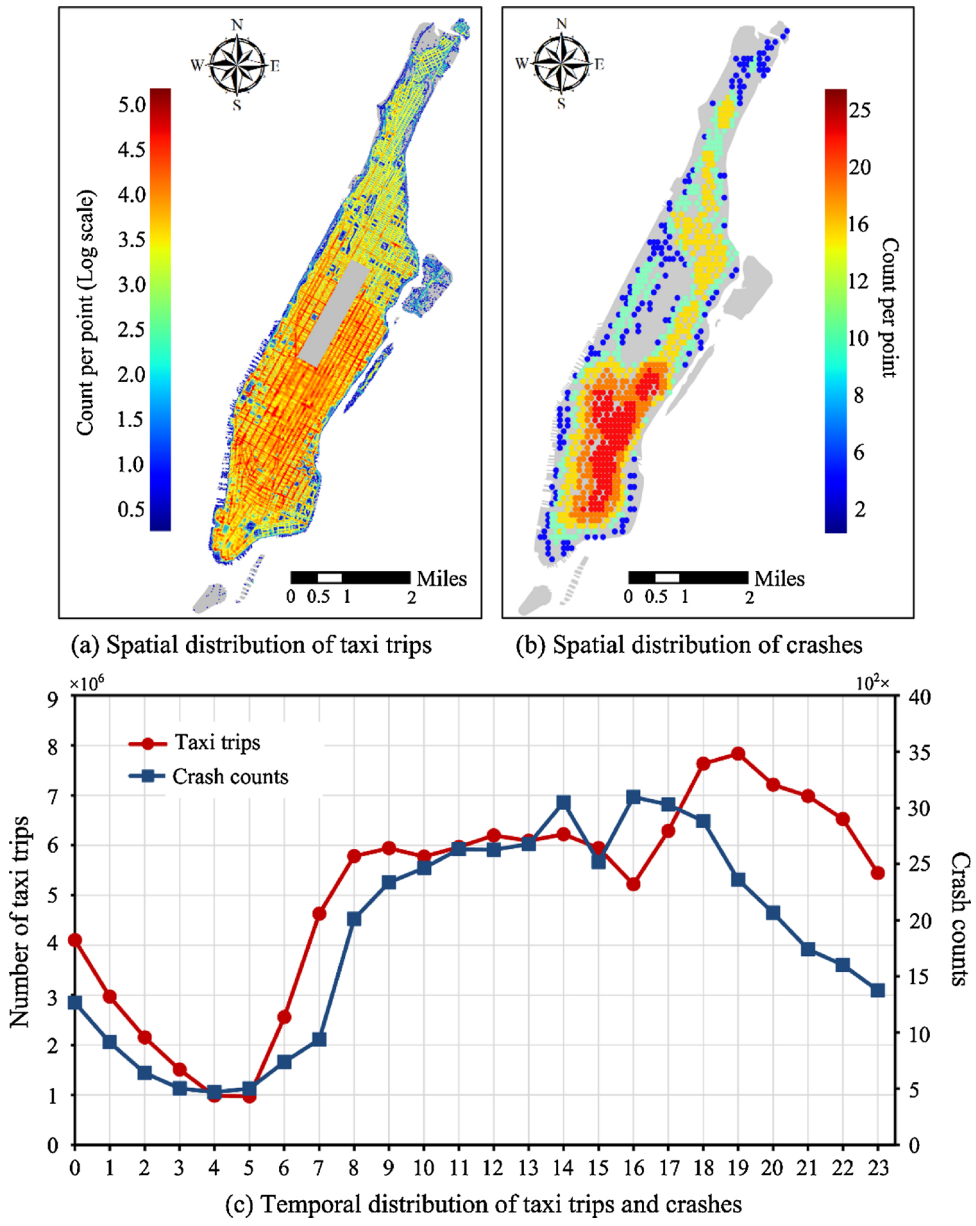


Fig. 5. The spatiotemporal distribution of taxi trips and crashes.

such as weather data, land use data and demographic data should also be incorporated in crash risk prediction models.

4.2. Results of STCL-Net analyses

In the present study, three general categories of short-term crash risk prediction models are developed: weekly crash risk model, daily crash risk model and hourly crash risk model. For each category three different types of models are developed: the model with 8 × 3 grid, the model with 15 × 5 grid and the model with 30 × 10 grid. Thus, in total nine citywide crash risk prediction tasks of different spatiotemporal resolution are developed and compared. The sample size of each crash risk prediction model are illustrated in Table 4.

For each developed model, the selected dataset is divided into training set and test set according to the ratio of 7:3. Fig. 6 illustrates the temporal variation of daily crash risk in Manhattan. The training dataset is from Jan 1 st, 2015 to Sep 10th, 2015, and the test dataset is from Sep 11th, 2015 to Dec 31th, 2015. As shown in Fig. 6, both the training dataset and test dataset have some time periods with abnormally high crash risk, leading to great challenges for the crash risk prediction.

Table 4
The sample size of each crash risk prediction model.

	8 × 3	15 × 5	30 × 10
Weekly crash risk model	1,056	3,300	13,200
Daily crash risk model	8,592	26,850	107,400
Hourly crash risk model	210,072	656,475	2,625,900

In this study, the proposed STCL-Net includes two CNN layers for spatial features extraction, two LSTM layers for temporal features extraction and two ConvLSTM layers for spatiotemporal features extraction. Since no general rules can be directly applied to set optimal values of hyper-parameters, the selection of hyper-parameters mainly rely on many previous studies (Krizhevsky et al., 2012; Deng, 2016; Lv et al., 2015; Hou and Edara, 2018; Dabiri and Heaslip, 2018; Ma et al., 2015a,b; Zhu et al., 2018; Ma et al., 2015a,b; Shi et al., 2015; Scherer et al., 2010; Van Lint et al., 2002; Hochreiter and Schmidhuber, 1997).

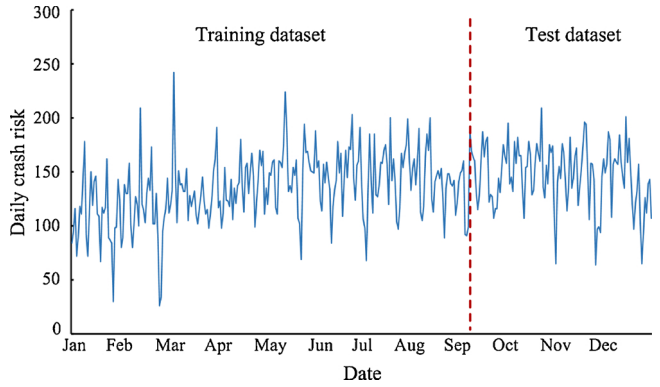


Fig. 6. Temporal variations of daily crash risk in Manhattan during the study period.

Specifically, referred to some famous CNN frameworks such as LeNet (Krizhevsky et al., 2012) and AlexNet (LeCun et al., 1998), the filter size of convolutional layer in this study is set as (3, 3), and the filter size of max pooling layer is set as (2, 2) which are typical kernel shapes to effectively reduce the spatial size of input feature maps. The depths of convolutional layers for different prediction tasks are test by assigning different values from 10 to 40 until the optimal results achieve. In the LSTM layers, different hidden cell sizes of 256, 512, 1024, and 2048 which were commonly used in previous studies (Ma et al., 2015a,b; Chen et al., 2016; Ma et al., 2015a,b; Van Lint et al., 2002; Hochreiter and Schmidhuber, 1997), are employed to determine the optimal sizes of hidden cell in different short-term crash risk prediction tasks.

The STCL-Net is trained based on the RMSprop optimizer (Tieleman and Hinton, 2012) in the back propagation process. RMSprop is an appropriate optimization technique for large-scale transportation network and has achieved remarkable performances in many previous studies (Zhu et al., 2018; Ma et al., 2015a,b; Shi et al., 2015). In the RMSprop optimizer, the learning rate is set as 0.001, the decay parameter is set as 0.9 and the batch size is set as 64. In this study, the loss function is the mean square error (MSE), and the proportion of validation data is set as 0.2 during the learning process. Moreover, to address the overfitting issue, the dropout layer and the early stopping criteria (Sarle, 1995) are also applied in this study. All the input variables are standardized through the min-max normalization before model training and validation. The proposed STCL-Net is developed and implemented on the basis of Keras framework with Tensorflow as backend. The experiments are conducted using Python 3.5 in a Windows 7 system with 16 GB RAM, and a GTX 1060 Graphics Processing Card is used to accelerate the model learning procedure.

The results of proposed STCL-Net in different short-term crash risk prediction tasks are shown in Table 5. For the prediction tasks under the same grid size and prediction time level, the comparisons of MSE and MAE suggest that the models incorporating all types of features outperform those only consider one or two types of features. This finding seems to confirm the benefits of incorporating multi-source datasets in citywide short-term crash risk prediction. In addition, the comparison of $MAPE^{*H}$ across different grid size and prediction time levels indicate that in general the prediction performance of the proposed STCL-Net on high crash risky regions and time periods decreases as the spatiotemporal resolution increases. This finding is expected because crash occurrence is a random event and shorter aggregated forecasting time in smaller region will pose greater challenge to the prediction task.

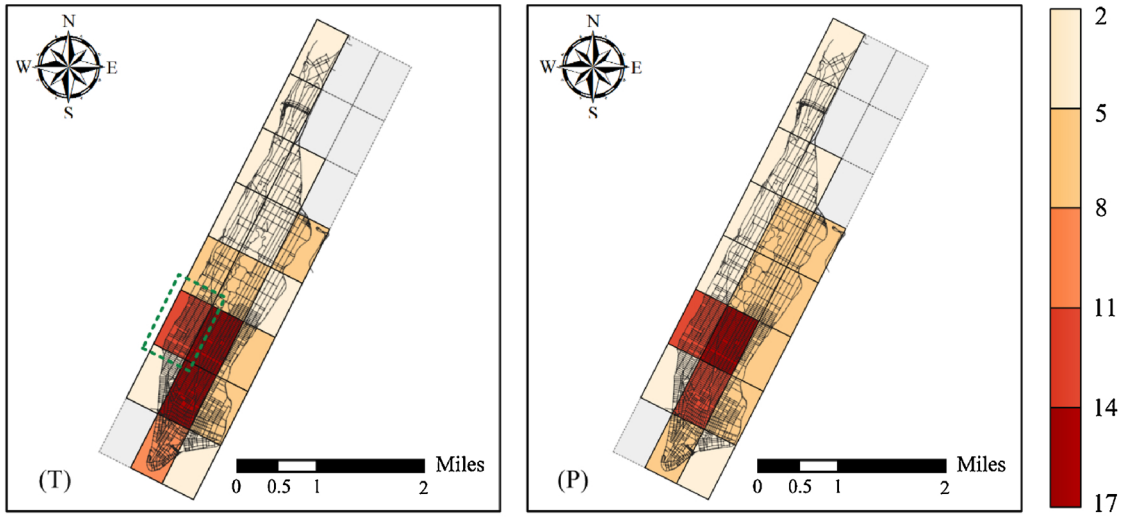
Moreover, Fig. 7 further visually depicts some examples of the ground truth and predicted value of daily crash risk. The deeper color indicates higher crash risk. In general, the crash risk is unbalanced across the space where grid cells around downtown area have much higher crash risk than

other grid cells. By further visually inspecting the green square area of Fig. 7, we find that when a large gird cell is divided into some small ones, the crash occurrence in each grid cell will become extremely rare and the distribution of crash risks among neighboring cells also become more unbalanced. This finding again explains the reason why the model performance decreases as the spatiotemporal resolution increases. In addition, the visualization of the developed crash risk prediction map can also help transportation authorities to clearly identify and forecast the grid cells with relatively high crash risk and to apply proactive countermeasures for these areas.

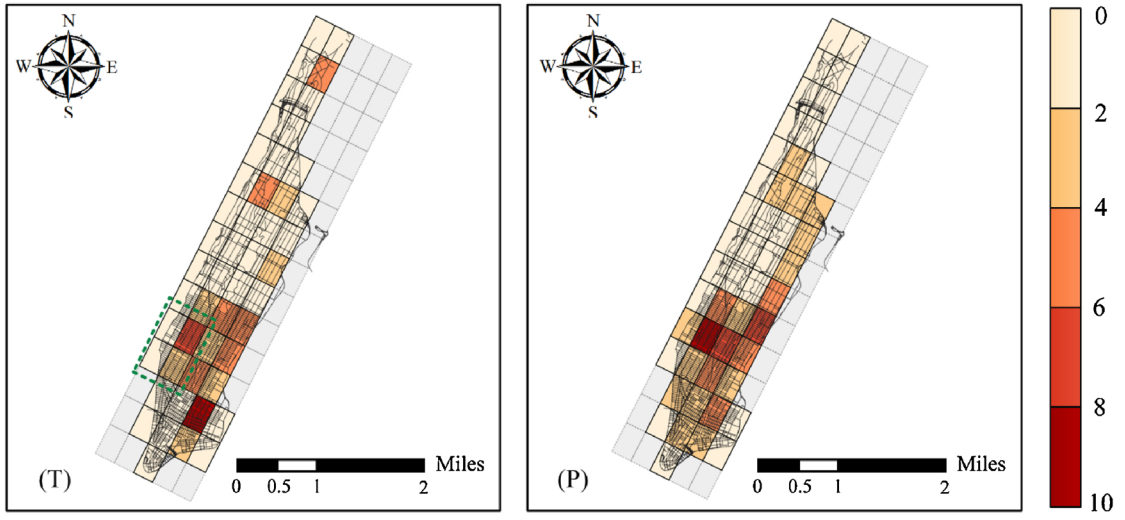
Table 5
Comparison of STCL-Net models incorporating different features.

Prediction time level	Grid size	Incorporated features	MSE	MAE	$MAPE^{*H}$ (%)
Weekly	8 × 3	Crash risk	131.719	8.124	14.068
		Crash risk + Taxi trips	129.158	7.796	12.093
		Crash risk + Taxi trips + Other features	89.25	7.143	8.368
	15 × 5	Crash risk	37.897	4.108	22.281
		Crash risk + Taxi trips	34.014	4.077	17.193
		Crash risk + Taxi trips + Other features	27.986	3.796	13.979
	30 × 10	Crash risk	6.909	1.663	32.051
		Crash risk + Taxi trips	6.841	1.658	31.819
		Crash risk + Taxi trips + Other features	6.04	1.625	25.248
Daily	8 × 3	Crash risk	13.225	2.635	28.488
		Crash risk + Taxi trips	13.119	2.618	27.006
		Crash risk + Taxi trips + Other features	9.715	2.37	19.048
	15 × 5	Crash risk	3.561	1.315	39.137
		Crash risk + Taxi trips	3.559	1.313	38.269
		Crash risk + Taxi trips + Other features	2.519	1.165	29.033
	30 × 10	Crash risk	0.818	0.562	60.987
		Crash risk + Taxi trips	0.782	0.555	60.122
		Crash risk + Taxi trips + Other features	0.606	0.482	49.606
Hourly	8 × 3	Crash risk	0.378	0.471	78.711
		Crash risk + Taxi trips	0.375	0.418	76.096
		Crash risk + Taxi trips + Other features	0.367	0.409	71.92
	15 × 5	Crash risk	0.119	0.144	85.696
		Crash risk + Taxi trips	0.119	0.15	85.401
		Crash risk + Taxi trips + Other features	0.116	0.134	83.729
	30 × 10	Crash risk	0.03	0.038	96.514
		Crash risk + Taxi trips	0.03	0.036	96.331
		Crash risk + Taxi trips + Other features	0.019	0.023	94.653

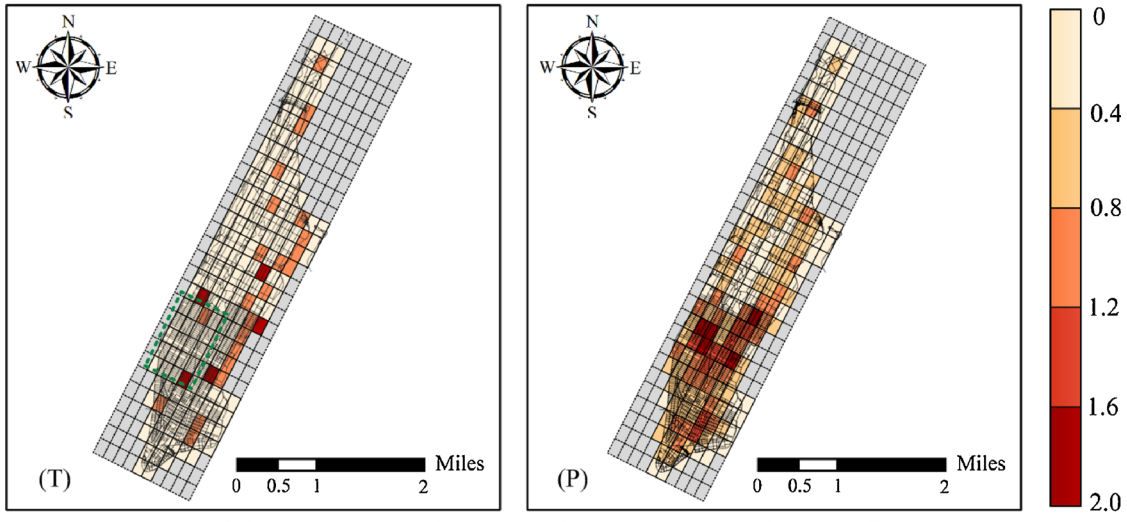
* H: MAPE covers the top 5% highest crash risk samples in the test dataset.



(a) Comparison of the ground truth (T) and prediction risk value (P) with 8X3 grid size



(b) Comparison of the ground truth (T) and prediction risk value (P) with 15X5 grid size



(c) Comparison of the ground truth (T) and prediction risk value (P) with 30X10 grid size

Fig. 7. Comparison of the truth (T) and predicted daily crash risk (P) for different grids.

Table 6
Comparison of different methods in terms of weekly crash risk prediction.

Grid Size									
Models	8×3			15×5			30×10		
	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)
STCL-Net	89.25	7.143	8.368	27.986	3.796	13.979	6.04	1.625	25.248
CNN	109.514	7.733	11.428	35.518	3.976	20.7	6.847	1.737	31.063
LSTM	121.036	7.625	17.351	32.979	4.28	21.357	7.147	1.784	32.51
ANN	132.577	8.124	15.571	31.453	3.92	23.703	7.43	1.718	37.334
GBRT	128.564	8.067	12.019	29.155	3.826	16.284	6.173	1.696	26.71
ARIMA	121.45	8.034	12.719	30.772	3.862	16.371	6.509	1.69	27.652
Random-parameter	112.506	7.817	11.799	29.601	3.822	15.71	6.345	1.685	27.856
Random-effects	101.825	7.344	10.947	26.549	3.597	14.31	5.676	1.579	25.711

(*continued on next page*)

Table 7
Comparison of different methods in terms of daily crash risk prediction.

Grid Size									
Models	8×3			15×5			30×10		
	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)
STCL-Net	9.715	2.37	19.048	2.519	1.165	29.033	0.606	0.482	49.606
CNN	11.856	2.573	21.02	3.528	1.303	36.925	0.805	0.537	62.739
LSTM	13.331	2.686	27.364	3.532	1.299	40.072	0.758	0.563	60.361
ANN	10.443	2.411	25.51	3.272	1.288	37.67	0.732	0.577	56.648
GBRT	12.778	2.605	24.144	3.835	1.367	38.37	0.707	0.554	55.826
ARIMA	13.935	2.763	24.336	3.835	1.383	38.848	0.878	0.586	61.505
Random-parameter	12.471	2.625	23.238	3.538	1.34	40.155	0.939	0.614	63.994
Random-effects	12.029	2.573	22.431	3.353	1.301	37.308	0.891	0.591	60.527

(*continued on next page*)

Table 8
Comparison of different methods in terms of hourly crash risk prediction.

Grid Size									
Models	8×3			15×5			30×10		
	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)	MSE	MAE	MAPE*H (%)
STCL-Net	0.367	0.409	71.92	0.116	0.134	83.729	0.019	0.023	94.653
CNN	0.403	0.408	74.413	0.12	0.144	84.745	0.031	0.028	96.99
LSTM	0.399	0.413	73.108	0.123	0.152	85.193	0.032	0.033	97.124
ANN	0.4	0.412	76.538	0.124	0.164	85.551	0.031	0.035	97.541
GBRT	0.381	0.419	76.447	0.124	0.162	85.643	0.029	0.035	95.813
ARIMA	0.459	0.414	77.52	0.14	0.159	86.5	0.034	0.044	97.789
Random-parameter	0.484	0.448	77.316	0.134	0.155	92.759	0.031	0.033	97.133
Random-effects	0.483	0.448	77.303	0.136	0.158	92.757	0.035	0.033	97.125
GWR	0.496	0.459	77.735	0.133	0.162	89.003	0.035	0.03	97.21

* H: MAPE covers the top 5% highest crash risk samples in the test dataset.

4.3. Results of model comparison

In this section, several benchmark methods are also tested and compared with the proposed STCL-Net model. The selected benchmark methods can be categorized into two types: econometric models and machine-learning models. The selected econometric models include autoregressive integrated moving average (ARIMA), random-parameter model, random-effects model, and geographically weighted regression (GWR) model, which were commonly used in previous studies of spatial analysis of crashes (Bao et al., 2017; Shi and Abdel-Aty, 2015; Bao et al., 2018; Ukkusuri et al., 2011). The selected machine-learning models include CNN, LSTM neural network, artificial neural network (ANN), and gradient boosting regression tree (GBRT), which were the state-of-art approaches in many previous studies of short-term traffic flow prediction (Ma et al., 2015a,b; Chan et al., 2012; Xia and Chen, 2017).

ARIMA is a widely-used time-series analysis method, which can integrate moving average model, the autoregressive part and the moving average part simultaneously (Box and Pierce, 1970). Random-effects model assumes that individual specific effects are uncorrelated with the independent variables, which can well account for the spatial correlation, temporal correlation or a combination of the two among observations (Naznin et al., 2016). Recently, random-parameter model and GWR model are two potential crash modeling techniques for addressing the spatial heterogeneity issue by allowing varying relationship between crash counts and predicting factors (Bao et al., 2017;

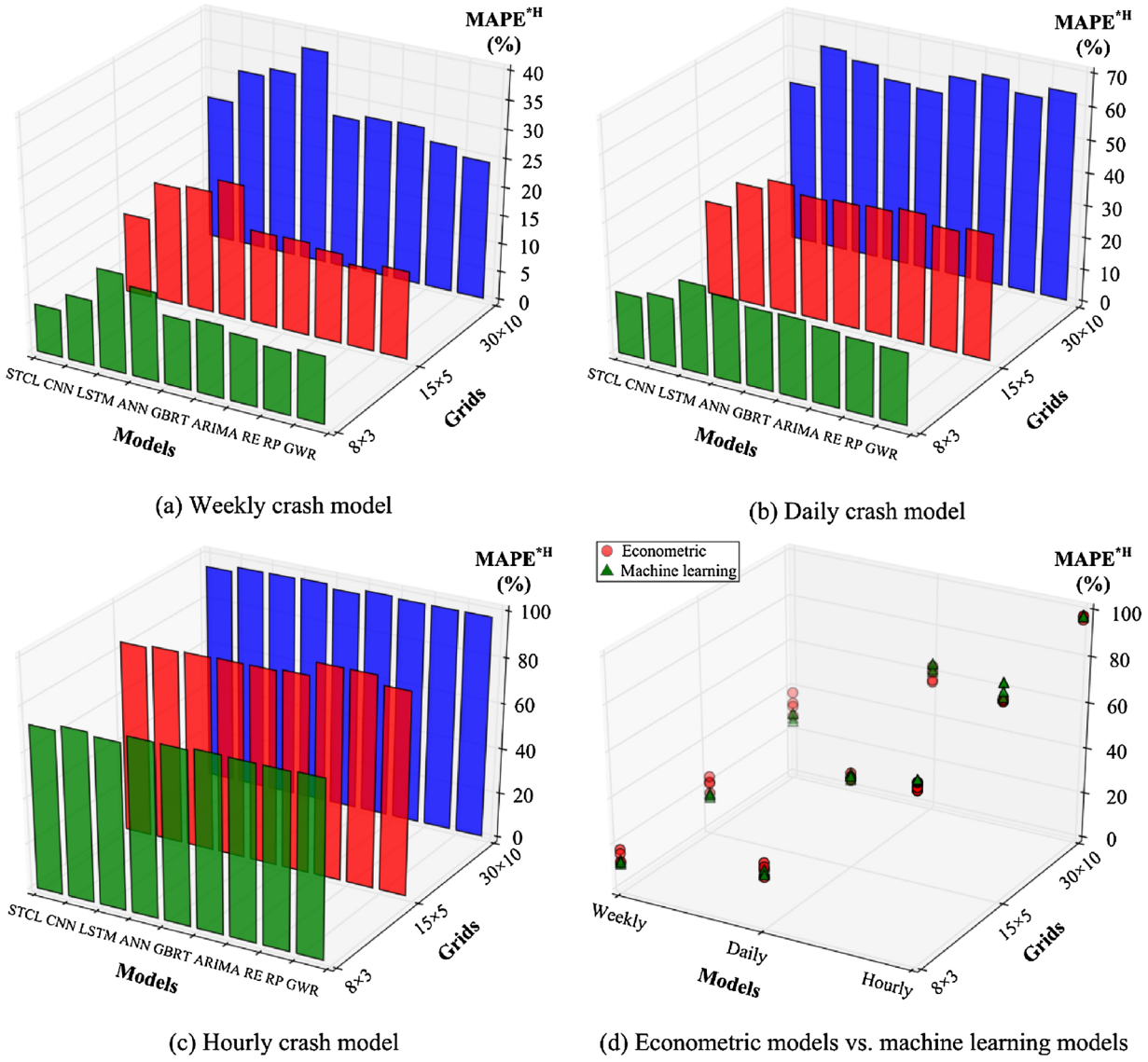


Fig. 8. Visualization of model comparison results.

Ukkusuri et al., 2011). The characteristics of CNN and LSTM neural network can be referred to section 3.2 and 3.3, respectively. ANN represents the traditional feed-forward neural network and consists of multiple fully-connected layers, which attempts to learn the features from multiple hidden layers. GBRT is a tree-based machine-learning method by assembling numerous prediction trees (Xia and Chen, 2017). GBRT has been applied in many fields of transportation researches due to the remarkable performances than other tree-based methods.

To conduct a fair comparison, all the benchmark methods are under fine-tuned with the same input data and number of training epochs. However, the data structure used in the STCL-Net cannot be directly applied in the benchmark methods. Some data processing needed to be conducted. More specifically, for the ANN, GBRT, GWR, ARIMA, random-parameter model and random- effects model, all the variables, including the historical crash risk, number of taxi trips, weather features, DVKT, land use, population and road-network-related attributes of a specific grid (i, j) are reshaped to a vector to predict the future crash risk value in grid (i, j). For the LSTM, all the variables in grid (i, j) are reshaped to a matrix with one axis as time step and another axis as the feature category. For the CNN, the data structure is similar to that in STCL-Net.

The comparison results of weekly crash model, daily crash model and hourly crash model are shown in Table 6–8 and Fig. 8, respectively. Note that the proposed STCL-Net is not included in Fig. 8(d). From the model comparison, the following findings are discussed:

(a) In general, the proposed STCL-Net outperform the selected benchmark models for different crash risk prediction tasks in terms of lower MSE, MAE and MAPE^{*H}. This finding confirms the superiority and feasibility of the proposed model, which can successfully capture both the spatial and temporal features among variables in citywide short-term crash risk prediction.
(b) The model performance decreases sharply with the increased spatiotemporal resolution of crash risk prediction tasks. For example, weekly crash risk prediction model with 8×3 grid exhibit the best performance while the hourly crash risk prediction model with 30×10 grid exhibit the worst performance.
(c) As shown in Fig. 8(d), for the weekly crash risk prediction task, econometric models generally exhibit better performance than machine-learning models. For the daily crash risk prediction task, machine-learning models achieve better results than econometric models. However, for the hourly crash risk prediction task, they both present poor prediction accuracy.

To further investigate the robustness and applicability of the proposed model, the prediction accuracy for automatic high crash risk identification are examined. In terms of label data, the predicted crash risk data are labelled as high crash risk or non-high crash risk according to the crash alert threshold value. In the present study, three different alarm threshold values are defined: the 95th-percentile, 90th-percentile, 85th-percentile crash risk in the sample, which were commonly applied in previous studies of crash hotspot identification (Yu et al., 2014; Park et al., 2014). The comparison results based on prediction accuracy rate (FPR) and false positive rate (FPR) metrics for weekly crash model, daily crash model and hourly crash model under three different alarm threshold values, are given in Table 9–11, respectively. From the comparison results, we can find that in general the proposed STCL-Net performs better than the selected benchmark models in the identification of high crash risk status in terms of higher PAR and lower FPR for all three crash risk prediction tasks. In addition, the prediction accuracy rate of the proposed model decreases as the spatiotemporal resolution of prediction task increases, while the false alarm rate of the proposed model increases as the spatiotemporal resolution of prediction task increases. For the prediction tasks with the same grid size and prediction time level, the prediction accuracy rate of the proposed model increases as the crash risk alarm threshold value increases, while the false alarm rate of the proposed model decreases as the crash risk alarm threshold value increases.

In addition, the findings of the present study could provide insightful suggestions for transportation engineers to select appropriate methods for different crash modeling problems. More specifically,

(a) For city planners, they focus on incorporating safety considerations into long-term transportation planning. The crash modeling is usually conducted at a relatively large spatial unit, such as TAZ and ZCTA. In this condition, the econometric models such as random-parameter model and GWR model are more preferable. The econometric model not only achieve better prediction results than machine-learning models, but also better reveal the impact of influencing factors on crash occurrences in each region.
(b) For the transportation agency, they would like to understand the dynamic change of crash risk in urban areas such that they could proactively allocate the police forces and remind the drivers about the crash hotspots dynamically. Thus, the crash risk prediction needs to be conducted at block-level within a relatively short time period, such as hourly or daily. In this context, machine-learning models will be more preferable due to the superior prediction performance.

5. Conclusions and discussions

The present study investigates how the deep learning approach contributes to citywide short-term crash risk prediction by leveraging multi-source datasets. This study uses the data collected from Manhattan in the New York City to illustrate the procedure. The following multiple datasets are collected: crash data, large-scale taxi GPS data, road network attributes, land use features, population data and weather data. A spatiotemporal convolutional long short-term memory network (STCL-Net) is proposed for predicting the citywide short-term crash risk. The proposed deep learning architecture is merged by multiple CNN layers, LSTM layers, and ConvLSTM layers, and could integrate a variety of explanatory variables collected from multi-source data. A total of nine prediction tasks are conducted and compared, including weekly, daily and hourly models with 8×3, 15×5 and 30×10 grids, respectively. The results suggest that the prediction performance of the proposed model decreases as the spatiotemporal resolution of prediction task increases. In addition, comparative analyses are also conducted to compare the performance of proposed STCL-Net that incorporates different explanatory variables. The results suggest that the proposed models with all types of variables achieve the best performance, which confirm the benefits of incorporating multi-source data in citywide short-term crash risk prediction.

Moreover, four commonly-used econometric models for spatial analysis of crashes (ARIMA, Random-parameter model, Random-effects model and GWR), and four state-of-the-art machine learning models for short-term traffic flow prediction (CNN, LSTM, ANN, and GBRT) are selected as benchmark methods to compare with the proposed STCL-Net for all the crash risk prediction tasks. The comparative analyses suggest that in general the proposed STCL-Net outperform the benchmark methods for different crash risk prediction tasks in terms of lower MSE, MAE and MAPE^{*H}. The prediction accuracy rate and false alarm rate of the developed models are also compared and the results indicate that the proposed spatiotemporal deep learning approach performs better at capturing the spatiotemporal characteristics for the citywide short-term crash risk prediction. In addition, the comparative analyses also reveal that econometric models perform better than machine-learning models in weekly crash risk prediction tasks, while they exhibit worse results than machine-learning models in daily crash risk prediction tasks. The results of this study can help transportation engineers to determine the right size of spatial and temporal scale for crash modeling. The potential engineering application are mainly twofold. First, for the city planners, the predicted crash risk results could help transportation agencies to regularly monitor region level safety and provide incentives to reduce the number of traffic casualties in a region's safety program. Second, for daily transportation management, the short-term crash risk results could reveal the crash hotspots in urban areas dynamically. The transportation agency could dynamically allocate the police forces, which is particularly important under limited human resources. In the future, with the increased prediction accuracy of crash models under high spatiotemporal resolutions, the results have great potential to be applied in vehicle routing assistance system, which could incorporate traffic safety into current navigation applications and provide drivers the safest and shortest route.

Even though the proposed spatiotemporal deep learning approach has exhibited great potential to citywide short-term crash risk prediction, several limitations are still needed to be addressed in this study:

(a) The large-scale taxi trip data may not be able to represent the general human mobility of the whole city. The inherent biases/limitations of taxi trip data may potentially affect the prediction performance of crash models.
(b) The model performances decrease sharply when the spatiotemporal resolution of prediction tasks increase. In particular, the hourly crash risk models exhibit quite poor prediction accuracy and are far away from practical application.
(c) Compared with traditional econometric models, the proposed STCL-Net works like a black box and falls short on the factor analysis, which is a common issue in most artificial intelligent methods.

The sample bias problem is a prevalent problem in many big data related studies (Bao et al., 2017; Xie et al., 2017; Hasan et al., 2013; Bao et al., 2018). Some of these limitations may go away since more and more on-board GPS data of private vehicles and other modes can be collected in urban areas. The reasons for the poor performance in hourly crash risk model are mainly twofold. First, due to limited available traffic sensors in urban areas, AADT on freeways and major arterials is used in hourly crash risk models instead of hourly traffic data, leading to a poor prediction of hourly crash risk. Second, when the spatiotemporal resolution of the prediction tasks improve, the zero-inflated problem will occur, leading to great challenges for crash risk prediction. In the future, other data sources such as the mobile phone data, social media data could also be combined with taxi data for better depicting the hourly traffic information and human mobility in urban areas and generating a better prediction of citywide short-term crash risk. Moreover, new spatiotemporal deep learning architecture should also be explored to addressing the zero-inflated issue. The authors recommend that future studies could focus on these issues.

Table 9
Comparison of prediction accuracy for weekly crash risk models under three different alarm threshold values.

Grid size									
Models	8×3						15×5		
	Crash alarm threshold PAR (%)	95th percentile FPR (%)	Crash alarm threshold PAR (%)	90th percentile FPR (%)	Crash alarm threshold PAR (%)	85th percentile FPR (%)	Crash alarm threshold PAR (%)	95th percentile FPR (%)	Crash alarm threshold PAR (%)
STCL-Net	99.21	0.00	97.69	1.58	97.22	1.69	97.78	0.47	95.61
CNN	99.07	0.00	97.3	1.65	97.3	1.65	97.33	0.78	95.46
LSTM	98.11	0.02	96.81	1.99	96.23	2.29	97.32	1.28	95.41
ANN	98.17	0.01	97.06	1.95	96.36	2.18	97.29	1.28	95.37
GBRT	98.17	0.01	97.1	1.94	96.54	2.01	97.61	0.58	95.53
ARIMA	98.18	0.01	97.06	1.89	96.49	2.12	97.58	0.61	95.54
RP	98.86	0.01	97.12	1.92	96.58	1.89	97.61	0.51	95.55
RE	99.06	0.00	97.26	1.69	96.69	1.79	97.67	0.47	95.52
GWR	98.86	0.01	97.16	1.89	96.69	1.79	97.61	0.48	95.55

Grid size									
Models	15×5			30×10					
	90th percentile FPR (%)	Crash alarm threshold PAR (%)	85th percentile FPR (%)	Crash alarm threshold PAR (%)	95th percentile FPR (%)	Crash alarm threshold PAR (%)	90th percentile FPR (%)	Crash alarm threshold PAR (%)	85th percentile FPR (%)
STCL-Net	1.16	94.42	1.58	95.48	1.01	93.56	2.32	88.78	3.37
CNN	1.36	93.52	2.76	94.92	1.28	92.93	2.49	86.89	3.83
LSTM	1.46	93.15	3.06	94.83	1.39	92.73	3.3	86.22	3.99
ANN	1.43	92.75	3.28	94.75	1.58	92.49	3.61	86.1	4.06
GBRT	1.26	94.31	2.18	95.27	0.97	93.27	2.35	87.93	3.68
ARIMA	1.29	94	2.66	95.11	1.1	93.08	2.45	87.42	3.71
RP	1.29	94.42	1.81	95.04	1.12	93.02	2.44	87.32	3.78
RE	1.17	94.36	2.02	95.21	0.82	93.37	2.35	88.49	3.61
GWR	1.29	94.39	1.91	95.52	0.84	94.98	2.3	90.26	3.21

Table 10
Comparison of prediction accuracy for daily crash risk models under three different alarm threshold values.

Grid size																		
Models	8×3						15×5						30×10					
	Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile		Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile		Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile	
	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)
STCL-Net	96.46	1.11	94.39	1.55	92.31	1.92	95.24	1.02	92.28	1.79	88.98	3.13	94.05	0.53	86.24	1.63	75.35	4.13
CNN	96.38	1.12	94.39	1.5	92.3	2.09	95.18	1.24	91.8	1.8	88.43	3.15	92.08	0.62	84.23	2.52	74.29	5.46
LSTM	96.13	1.75	93.17	2.21	91.16	2.75	94.4	1.41	91.18	1.96	87.59	3.82	93.94	0.53	86.16	1.82	74.81	4.43
ANN	96.21	1.66	93.25	2.13	91.59	2.54	94.64	1.35	91.78	1.81	88.36	3.36	94.04	0.46	86.2	1.77	75.04	4.45
GBRT	96.31	1.59	94.26	2.07	92.08	2.29	94.54	1.32	91.68	1.77	88.38	3.38	94.01	0.48	86.2	1.68	75.75	4.2
ARIMA	96.25	1.4	93.85	2.11	91.77	2.37	94.42	1.39	91.15	1.87	87.71	3.74	92.09	0.57	84.06	2.38	74.35	5.03
RP	96.32	1.19	94.24	1.84	92.06	2.14	93.64	1.44	91.04	1.98	87.39	3.94	92.02	0.71	83.61	2.72	73.61	6.12
RE	96.38	1.14	94.28	1.54	92.11	2.13	94.67	1.29	91.73	1.8	88.39	3.33	92.19	0.54	84.26	1.86	74.56	4.6
GWR	96.36	1.15	94.35	1.53	92.13	2.13	94.47	1.39	91.63	1.8	88.2	3.43	92.1	0.65	84.09	2.65	74.13	5.54

Table 11
Comparison of prediction accuracy for hourly crash risk models under three different alarm threshold values.

Grid size																		
Models	8×3						15×5						30×10					
	Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile		Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile		Crash alarm threshold 95th percentile		Crash alarm threshold 90th percentile		Crash alarm threshold 85th percentile	
	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)	PAR (%)	FPR (%)
STCL-Net	93.72	0.13	80.74	3.46	80.74	3.46	92.65	0.22	76	3.58	76	3.58	81.58	0.34	71.02	4.17	71.02	4.17
CNN	93.73	0.13	80.74	3.73	80.74	3.73	92.65	0.22	75.68	3.6	75.68	3.6	81.53	0.36	70.88	4.21	70.88	4.21
LSTM	93.68	0.13	80.46	3.67	80.46	3.67	92.66	0.21	75.39	3.6	75.39	3.6	80.94	0.36	70.85	4.21	70.85	4.21
ANN	93.43	0.1	79.83	4.16	79.83	4.16	92.57	0.23	75.38	3.71	75.38	3.71	79.76	0.5	69.77	5.56	69.77	5.56
GBRT	93.53	0.12	80.39	3.92	80.39	3.92	92.57	0.23	75.36	3.76	75.36	3.76	81.55	0.36	70.96	4.21	70.96	4.21
ARIMA	93.16	0.08	76.53	5.28	76.53	5.28	92.51	0.27	75.27	3.84	75.27	3.84	79.55	0.51	68.69	5.76	68.69	5.76
RP	93.34	0.09	76.96	5.09	76.96	5.09	92.42	0.31	73.4	4.16	73.4	4.16	80.58	0.42	70.58	4.68	70.58	4.68
RE	93.34	0.09	76.98	5.08	76.98	5.08	92.46	0.3	73.52	4.13	73.52	4.13	80.53	0.43	70.33	4.8	70.33	4.8
GWR	93.2	0.09	76.54	5.09	76.54	5.09	92.5	0.3	73.54	4.12	73.54	4.12	79.8	0.46	69.8	5.21	69.8	5.21

Acknowledgements

This research is supported by the Projects of International Cooperation and Exchange of the National Natural Science Foundation of China (No. 51561135003) and the Scientific Research Foundation of Graduate School of Southeast University (No. YBJJ1790). In addition, the authors thank China Scholarship Council for supporting Jie's one-year study abroad at Purdue University.

References

Abdel-Aty, M., Hassan, H., Ahmed, M., Al-Ghamdi, A., 2012. Real-time prediction of visibility-related crashes. Transp. Res. Part C 24, 288–298.
Bao, J., Liu, P., Yu, H., Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accid. Anal. Prev. 106, 358–369.
Bao, J., Liu, P., Qin, X., Zhou, H., 2018. Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. Accid. Anal. Prev. 120, 281–294.
Box, G.E., Pierce, D.A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Am. Stat. Assoc. 65, 1509–1526.
Chan, K.Y., Dillon, T.S., Singh, J., Chang, E., 2012. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. IEEE Trans. Intell. Transp. Syst. 13 (2), 644–654.
Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mob. Netw. Appl. 19, 171–209.
Chen, Q., Song, X., Yamada, H., Shibasaki, R., 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In Proceedings of the 30th AAAI Conference on Artificial Intelligence.
Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C 86, 360–371.
Deng, L., 2016. Deep learning: from speech recognition to language and multimodal processing. APSIPA Trans. Signal Inf. Process. 5, 1–15.
Faghih-Imani, A., Eluru, N., El-Geneidy, A.M., Rabbat, M., Haq, U., 2014. How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. J. Transp. Geogr. 41, 306–314.
Federal Highway Administration, 2005. Safetea-LU: Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users. U.S. Department of Transportation.
González, M.C., Hidalgo, C.A., Barabási, A.L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.
Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151, 304–318.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hou, Y., Edara, P., 2018. Network scale travel time prediction using deep learning. In Proceedings of the 97th TRB Annual Meeting.
Huang, H.L., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida bayesian spatial modeling. Transp. Res. Rec.: J. Transp. Res. Board 2148, 27–37.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.
Li, Z.B., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using geographically weighted poisson regression for county-level crash modeling in California. Saf. Sci. 58, 89–97.
Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. 16 (2), 865–873.
Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015a. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transp. Res. Part C 54, 187–197.
Ma, X., Yu, H., Wang, Y., Wang, Y., 2015b. Large-scale transportation network congestion evolution prediction using deep learning theory. PLoS One 10 (3), 1–17.
Naznin, F., Currie, G., Logan, D., Sarvi, M., 2016. Application of a random effects negative binomial model to examine tram-involved crash frequency on route sections in Melbourne, Australia. Accid. Anal. Prev. 92, 15–21.
NCHRP (National Cooperative Highway Research Program), 2010. PLANSAFE: Forecasting the Safety Impacts of Socio-demographic Changes and Safety Countermeasures. Transportation Research Board. NCHRP, pp. 8–44.
Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. Accid. Anal. Prev. 35, 599–611.
Noland, R.B., Smart, M.J., Guo, Z., 2016. Bikeshare trip generation in New York City. Transp. Res. Part A 94, 164–181.
Park, B., Lord, D., Lee, C., 2014. Finite mixture modeling for vehicle crash data with application to hotspot identification. Accid. Anal. Prev. 71, 319–326.
Ren, H., Song, Y., Liu, J., Hu, Y. and Lei, J. (2017). A Deep Learning Approach to the Prediction of Short-term Traffic Accident Risk. arXiv: 1710.09543.
Rhee, K., Kim, J., Lee, Y., Ulfarsson, G.F., 2016. Spatial regression analysis of traffic crashes in Seoul. Accid. Anal. Prev. 91, 190–199.
Sarle, W.S., 1995. Stopped training and other remedies for overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics 352–360.
Scherer, D., Müller, A., Behnke, S., 2010. Evaluation of pooling operations in convolutional architectures for object recognition. Proceedings of 20th International Conference on Artificial Neural Networks.
Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C 58, 380–394.
Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Proceedings of 29th Annual Conference on Neural Information Processing Systems (NIPS).
Tieleman, Tijmen, Hinton, G., 2012. Lecture 6.5-RMSprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning. pp. 26–31.
Ukkusuri, S., Hasan, S., Abdul, A.H., 2011. Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transp. Res. Rec.: J. Transp. Res. Board 2237, 98–106.
Van Lint, J.W.C., Hoogendoorn, S.P., Zuylen, H.V., 2002. Freeway travel time prediction with state-spaced neural networks: modeling state-space dynamics with recurrent neural networks. Transp. Res. Rec.: J. Transp. Res. Board 1811, 30–39.
Xia, Y., Chen, J., 2017. Traffic flow forecasting method based on gradient boosting decision tree. Proceedings of the 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology 130, 413–416.
Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots. Risk Anal. 37 (8), 1459–1476.
Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66, 80–88.
Zhu, L., Guo, F., Krishnan, R., Polak, J., 2018. The use of convolutional neural networks for traffic incident detection at a network level. In Proceedings of the 96th TRB Annual Meeting.