

# A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction

Honglei Ren\*, You Song\*, Jingwen Wang\*, Yucheng Hu<sup>+</sup>, and Jinzhi Lei<sup>+</sup>

\*School of Software, Beihang University, Beijing, China,  
songyou@buaa.edu.cn, renhongleiz@buaa.edu.cn, wangjingwen@buaa.edu.cn

<sup>+</sup>Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University, Beijing, China,  
huyc@tsinghua.edu.cn, jzlei@mail.tsinghua.edu.cn

**Abstract**—With the rapid development of urbanization, the boom of vehicle numbers has resulted in serious traffic accidents, which led to casualties and huge economic losses. The ability to predict the risk of traffic accident is important in the prevention of the occurrence of accidents and to reduce the damages caused by accidents in a proactive way. However, traffic accident risk prediction with high spatiotemporal resolution is difficult, mainly due to the complex traffic environment, human behavior, and lack of real-time traffic-related data. In this study, we collected big traffic accident data. By analyzing the spatial and temporal patterns of traffic accident frequency, we presented the spatiotemporal correlation of traffic accidents. Based on the patterns we found in analysis, we proposed a high accurate deep learning model based on recurrent neural network toward the prediction of traffic accident risk. The predictive accident risk can be potential applied to the traffic accident warning system. The proposed method can be integrated into an intelligent traffic control system toward a more reasonable traffic prediction and command organization.

## I. INTRODUCTION

In modern society, the rapid development of urbanization has resulted in the boom of vehicles, causing a number of problems, such as traffic congestion, air pollution, and traffic accidents. These problems have caused huge economic loss as well as human casualties. According to *Global Status Report on Road Safety*, published by World Health Organization in 2015, about 1.25 million people were killed in traffic accidents every year. With the help of big traffic data and deep learning, real-time traffic flow prediction has enabled people to avoid traffic jam by choosing less congested routes. Big traffic data and deep learning may also provide a promising solution to predict or reduce the risk of traffic accidents.

One important task in traffic accident prevention is to build an effective traffic accident risk prediction system. If the traffic accident risk in a certain region can be predicted, we can disseminate this information to the nearby drivers to alert them or make them choose a less hazardous road. However, accurate prediction of traffic accident risk is very difficult because many related factors could affect traffic accident. For example, different regions have tremendous difference on traffic accident rate. In addition, poor weather condition such as snow or fog can reduce road visibility and traffic capacity, thus increase the change of traffic accidents. Traffic accident rate varies at different time of a day, possibly related to the physical

condition of the drivers. Although many researchers have focused on the identification of key factors associated with traffic accident [1], effective prediction of the traffic accident risk dynamically remains to be a challenge problem.

With the development of deep learning, methods based on deep learning and big data have shown favorable results in traffic related problems, such as traffic flow prediction [2], arrival time estimation [3], origin-destination forecasting [4], etc. As for traffic accident risk prediction based on deep learning, to our best knowledge, the only work is done by Chen et. al., who use human mobility features extracted from Stack denoise Autoencoder to infer traffic accident risk in Japan [5]. However, they did not consider the periodical patterns and the spatial distribution patterns of traffic accidents. In particular, traffic accidents may closely related to the day of week. Other important factors they missed are weather condition, air quality, etc. To improve the power of traffic accident risk prediction, it is important to combine all these factors into a comprehensive model.

In this paper, we collected big traffic accident data and built a deep model for traffic accident risk prediction based on recurrent neural network. By analyzing the spatial and temporal patterns of traffic accident frequency, we presented the spatiotemporal correlation of traffic accidents. Based on the patterns we found in analysis, we proposed a high accurate deep learning model for traffic accident risk prediction. The model can learn deep connections between traffic accidents and its spatial-temporal patterns. As a potential application, the traffic accident prediction system based on our method can be used to help traffic enforcement department to allocate police forces in advance of traffic accidents.

The rest of this paper is organized as follows: Section 2 introduces some previous works that are related with the present one. Section 3 describes the data source and the pattern analysis result of traffic accidents. Section 4 introduces our deep learning model for traffic accident risk prediction. Section 5 shows the results of experiment. Section 6 gives the conclusions and future works.

## II. RELATED WORK

1) *Identification of Traffic Accident Trigger*: Tremendous efforts have been devoted to the identification of key con-

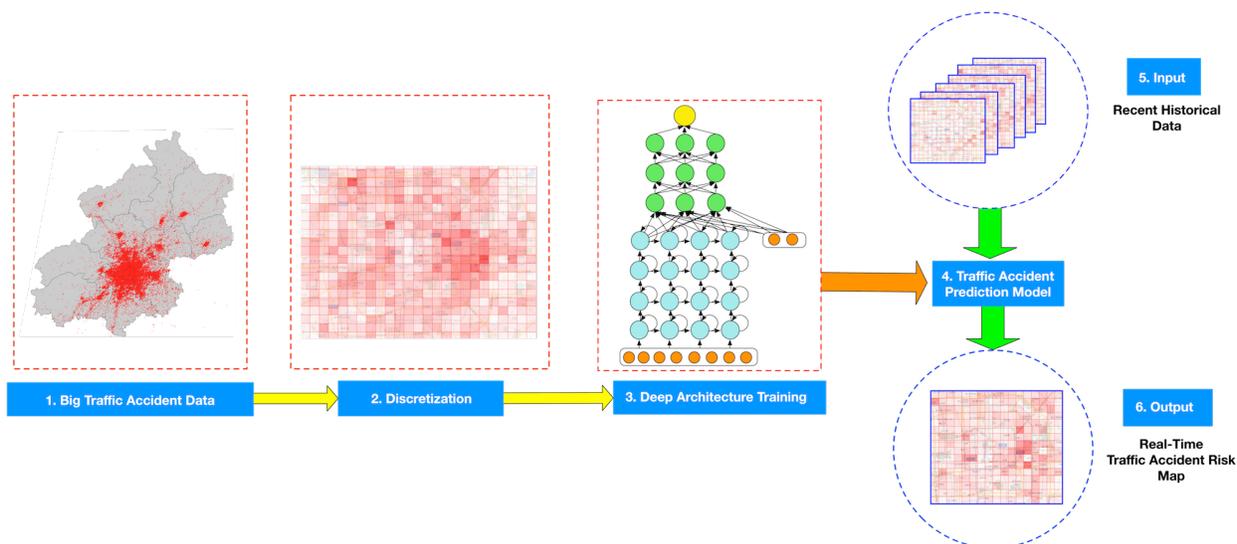


Fig. 1. Workflow of our traffic accident risk prediction method. First, the big traffic accident data is collected. Second, the data is discretized in space and time, and then feed to the deep model for training. After training, we feed recent historical data to the trained model and then obtained the real-time traffic accident risk prediction.

ditions or particular traffic patterns that could lead to traffic accident. For instance, Oh proposed the assumption that disruptive traffic flow is a trigger to crash [6]. Based on the loop detector data and crash data, they found that 5-min standard deviation of speeds right before a traffic accident is an effective indicator of crash. Although different crash indicators have been proposed, they could not meet the requirement of accurate accident prediction because numerous factors have complex connections with traffic accidents.

2) *Real-time Traffic Accident Prediction*: With the development of machine learning, many researchers start to focus on real-time traffic accident prediction. Lv chose feature variables based on Euclidean metric and utilized k-nearest neighbor method to predict traffic accident [8]. Park collected big traffic accident data of highway in Seoul and build a prediction workflow based on k-means cluster analysis and logistic regression [9]. Recently, Chen used human mobility data in Japan and build a Stack denoise Autoencoder to infer the real-time traffic risk [5]. One limitation of these works is that, they did not incorporate several importance factors such as traffic flow, weather condition, air quality into their model. Without these information, the predictive power of the model could be weakened.

3) *Deep Learning*: The success of deep learning has proved its power in discovering intricate structures in high-dimensional data. It has been widely used as the state-of-the-art technique in image recognition speech recognition, natural language understanding, etc. As for researches on intelligent transportation system, a number of studies focus on traffic flow prediction based on deep learning [2]. In a longer time scale, some studies try to predict the congestion evolution of large-scale transportation network [10]. Another interesting application utilized deep reinforcement learning to control the

timing of traffic signal [11].

### III. PATTERN ANALYSIS OF TRAFFIC ACCIDENT

#### A. Big Traffic Accident Data

In this study, to predict traffic accident risk, the traffic accident records of Beijing in 2016 and 2017 was collected. Each record contains the time, GPS (Global Positioning System) coordinate of the accident event.

#### B. Data Preprocessing

Before we analyze the pattern of accident, and build machine learning model, a proper data structure is necessary. Therefore, we first preprocess our raw data by discretization.

The traffic accident data was first discretized in space and time. The temporal resolution was 1 hour for different time horizon of prediction, and spatial resolution dimension was  $1000\text{m} \times 1000\text{m}$  in uniform grids.

After discretization, we obtained a matrix  $S$  whose element  $S_{r,t}$  is the count of traffic accidents happened within region  $r$  and time slot  $t$ .

#### C. Spatial Distribution of Traffic Accident

To explore whether traffic accident frequency is associated with the geographical position of a region, we plot the heatmap of traffic accident frequency in Beijing in 2016 (Figure 2). As shown in Figure 2, the traffic accident frequency is not uniform distributed, and it is highly related with the geographical position of a region. Usually, the highest traffic accident region lies in the major commercial and business areas.

#### D. Temporal Pattern of Traffic Accident

To explore the temporal patterns of the traffic accident frequency, we first checked whether everyday's traffic accident count varies in different time period. Figure 3 gives the

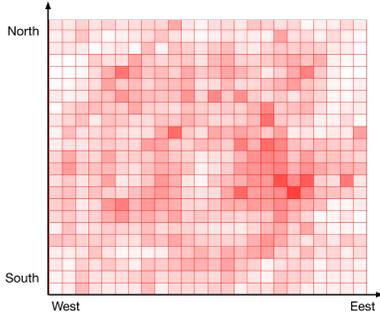


Fig. 2. The heatmap of traffic accident frequency in Beijing in 2016 with 1000m\*1000m spatial resolution. Deeper red indicates higher frequencies of traffic accident.

scatter and box-plot of the everyday's traffic accident count for different time periods of Beijing. Obviously, the traffic accident patterns change drastically for different time period of a day. Specifically, traffic accident is more frequent at rush hours than that at off-peaks.

The time periods in Figure 3 is defined according to working time pattern and Chinese lifestyle [1]: 00:00–06:59 (mid-night to dawn), 07:00–08:59 (morning rush hours), 09:00–11:59 (morning working hours), 12:00–13:59 (lunch break), 14:00–16:59 (afternoon working hours), 17:00–19:59 (afternoon rushing hours), and 20:00–23:59 (nighttime).

Beside temporal patterns for different time period, we also want to know whether weekly periodic patterns exist in traffic accident frequency. Therefore, we plot a two week's histogram of hourly traffic accident count (Figure 4). It can be observed that the patterns of histograms are similar for the same day of week and between weekdays.

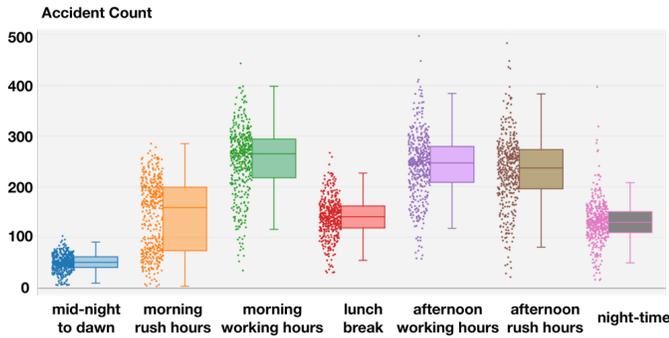


Fig. 3. Scatter and box-plot of everyday's traffic accident count for different time periods in Beijing.

To quantify the spatio-temporal correlation between traffic accident, we first defined the spatial correlation for a given time  $t$  as follows:

$$C(k, t) = \frac{\sum_{i,j} (a_{i,j,t} - \bar{a}_t)(a_{i',j',t} - \bar{a}_t)}{\sum_{i,j} (a_{i,j,t} - \bar{a}_t)^2} \quad (1)$$

where

$$\bar{a}_t = \frac{\sum_{i,j} a_{i,j,t}}{M * N}$$

The  $C(k, t)$  in Eq.(1) is the spatial correlation with a  $k$  Manhattan distance for a given time  $t$ .  $a_{i,j,t}$  is the traffic accident count happened in grid  $(i, j)$  and time  $t$ .  $\bar{a}_t$  is the average traffic accident count of all grids at time  $t$ .  $M$  and  $N$  are the number of grids along the longitude and latitude.

Based on the spatial correlation defined by Eq.(1), the spatio-temporal correlation can be written as Eq.(2).

$$f(k, \tau) = \frac{\sum_t (C(k, t) - \bar{C}(k))(C(k, t + \tau) - \bar{C}(k))}{\sum_t (C(k, t) - \bar{C}(k))^2} \quad (2)$$

where  $f(k, \tau)$  is the correlation of two grids with a  $k$  Manhattan distance and time interval  $\tau$ .

Figure 5 shows the contour map of the spatio-temporal correlation of traffic accident. It can be observed that the correlation shows a strong temporal periodic pattern, and the period is around 24 hours. Traffic accidents have about 0.4 ~ 0.5 correlation if their Manhattan distance is within 4 km, and time interval is the multiples of 24 hours.

#### IV. DEEP MODEL OF TRAFFIC ACCIDENT RISK PREDICTION

As Chen et. al. has documented, after some analysis of traffic accident data, we find it is difficult to predict whether traffic accident will happen or not directly, because complex factors can affect traffic accident, and some factors, such as the distraction of drivers, can not be observed and collected in advance [5]. Figure 6 gives the dimensionality reduction result by t-SNE when we predict whether traffic accident happen or not directly, the input data is the sequence of traffic accident count for a region. Obviously, the red points (accident) and black points (non-accident) are inseparable, and that means it is hard to predict whether traffic accident happen or not directly. Therefore, we try to predict traffic accident frequency(risk), that is average traffic accident count per hour for the same time of recent days (3 days, 7 days, 30 days, etc.). For instance, if 5 accidents happened during 8:00-9:00 a.m. in last 3 days, then today's traffic accident frequency during 8:00-9:00 a.m is  $\frac{5 \text{ times}}{3 \text{ hours}} \approx 1.67 \text{ times/hour}$ .

Our method is illustrated in Figure 1. First, we discretized the big traffic accident data in space and time, so that it can be processed by machine learning algorithm. Then we constructed a deep model on the basis of recurrent neural network to infer traffic accident risk, and input the processed data into it. After the data training, we input the recent traffic accident frequency data into the trained model, and then obtained the predicted accident risk map from the output.

##### A. Model

In this subsection, we will introduce our Traffic Accident Risk Prediction Method based on LSTM (TARPLM), and Figure 7 illustrates the deep model of TARPLM. The input layers are consisted of two parts. The first input is the sequence of recent traffic accident frequency, and it is input to the first LSTM layer. The second input contains the longitude and latitude of the region center that we expected to predict,

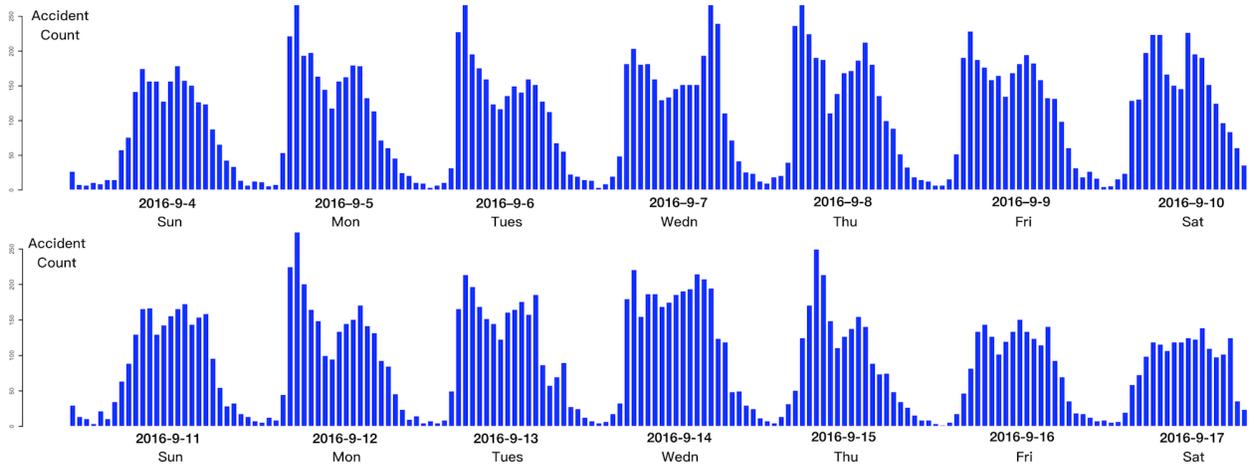


Fig. 4. Histogram of hourly traffic accident count from 2016-09-04 to 2016-09-17 (two weeks)

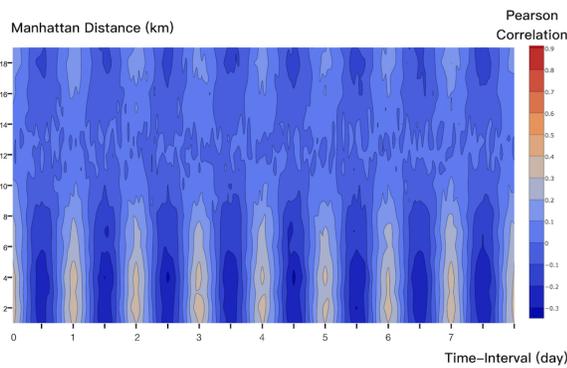


Fig. 5. The contour map of the spatio-temporal correlation of traffic accident. The horizontal and vertical axis are the time-interval and Manhattan distance of two grids, respectively.

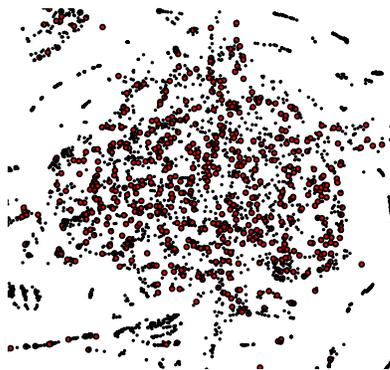


Fig. 6. Dimensionality reduction and visualization with t-SNE, the original data is the sequence of traffic accident count for a region. The red points are the data with traffic accident, and the black points are the traffic accident free data

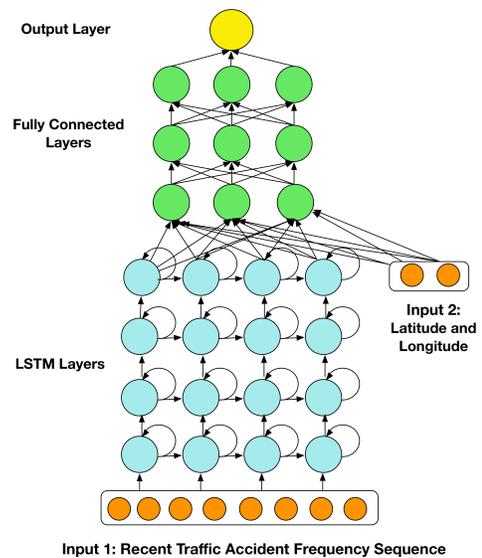


Fig. 7. The Deep Model of Traffic Accident Risk Prediction Method based on LSTM. The model consisted of 2 separated input layers, 4 LSTM layers, 3 fully connected layers and 1 output layer. The first input layer is made up of the sequence of recent traffic accident frequency. The second input layer contains the longitude and latitude of the region center that we expected to predict.

and it directly input into fully connected layers. The hidden layers of deep model is consisted of 4 LSTM layers and 3 fully connected layers sequentially. The last layer of model is output layer, which outputs the predicted traffic accident risk(frequency) for the given input.

To avoid overfitting, we add a dropout layer with 0.5 dropout rate between each two fully connected layers. The activation function of fully connected layers and output layer is Rectified Linear Units (RELU), which can be denoted as  $\max(0, x)$  mathematically.

The reason why we chose LSTM is that LSTM can capture the periodic feature of traffic accident, and traditional RNNs shows poor performance and intrinsic difficulties in training

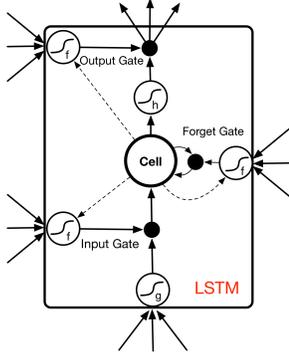


Fig. 8. The structure of LSTM Cell, which is consisted of an input gate, a neuron with a self-recurrent link, a forget gate and an output gate.

when it has long time period. These weaknesses have been proved in researches related with traffic flow prediction[10]. On another hand, the explicit memory cell in LSTM can avoid the problems of gradient vanish or gradient explosion existed in traditional RNNs. The structure of LSTM is similar to traditional RNNs, and it consisted of one input layer, one or several hidden layer and one output layer. The core concept of LSTM is its memory cell in hidden layer, it contains 4 major parts: an input gate, a neuron with a self-recurrent link, a forget gate and an output gate, and its inner structure is shown in Figure 8.

## V. EXPERIMENTS AND RESULTS

In this section, we compare our TARPML method with several baseline models, including Lasso, Support Vector Regression, Random Forest Regression, etc. All of the experiments are performed by a PC (CPU: Intel Xeon(R) CPU E5-2609, 32GB memory, GPU: Tesla K20C).

### A. Experimental Setup

Because our model is temporal related, we arrange the data chronologically. We chose the data from 2016-01-01 to 2017-04-01 as the training data, and the data from 2017-04-01 to 2017-08-20 is for testing. The last 20% of training data is used as validation data. The sample size of training, validation, testing is 1590958, 397740 and 233850 respectively.

The architecture of TARPML are built upon Keras, which is a Python Deep Learning library. We chose mean squared error as objective function of optimization, and selected RMSProp as the optimizer.

By comparing the Root Mean Square Error (RMSE) of TARPML method with different input sequence length (Table I), we finally chose 100 as the best sequence length to input. The number of neurons of the LSTM layers are 100, 200, 200, 200, respectively, and the number of neurons of each hidden layer is 200.

### B. Performance Evaluation

1) *Evaluation Metrics*: To evaluate the accuracy and precision of the prediction, we selected Mean Absolute Error

TABLE I  
PERFORMANCE COMPARASION OF TARPML METHOD WITH DIFFERENT INPUT SEQUENCE LENGTHS

Sequence Length	10	20	50	100
1 day	0.119	0.122	0.115	0.105
3 days	0.042	0.041	0.038	0.034
7 days	0.022	0.021	0.018	0.015
30 days	0.006	0.005	0.005	0.004

TABLE II  
PERFORMANCE FOR 3-DAY TRAFFIC ACCIDENT FREQUENCY PREDICTION

Method	MAE	MSE	RMSE
Lasso	0.046	0.006	0.076
SVR	0.066	0.006	0.075
DTR	0.021	0.004	0.058
ARMA	0.058	0.049	0.169
<b>TARPML</b>	0.014	0.001	0.034

(MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as our metrics. They are defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |r_i - \hat{r}_i| \quad (3)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (r_i - \hat{r}_i)^2 \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - \hat{r}_i)^2} \quad (5)$$

where  $n$  is the sample size,  $r_i$  and  $\hat{r}_i$  are real and predicted risk (accident frequency) respectively.

2) *Baseline Models*: We selected several traditional machine learning models as our baseline models to compare the prediction performance with our TARPML method. The baseline models we selected are Lasso, Support Vector Regression (SVR), Decision Tree Regression (DTR) and Autoregressive Moving Average Model (ARMA). All these models were implemented by scikit-learn, a Python module that implemented lots of state-of-the-art machine learning algorithms, and the default parameters of baseline models were used.

3) *Performance Evaluation*: We compared the predictability of our model with that of baselines, and Table II demonstrates their MAE, MRE and RMSE values for 3-day traffic accident frequency (Its definition can be found at Section IV). The table shows that our model outperforms than other models, and have less prediction errors.

4) *Simulation Results*: To evaluate the effectiveness of our model, here we selected 2017-07-10 (Monday) as a example for comparing the prediction results of different models. Figure 9 (a) - (e) are the real traffic accident risk map (a) and the predicted results of different models (b) - (e), and it can be

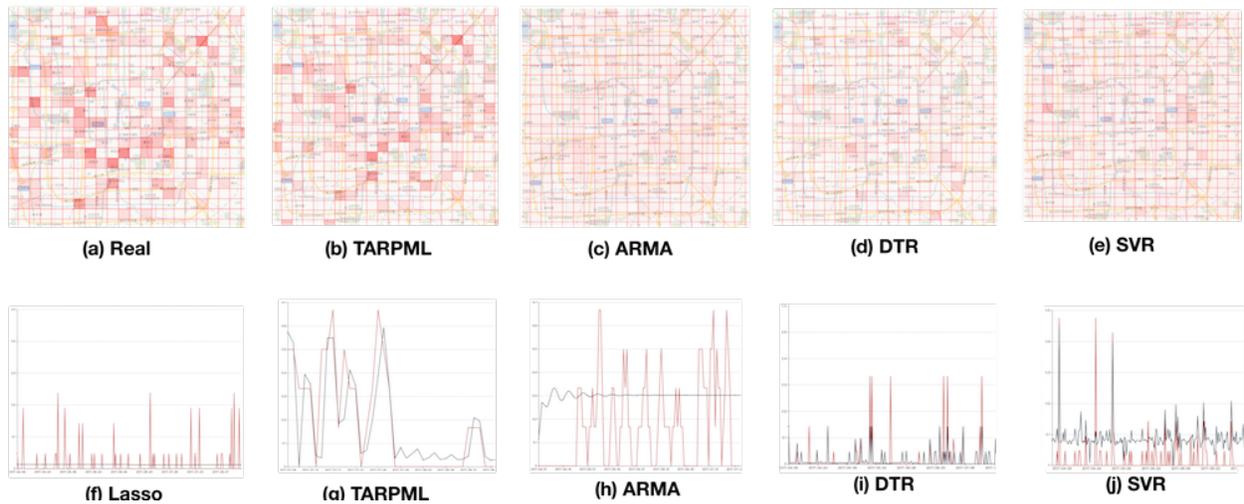


Fig. 9. Comparison of the real traffic accident risk map (a) and the predicted results of different models (b) - (e). (f) - (j) are the predicted risk curve from different models and its corresponding real traffic risk curve.

clearly seen that TARPML model is far better than other models. Figure 9(f) - (j) are the predicted risk curve from different models and its corresponding real traffic risk curve. It can also be observed that the predicted curve from TARPML model are more accurate than others.

## VI. CONCLUSION

In this paper, we collected big traffic accident data, and built a deep learning model based on LSTM for predicting traffic accident risk. Based on the pattern analysis result, it can be observed that the traffic accident risk are not uniformly distributed in space and time. It shows strong periodical temporal patterns and regional spatial correlation. According to the dimensionality reduction result (t-SNE, Figure 6), it is hard to predict traffic accident directly. Therefore, we defined the traffic accident risk by its frequency, and built a deep learning model based on LSTM to capture its spatial and temporal patterns. The performance comparison based on RMSE (Table II) and the predicted risk map (Figure 9) shows the accuracy and effectiveness of our model. This study therefore indicates that benefits gained from temporal-spatial features, big traffic accident data and deep recurrent neural network can bring accurate traffic accident risk prediction. Our method can be easily applied to the traffic accident warning system and help people avoiding traffic accident by choosing safer regions.

However, due to the complexity of traffic accident, our study has some limitations in following aspects. First, here we only utilized the traffic accident data itself for prediction. However, other related data, such as traffic flow, human mobility, road characteristic and special events, maybe significant to traffic accident risk prediction as well. Second, our prediction results are coarse-grained, and can not provide road level accident risk prediction. But it can be easily applied to the road network based prediction. Therefore, future work combined with structure of urban road network and comprehensive

factors related with traffic accident will be promising to make better prediction result.

## REFERENCES

- [1] G. Zhang, K. K. W. Yau, and G. Chen, "Risk factors associated with traffic violations and accident severity in China." *Accident; analysis and prevention*, vol. 59, pp. 18–25, Oct. 2013.
- [2] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," *arXiv preprint arXiv:1612.01022*, 2016.
- [3] C. Bai, Z. Peng, Q. Lu, and J. Sun, "Dynamic bus travel time prediction models on road with multiple bus routes," *Computational Intelligence and Neuroscience*, vol. 2015, pp. 432 389–432 389, 2015.
- [4] F. Toqué, E. Côme, M. K. El Mahrsi, and L. Oukhellou, "Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 1071–1076.
- [5] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," 2016.
- [6] C. Oh, J.-S. Oh, S. Ritchie, and M. Chang, "Real-time estimation of freeway accident likelihood," in *80th Annual Meeting of the Transportation Research Board, Washington, DC*, 2001.
- [7] M. Abdel-Aty, N. Uddin, A. Pande, F. Abdalla, and L. Hsia, "Predicting freeway crashes from loop detector data by matched case-control logistic regression," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1897, pp. 88–95, 2004.
- [8] Y. Lv, S. Tang, and H. Zhao, "Real-time highway traffic accident prediction based on the k-nearest neighbor method," in *Measuring Technology and Mechatronics Automation, 2009. ICMTMA'09. International Conference on*, vol. 3. IEEE, 2009, pp. 547–550.
- [9] S.-h. Park, S.-m. Kim, and Y.-g. Ha, "Highway traffic accident prediction using vds big data analysis," *The Journal of Supercomputing*, vol. 72, no. 7, pp. 2815–2831, 2016.
- [10] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS one*, vol. 10, no. 3, p. e0119044, 2015.
- [11] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 3, pp. 247–254, 2016.