Ref. Ares(2020)7961544 - 27/12/2020

Marie Skłodowska-Curie Actions

Research and Innovation Staff Exchange (RISE)

www.cosafe.org.uk

Cooperative Connected Intelligent Vehicles for Safe and Efficient Road Transport

Grant Agreement no: 824019

D5.1: V2X Traffic Modelling and Prediction model

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 824019

DELIVERABLE DOCUMENT SHEET	
Project Title	*COSAFE*
Project Start Date	*30/11/2018*
Project Duration	*48 months*
Work Package	*WP5: Service-oriented V2X Network and Resource Management*
Deliverable #	*D5.1: V2X Traffic Modelling and Prediction model.*
Deliverable Due Date	*31/12/2020*
Dissemination Level	*Public*
WP Leader	*Ranplan wireless network design*
Leading Editors	*Bo Ma (Ranplan), Fuyou Li (Ranplan), Zhezi Zhao (Ranplan), Kan Lin (Ranplan), Ke Zhang (University of Electronics Science and Technology of China), Jianhua He (University of Essex), Kaisheng Huang (Tsinghua University).*
Reviewers	*Jie Zhang (Ranplan), Laura Li (Birmingham City Council), Jianye Huang (Tsinghua University), Supeng Leng (University of Electronics Science and Technology of China).*
Document Reference	*COSAFE-WP5-D5.1-Final*
Contact:	*Dr Kan Lin, Ranplan wireless network design* *Dr Jianhua He, j.he@essex.ac.uk, University of Essex*

QUALITY CONTROL ASSESSMENT SHEET			
Issue	**Date**	**Comment**	**Leading Author**
V1	2020-11-15	First Draft	Bo Ma, Fuyou Li, Zhezi Zhao
V2	2020-12-16	Second Draft	Kan Lin
V-Final	2020-12-25	Final Draft	Jianhua He

Short Names for COSAFE Partners

Partner	**University of Essex**	**Aston University**	**University of Oslo**	**University of Göttingen**	**Ranplan Wireless**	**Birmingham City Council**	**Tsinghua University**	**University of Electronics Sci. & Tech. of China**	**Forward Innovation**
Short name	**UESSEX**	**ASTON**	**OSLO**	**UGOE**	**RPN**	**BCC**	**THU**	**UESTC**	**FWI**
Leader	Dr J. He	Dr Z. Tang	Prof Y. Zhang	Prof X. Fu	Prof J. Zhang	Mrs N. Bryant	Prof K. Huang	Prof S. Leng	Mr Z. Xiong

DISCLAIMER

Table of Contents

1. Executive Summary

In the COSAFE project, WP5 is dedicated to model and predict CAV application traffic; develop efficient network resource management schemes to adapt to road and traffic conditions, in order maximize CAV application performance.

The leading partner (Ranplan wireless network design) and the supporting partners have followed the original research plan, and conducted research as outlined in the proposal. The partners involved in WP5 have successfully achieved the research targets which have been set for the first stage for the research for WP5.

In this report, we will present the research progresses we have made with Task 5.1 on the development V2X traffic modelling and prediction model, and natural language processing (NLP) for the event detection in social media, deep learning for the traffic prediction. The report presents the research development and implementation of relevant algorithms. And the application of the technologies and research outcomes will also be presented.

2. Introduction

V2X networks are shaped by both road network and communication network traffic. Cellular network takes the majority in the communication network so as its traffic. The cellular network traffic records users' registration to base stations no matter they are pedestrians or drivers, so we can know how many users are in this region as well as their trajectories. This is especially important in urban regions because there is still no geo-location data having such a long-term and full-scale monitoring of citizens' traffic (both road and communication traffic).

Counting the number of citizens according to the base stations on the roads can offer an overview of road traffic. Moreover, monitoring the cellular network usage can quantify the communication traffic of diverse services. In that case, we can take advantages of this kind of data to quantify the correlation on the road traffic and communication traffic. Here, an assumption is made that the CIV traffic is positively correlated to the general road traffic. The results can be exploited in the network resource management and CIV applications.

With this correlation, stochastic and machine learning models would be built to fit the traffic data. Both spatial and temporal dimensions are added to describe the urban traffic. In other words, space-time dependency of the traffic pattern is taken into account. Other WPs can utilize the built models for their simulation. Also, the machine learning models are able to predict future traffic distribution. That means optimization can be made proactively to improve the energy- and time-efficiency. Deep learning algorithm are used to predict the complex long term spatial-temporal distribution of the communication traffic and short-term traffic size.

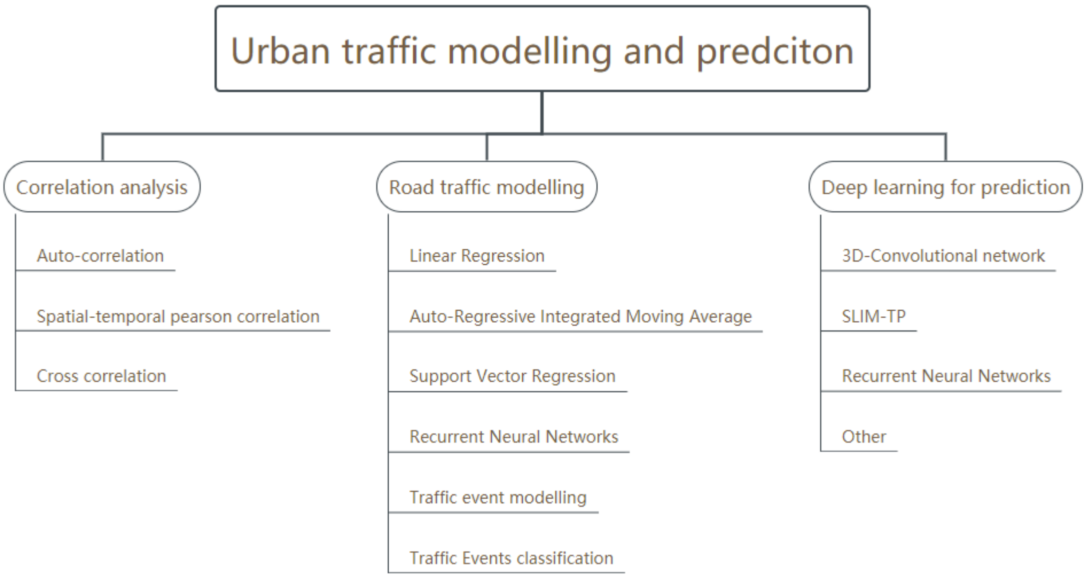


Figure 1 Report organization

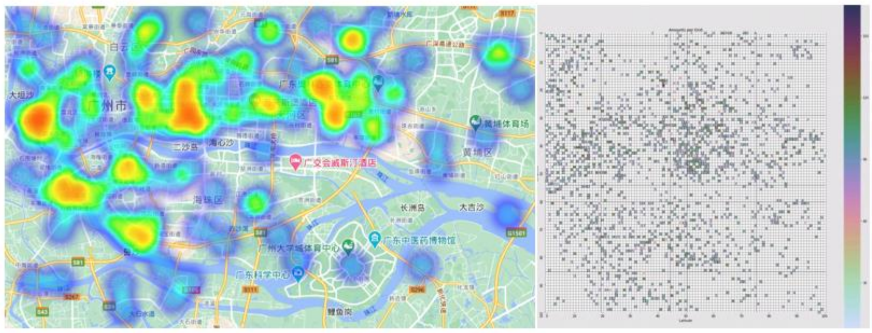
In summary, this report contributes to the following three aspects:

- Analyse the statistical correlation on the road traffic and communication traffic for V2X networks, which are shaped by both road network traffic and communication network traffic.
- Model road traffic for various CIV applications under urban scenarios, taking into account the space-time dependence.
- Develop deep learning algorithms to predict the complex long term spatial-temporal distribution of the road traffic, communication traffic and short-term CIV application traffic.

3. Dataset description

The dataset is collected from one of the major network operators named China Unicom in Guangzhou, a major city in China. We mesh the heatmap of the downtown, shown in Figure 2 (a), which is 22 km length and 15 km width as target area into 201 X 201 same size grids.

The road traffic dataset records 70 thousand activated user equipment (UE) logs in the same spatial-temporal range, including the anonymized IDs, location and visiting time. The crowd (or users) traffic is collected when the devices generate any type of data including call, SMS and network data in the certain area. Figure 2 (b) is the detail of grid heat-maps. The grid heat map takes more advantageous in terms of area size statistical and demographics.



(a) (b)

Figure 2 (a) The heatmap of crowd traffic in Guangzhou downtown (b) The grid heatmap of crowd traffic in Guangzhou downtown

We use WeChat to represent the communication traffic because it is the most frequently used APP in Guangzhou. The WeChat dataset shows around 0.82 billion counts of 'WeChat traffic stamps' in each grid in every hour from March 1st to March 31st, 2019. These stamps are obtained by slicing the traffic load evenly per two seconds, which can be seen as the equal size of data blocks.

4. Correlation analysis on the road traffic and communication traffic

4.1 Auto-correlation

To show the temporal correlation of mobile network traffic, we conduct the autocorrelation analysis on mobile network traffic. The autocorrelation is defined as:

$$r_{i,k} = \frac{\sum_{t=1}^{T-k}(d_{i,t+k} - \bar{d_\iota})(d_{i,t} - \bar{d_\iota})}{\sum_{t=1}^{T}(d_{i,t} - \bar{d_\iota})^2}, 0 \leq k \leq T$$

Where t denotes the time and $\bar{d_\iota}$ represents the mean traffic value of cell i.

Figure 3 shows the autocorrelation of the cellular traffic of a cell. We can see the autocorrelation reaches high with the lag of the multiples of every 24 hours, meaning that the mobile traffic patterns are similar from day to day. Therefore, the results of the autocorrelation analysis show that mobile network traffic is temporal correlated.

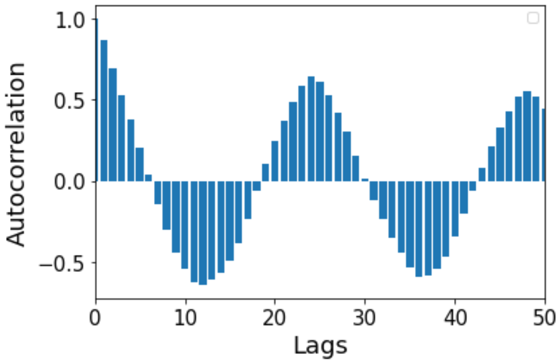


Figure 3 Auto-correlation of cellular traffic

4.2 Pearson correlation of road traffic

For the spatial correlation, we conducted Pearson correlation analysis, which is a widely adopted metric in measuring spatial correlations. The Pearson correlation is calculated as:

$$\rho_{d_i, d_{i'}} = \frac{cov(d_i, d_j)}{\sigma_{d_i}\sigma_{d_j}}$$

where $\sigma_{d_i}$is the standard deviation of cell i and $\sigma_{d_j}$is the standard deviation of cell $j cov(d_i,d_j)$is the covariance between d_i and d_j, which is defined as:

$$cov(d_i,d_j) = E\left[(d_i - \mu_{d_i})(d_j - \mu_{d_j})\right]$$

with μ_{d_i}being the mean of d_i, and E[.] represents the expectation.

The Pearson correlation ranges from -1 to 1, where -1 indicates entirely negative correlated, 0 means no correlation, and 1 represents entirely positive correlated. We analyze the Pearson correlation between the network traffic of a cell and its adjacent cells. Figure 4 shows the Pearson correlation of mobile network traffic of some cells, in the form of a heatmap. The diagonal elements represent the autocorrelation of the cell, whose Pearson correlation is always 1. The elements are closer to the diagonal means the shorter geographic distance to the diagonal cell. This heatmap shows that adjacent cells usually have higher correlations, indicating that the cellular traffic among neighbouring cells has similar traffic patterns. This shows that spatial correlations exist among cells.

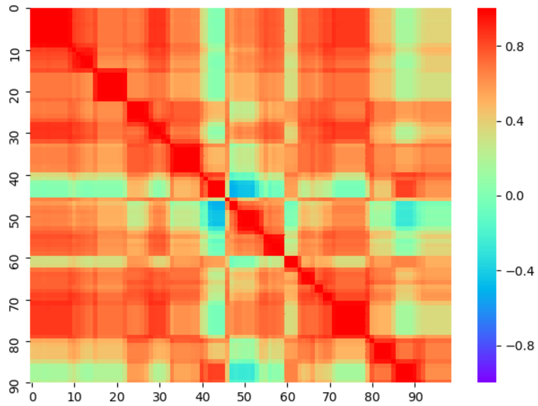
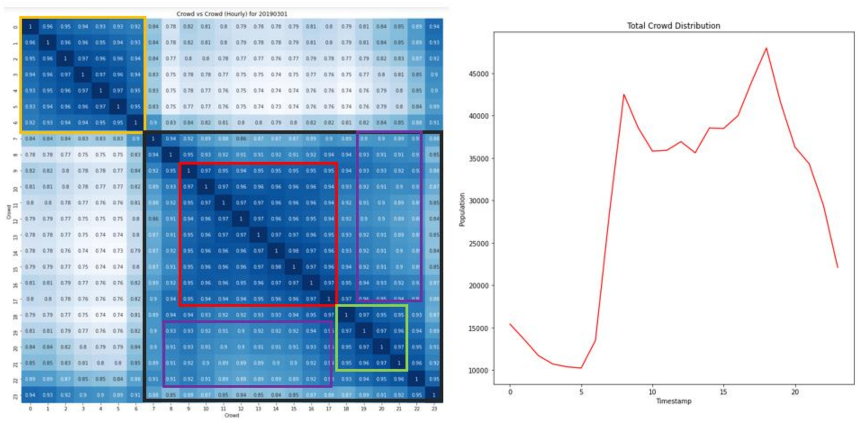


Figure 4 Spatial correlation heatmap

By meshing the crowd distribution map, we can quantify the change of population in the micro area (per grid) to discover the spatiotemporal features of data. Through the visualization of the heatmap, we will further quantify the patterns and find a suitable prediction method based on the analysis results.

Since the observation object is a heatmap of each timestamp, we first need to find a comparison method to quantify the similarity of the maps. Pearson correlation coefficient, a statistic for measuring vector similarity and correlation, is the method we applied. By converting a grid heatmap into a matrix and flattening it, we can transform this matrix into a one-dimensional array. By calculating the Pearson coefficients of the arrays, we can obtain the similarity of the heatmaps at each timestamp to figure out the periodic pattern of the population distribution.



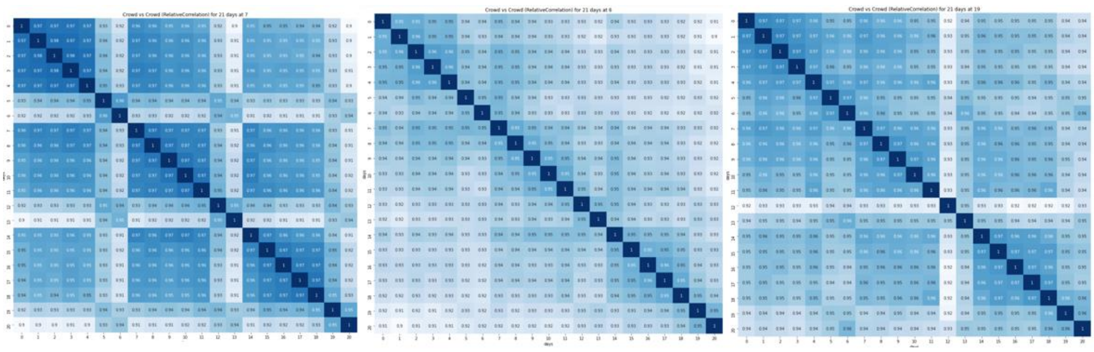
(A) (B)

Figure 5 (a) The similarity between each timestamp on 1st March (b) Total population on 1st March

Figure 5 (a) shows the similarity between each timestamp in a day. There are two crowd distributions which are before 7 am (yellow frame) and after 7 am (black frame). Moreover, the similarity coefficients from 9 am to 5 pm (red frame) and 6 pm to 9 pm (green frame) are higher, which means that 7 am is the turning points of urban population distribution, and the population distribution partially changes at 6 pm. The latter deduction is proved by the purple frames where the coefficient is lower than red and green frames. It verifies that the total population distribution has a periodic pattern, but the changes of some sub-regions are out-sync. Furthermore, as can be seen from Figure 5 (b), in a specific period, the population distribution is relatively fixed, while the number of users fluctuates. It indicates that not only do we need to use spatial algorithms to capture the characteristics of population distribution, but also need to use temporal algorithms to predict the specific population in each timestamp.

Figure 6 shows the similarities at three timestamps in 21 days, respectively. At 7 am, it can be clearly distinguished that nine square areas have higher coefficients, which represent 15 weekdays. It means that in the weekdays, the heatmaps at 7 am exhibit a similar distribution, that most users are in the workplace while weekends are not. According to Figure 7, from 7 am to 6 pm (working time), the similarity matrices are the same as Figure 7 (a). However, when the time is before 7 am or after 7 pm, similar to Figure 7 (b) and (c), there is no significant high coefficient area. The reason is that during this period, most users are distributed in residential areas, and such distribution will not change in working days or weekends. Hence from 8 pm, the high coefficient area of similarity matrix is gradually blurred.

Meanwhile, as the number of users decreases in the midnight (2 am to 5 am) and the randomness increases, the average coefficient drops down as well, which can be seen the blue frame and red frame in Figure 7. There are more details in Figure 7. The coefficient on Saturday is higher than it on Sunday, especially at 7 am and 8 am. On the other hand, the coefficient at 8 am in weekdays are the highest in every week. These require more perspective analysis, such as event detection or crowd behaviour observation, which can be associated with other work pages.



(a) (b) (c)

Figure 6 (a) The similarity at 7 am in 21 days (b) The similarity at 6 am in 21 days (c) The similarity at 7 pm in 21 days (Start from 11st March)

According to the above analysis, the correlation numerates obtained by the Pearson coefficient method can help us to understand the spatiotemporal features of Crowd traffic dataset. The population distribution changes periodically and depends on temporal attribution such as weekday or weekend, peak time or off-peak time. Meanwhile, once the population distribution stabilizes, it will continue for a period (hours) which means that the population fluctuation trends of each grid will be similar. The spatial algorithm can be applied to predict the target value through adjacent grids data. If predicting separately in temporal and spatial, it will be very complicated and time-consuming. Therefore, we need a spatiotemporal algorithm which can directly generate the grid heatmap as the result.

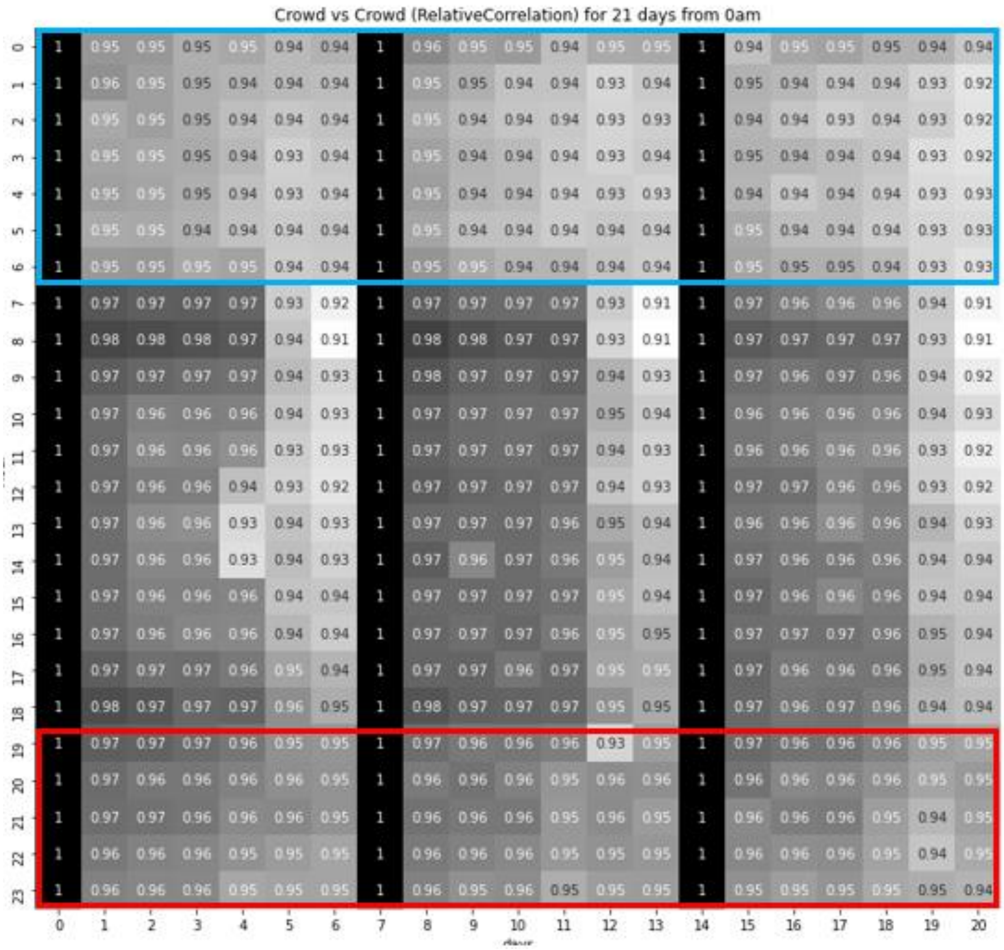


Figure 7 The hourly similarity between each Monday and the other six days in 21 days (Start from 11st March)

4.3 Pearson correlation of communication traffic (WeChat traffic)

WeChat traffic demand is highly bursty which can be drawn by comparing the crowd traffic in the same grid. We introduce conditional entropy function [1] to describe the randomness by quantifying the predictability of traffic. In Figure 8 (b), the entropy of crowd traffic as a comparison item is smaller than that of WeChat, which means that the communication application traffic series (WeChat traffic) is more difficult to forecast. Existing model-driven methods fail to model the burstiness factors in the sequences as aforementioned. Besides, machine learning-based methods like LSTM cannot achieve promising results either if they only depend on such high randomness sequences. More dimensional dependencies need to be explored to help model the burstiness.

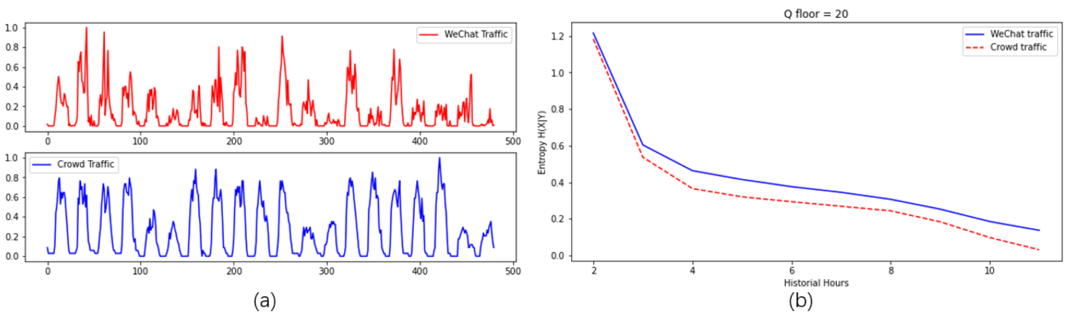


Figure 8 (a) WeChat traffic and Crowd traffic in a specific grid (b) The conditional entropies of WeChat traffic and Crowd traffic

Different from the high burstiness of data in time series, the spatial distribution of WeChat traffic reveals strong correlation in terms of Pearson correlation coefficient, which is a widely adopted metric for measuring spatial correlations. As can be seen from Figure 9 (a) and (b), it shows that WeChat distribution has a strong periodic correlation in specific time intervals like weekday and work hour, which is similar to the characteristics of crowd distribution mentioned in the previous section. It is a hint that modelling the data traffic burstiness from spatial dependence by training adjacent grids data together is a possible way to enhance the performance of prediction. The convolutional layer is suitable for exploring local grid spatial features, while it fails with modelling the temporal characteristics. Hence, the proper algorithm in IM traffic forecasting has the capability to simultaneously model the spatiotemporal dependence of data.

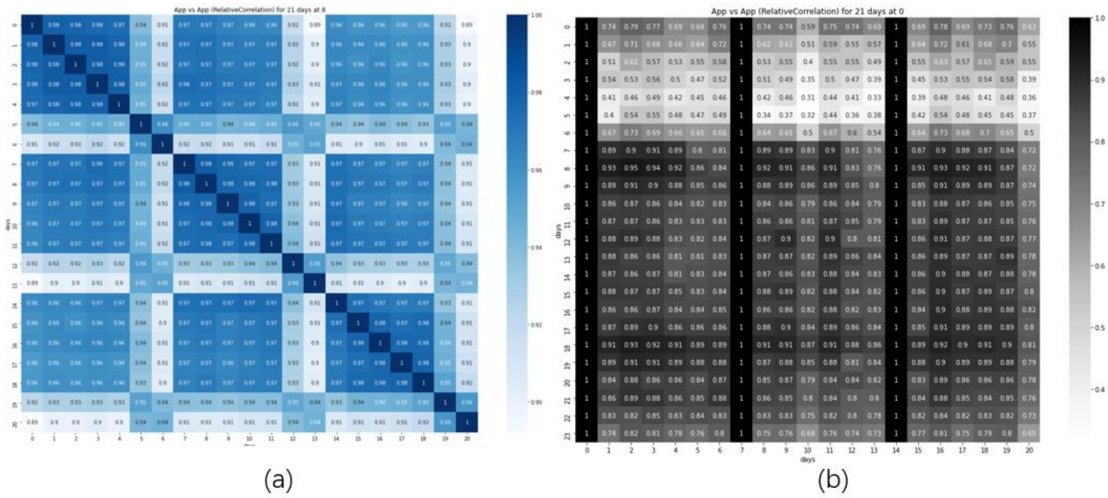


Figure 9 (a) The similarity at 8 am in 21 days of WeChat traffic (b) The hourly similarity between each Monday and the other six days in 21 days of WeChat traffic

4.4 Cross correlation between road traffic and network traffic

Figure 10 (a) shows the hourly correlation between WeChat and crowd flow in terms of Pearson coefficient in 21 days. It can be found that its distribution shares a similar pattern and periodicity with both WeChat distribution and crowd distribution. On the other hand, in Figure 10 (b), it illustrates that using population traffic to help predict next step WeChat data can increase predictability in terms of conditional entropy. Furthermore, applying the multi-grids population traffic as input can achieve the effect of optimizing WeChat forecasting performance as well which is shown in Figure 10 (c) . The above two perspectives could help to increase the possibility of burstiness detection especially when in WeChat traffic forecasting. Existing works introduce some external information like POI, BS distribution etc.. These contexts are treated as static, which is mainly used in labelling data and clustering in the model. These information fails to model the high burstiness of IM traffic caused by users' short distance movement. Oppositely, crowd distribution data is dynamic along with the network traffic which is an appreciate factor to explore the changes of IM traffic. Therefore, another synchronous training method should be proposed to improve the performance of the model.

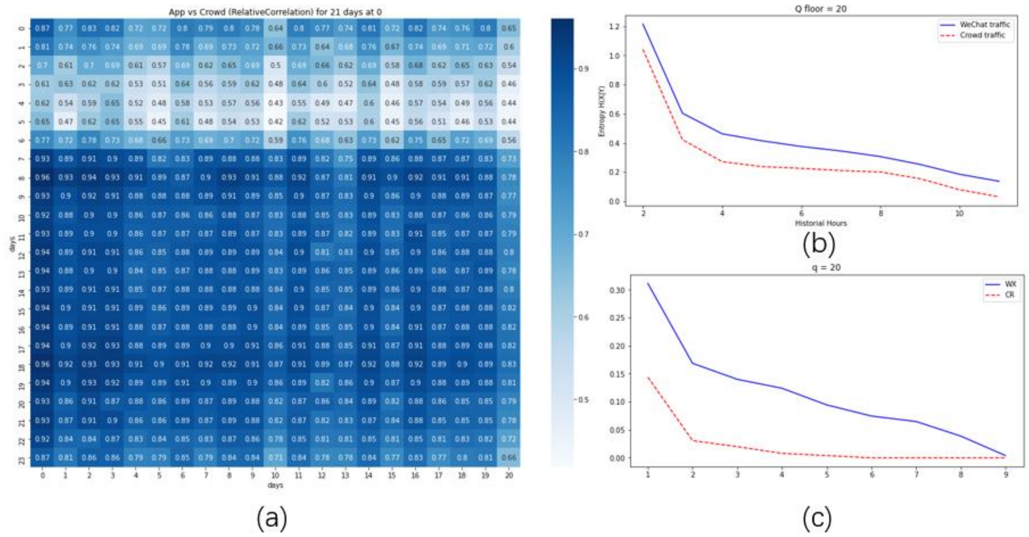


Figure 10 (a) Cross correlation between WeChat traffic and crowd traffic (b) The cross conditional entropies of WeChat traffic and crowd traffic to WeChat traffic (c) The cross spatial conditional entropies of WeChat traffic and crowd traffic to WeChat traffic

5. Modelling spatial-temporal traffic

5.1 General models (LR, ARIMA, SVR, RNN)

This part provides an attempt to model the temporal traffic of different regions. We provide Linear Regression (LR), Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR) as well as recurrent neural networks (RNN) to model the fluctuation of traffic. Then the performance is compared. To improve the performance of modelling and reduce the requirement of large amount of data, we also design a meta learning for this job.

From the temporal analysis in the last section, the mobile network traffic shows periodicity with the period of one day, this periodicity indicates that the mobile network traffic is predictable. We first adopt deep Long Short-Term Memory (LSTM) network [2], a kind of deep learning techniques, to address this. LSTM is a kind of recurrent neural network (RNN) The motivation behind this is that LSTM has excellent ability in modeling long-term dependencies. The input of the LSTM network consists of three features: one is the normalized traffic load, and the other two are the time of day (an hour in a day) and day of the week (from Monday to Sunday). The output is the predicted traffic load.

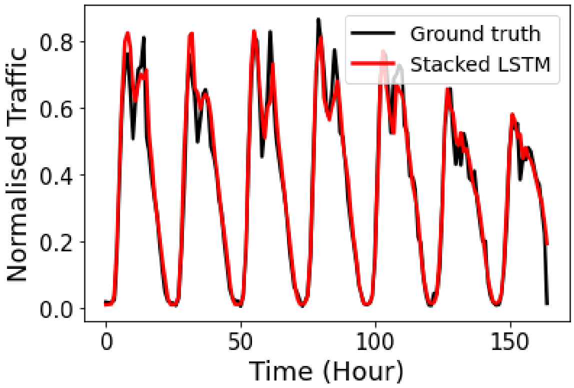


Figure 11 Predicted mobile network traffic using LSTM

Figure 11 shows the predicted traffic load by the LSTM, along with real traffic load. From the results, we can see that LSTM performs well in predicting cellular network traffic. To further evaluate its performance, we introduce several methods that are widely used in dealing with time-series prediction tasks for comparison. These methods include Linear Regression (LR), Auto-Regressive Integrated Moving Average (ARIMA), and Support Vector Regression (SVR). Two evaluation metrics are adopted: Root Mean Square Error (RMSE) and R-squared (R2), which are defined as follows:

$$NRMSE_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{d_{i,t}} - d_{i,t}\right)^2}$$

$$R2 = 1 - \frac{1}{T}\frac{\sum_{t=1}^{T}\left(\widehat{d_{i,t}} - d_{i,t}\right)^2}{\sum_{t=1}^{T}\left(\overline{d_{i,t}} - d_{i,t}\right)^2}$$

RMSE measures the errors between predictions and the ground truth, and R2 reflects the fitting degree between the ground truth and predicted values, ranging from $(-\infty, 1]$. Figure 12 shows the results of the comparison. It can be easily found that LSTM has the lowest prediction error and the highest fitting level. Thus, we can conclude that LSTM achieves the best performance in predicting network traffic load.

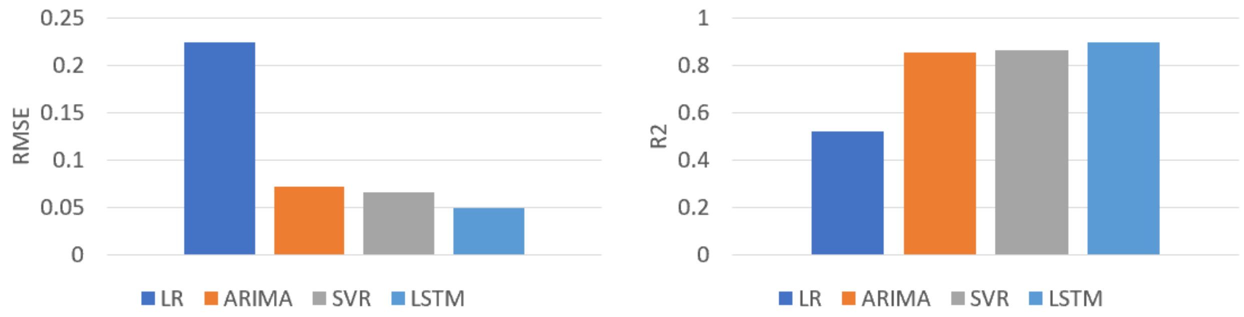


Figure 12 Performance comparisons between LSTM and other methods

Although the LSTM achieves competitive prediction performance, as a deep learning algorithm, it requires lots of samples to train the network and has high computation complexity. To improve the training and computation efficiency, we propose a new cellular traffic prediction framework based on the meta-learning technique, named as meta-learning based cellular level mobile network traffic prediction framework (MLCTPF). Meta-learning is a type of deep learning concept which is also named "learning to learn". It aims to exploit meta knowledge to enable the model to learn to solve problems. A meta-learning framework usually consists of two parts: multiple base-learners and a meta-learner. Base-learners are underlying learners, and each base-learner is designated to one learning task. A meta-learning system usually has many base-learners which learns a meta-level knowledge across tasks, and it produces the weight metrics which parameterizes the main policy, to achieve rapid generalization.

5.2 Improved model (meta-learning, KNN)

Motivated by this idea of meta-learning, we proposed a meta-learning-based framework. We employ LSTM networks as base-learners, and a K-nearest neighbors (KNN) based network as meta-learner. Each deep LSTM network addresses one traffic load prediction task, also named as the base-task, of one cell. For the meta-learner, it aims to learn to learn the proper prediction models for different base learning tasks. In addition, we select the frequency characteristics of the traffic load in the frequency domain as meta-features. More specifically, we have found that five pairs of frequency components are significant; thus we choose these five pairs as main frequency components, constituting the meta-features of each base task. We measure the similarity of the traffic patterns between cells by considering their meta-features as two vectors respectively and calculate the Euclidean distance between the two vectors. Closer distance means a higher similarity between the traffic patterns of the two cells. We have statistically and mathematically proved that when the distance of meta-features between two cells has a bound, transferring a well-trained LSTM model for one base task to another cell, the prediction error is also bounded.

For the base tasks handled by base learners, both their meta-features and model parameters of base-learners are accumulated as the meta-knowledge of the meta-learner. For a new base task, its meta-features are extracted and input to the meta-learner, then the meta-learner will utilize its meta-knowledge to predict the model parameters of a deep LSTM network for this base task.

To demonstrate the performance of our proposed framework, we compared the performance of the MLCTPF model with four baseline methods which are widely used in time series prediction tasks: linear regression (LR), auto-regressive integrated moving average model (ARIMA), support vector regression (SVR) and traditional LSTM (see Figure 13). We chose three evaluation metrics: normalized root mean square error (NRMSE), normalized mean absolute error (NMAE), and R-squared (R2). The first two metrics measure the prediction error, and the last metric evaluates the fitting degree between the ground truth and predicted values. The following figure shows the performance of our proposed model and baseline methods. We can find that the proposed MLCTPF has the lowest prediction error and the highest fitting degree among these methods.

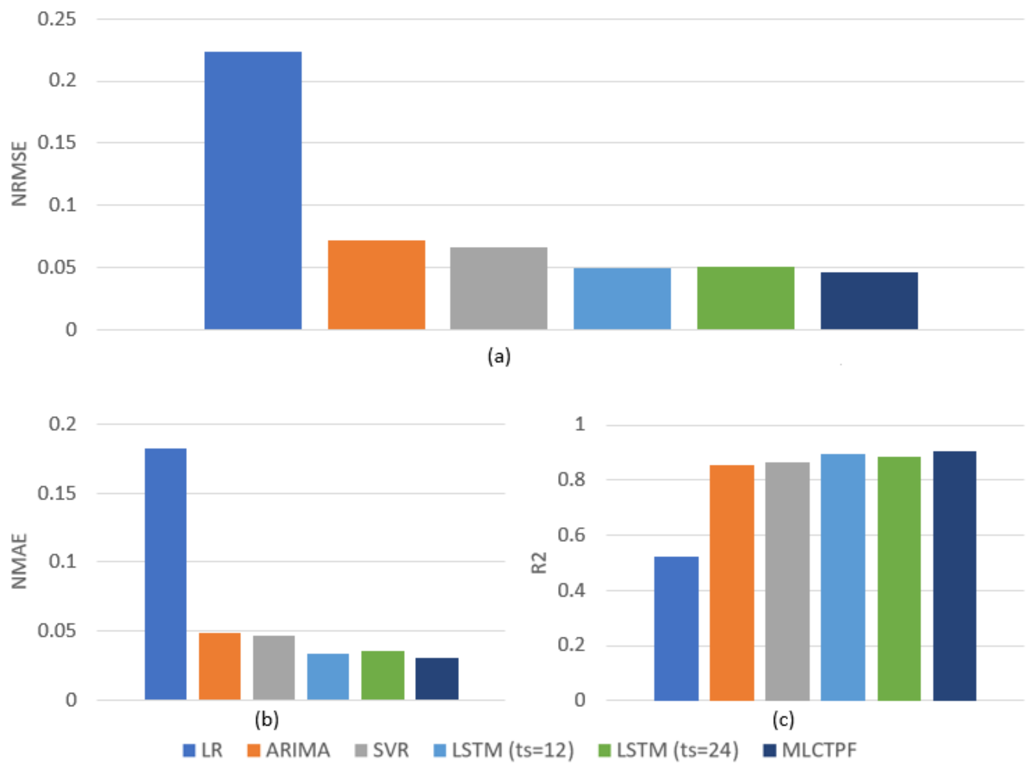


Figure 13 Performance of MLCTPF and the baseline methods in terms of (a) NRMSE; (b) NMAE 2; (c) R2

To further show the superiority of MLCTPF, we compare the training efficiency of MLCTPF with the traditional LSTM network. Figure 14 shows the change of the NRMSE of MLCTPF and LSTM during the training process, under two training data availability settings. The results show that, under both cases, MLCTPF converges much faster than LSTM, and achieves lower NRMSE. MLCTPF significantly reduces the number of epochs, meaning that the computation and training time is reduced; thus, a significant improvement in terms of learning efficiency is achieved.

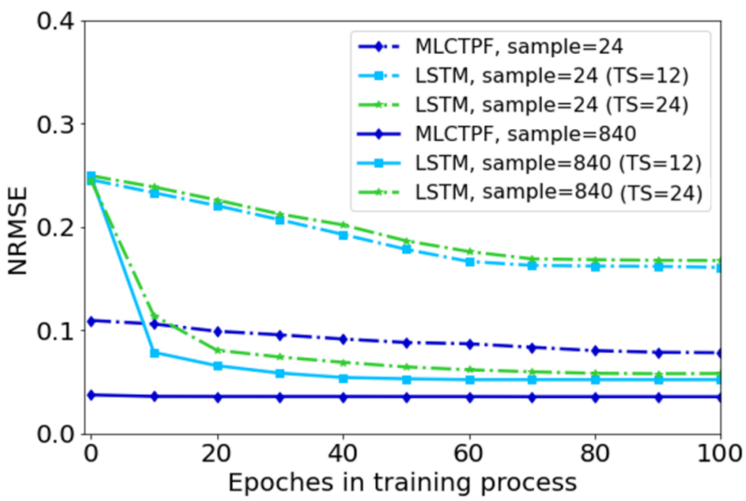


Figure 14 Performance of MLCTPF and deep LSTM networks with different number of training epochs

5.3 Road traffic event model (anomaly detection)

As the road-traffic management becomes more intelligent, recent researches began to pay attention to the data and information generated by citizens in the transportation process, and tried to use this information to discover and solve problems. There are two types of the traffic big data. One is the original traffic data that can directly reflect traffic conditions, such as cars' geographic location information, the traffic flow, etc.

The other is heterogeneous data that indirectly reflects traffic conditions and travelers' experience, such as traffic information on social networks. There exists discussion of the traffic problems on these platforms. The analysis of the former kind of data can reflect the problems existing in traffic from the specific behaviors, while the analysis of the latter one mainly focuses on the elusive factors such as travelers' views, feelings and attitudes under the traffic conditions, and tries to further analyze the existing problems that are frequently discussed. As a new communication channel, social media plays an important role in information communication. There are also peer-to-peer communication and group communication within social media, which fundamentally changes the form and efficiency of information dissemination. After observation, we conclude that social media carries the functions of information release, group discussion, topic discussion and official people communication in the communication of traffic related information. Social media has become an important intermediary channel for the release, dissemination and feedback of traffic information.

Data driven traffic management is indispensable for the above two types of data analysis. This work package focuses on the second type of data. We will crawl Guangzhou traffic related data information from Weibo, and then describe and analyze it. A Weibo big-data platform has been built to bring convenience for our researches. The process is shown in the following figure, two main processes are included: preparation and data analysis.

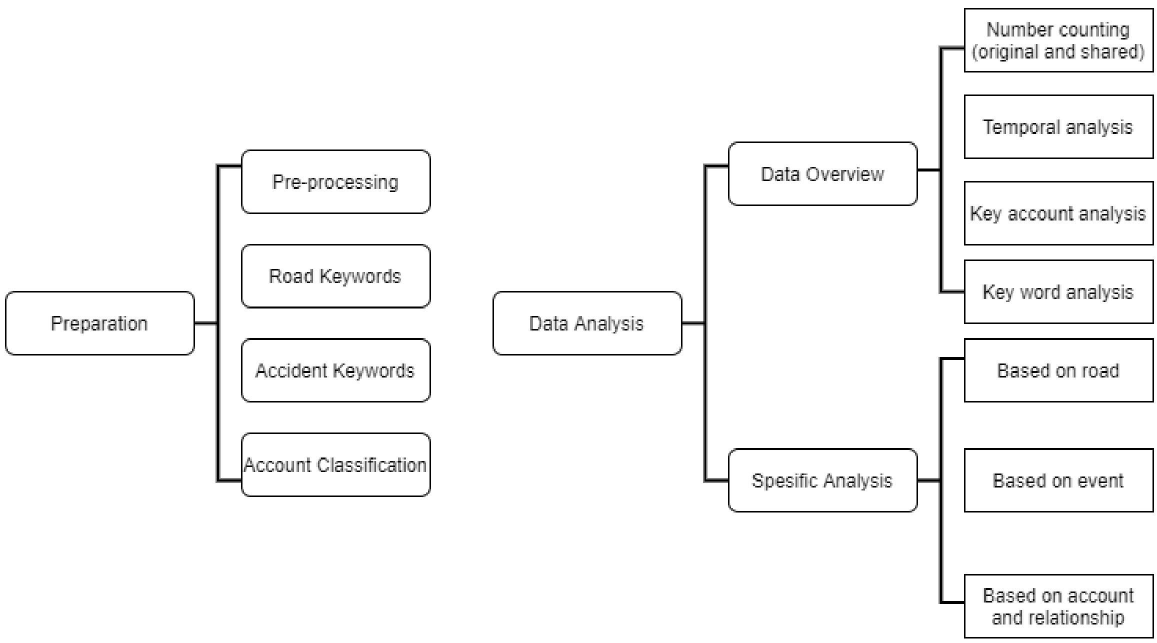


Figure 15. The processing of event classification from heterogeneous unstructured data

5.3.1 Social network data pre-processing

In this package, our purpose is to use online social network (OSN) data to investigate the status of Guangzhou traffic and citizen travel experience. The OSN used for traffic analysis in Guangzhou is Weibo, which is one of the most popular OSN in China.

By analysing the features of Weibo, we summarise the following aspects to be considered in the analysis:

1.The time Weibo is posted

(1) What happened during the period when large number of weibo was posted? (i.e. event tag)

(2) The relationship with holidays (for road conditions broadcast, traffic control)

(3) If special time periods outside the holidays exist

(4) If regularity exists

2.Geo-location when Weibo is posted

Extracting popular places from the geo-location information with the Weibo or location indicated in the Weibo content

3.Repost, like, and comment

(a)The comment information is considered as the most important factor among the three because it indicates the interaction behaviour among users.

(b)The user whose Weibo has high repost index.

4.User ID

We mainly focus on searching for the official account of the traffic police since some traffic police related accounts will post real-time traffic status

5.Weibo content keywords

With the help of Weibo content analysis of certain types of events, the common tag keywords of similar events can be summarized to facilitate subsequent detailed screening and building word cloud maps.

6.Weibo topic

The topics with the highest popularity are summarized and combined with aspect No.5 above for analysis.

We built the keyword library for the traffic analysis, and the following figure gives an example of the keyword library. Based on different logical combinations between phrases, including OR, AND, NOR etc, it can locate the Weibo content of reporting specific region traffic volume in Guangzhou from the database. As can be seen from the following figure, there are five labels in the library, which are city, area, warning keywords, traffic keywords and exclusion keywords. In the first two labels, there are many local street or district name in Guangzhou, which is to locate the content only relate with this city. The third and forth labels are discovering the various traffic accidents, like crush, traffic jam, no entry, etc. The difference between them is in the Weibo content, the warning keywords only refer to traffic accident, but traffic keywords may relate to other objects like movie posters, house selling, restaurant advertisement, etc. Therefore, the exclusion keywords are added to eliminate those 'noise data'. The main logic combination is shown as red.

地区 Regions （"广州" OR "广州交通" OR "广州交警" OR "天河" OR "白云区"OR "黄浦区" OR "越秀" OR "增城区" OR "荔湾" OR "番禺区" OR "花都区" OR "从化区" OR "海珠" OR "羊城" OR "白云机场" OR "广州南站" OR "广州北站" OR "广州东站"） 地点 Locations （"街" OR "道" OR "路" OR "桥" OR "广深" OR "广惠" OR "广珠" OR "出行" OR "路口" OR "站" OR "巷" OR "弄" OR "广场"）	Local street or district name in Guangzhou, which is to locate the content only relate with this city.

路况敏感词 (Road Traffic) AND (Real-Time OR Update OR Information) （"路况"）AND （"实时" OR "更新" OR "最新" OR "信息"） 交通状况 Traffic Conditions (e.g., car crash, traffic jam) （"事故" OR "车祸" OR "撞车" OR "刮蹭" OR "追尾" OR "交通事故" OR "施工" OR "修路" OR "道路维修" OR "绕行" OR "高峰" OR "塞车" OR "拥堵" OR "堵车" OR "行驶缓慢" OR "缓行" OR "慢行" OR "客流量大" OR "交通管制" OR "封闭" OR "限号" OR "限行" OR "通畅" OR "顺畅" OR "通畅"）	The various traffic accidents, like crush, traffic jam, no entry, etc.
排除项（非）Exceptions （"江西" OR "山东" OR "南京" OR "上海" OR "广西" OR "深圳" OR "陕西" OR "天津" OR "哈尔滨" OR "青岛" OR "楼盘" OR "房子" OR "广州路" OR "天河机场" OR "号线" OR "口感" OR "小吃" OR "标志" OR "酒店" OR "离开"）	the exclusion keywords are added to eliminate those 'noise data'.
地区 AND 地点 AND （路况敏感词 OR 交通状况） NOT 排除项 **Regions AND Locations AND (Road Traffic OR Traffic Conditions) NOT Exceptions**	The main logic combination

Figure 16: Keyword library for Guangzhou traffic analysis

According to the above keywords, our Weibo big-data platform finally collected 43598 traffic-related Weibos from 17308 users from Jan. 2018 to Jun. 2019, and it can also continue collecting real-time social network data if they satisfy our designed logic combinations. This dataset becomes the training and testing data of our designed event detection and classification models. Then, a basic NLP tool on the big-data platform is used to count the frequency of words and gives the following 'word cloud'. We can find that several key words have been frequently shown up in the contents of Weibo, like peak-time, car crash, high way, and police. It is important to fast track where and when these high-frequency words show up, then we can have early alerts of the traffic problems.

Here, we will use NLP to detect and classify events based on their relevant location and time (see Figure 17).

To detect the traffic events, we firstly designed an anomaly detection algorithm to find if some data becomes different than it usually should be. There are two main steps of anomaly detection:

• Model the regularity (trend and seasonality) of the training data set.

• Detect the anomaly by finding the outliers in the modelled regularity.

In this work, the temporal traffic is used to model the regularity and detect the outliers. The following figure illustrates the process. The detailed steps are shown as follow:

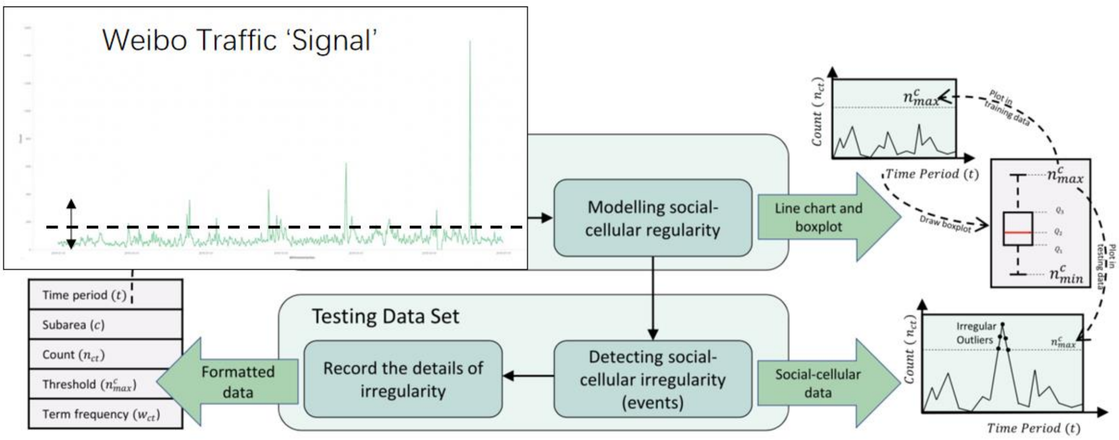


Figure 17 The process of building regularity and detecting network anomaly (irregularity).

Firstly, the Weibo data is counted along with time, so it is transformed to a new form, Weibo Traffic 'Signal'. This part denotes the number of Weibos as n and it changes along with time periods t.

Then a line chart of temporally changing traffic and its corresponding box-plot can be generated. Each box-plot offers a box of the majority and a maximum threshold ($n_{\max}^{c}$). Such a threshold describes the regular traffic range that regular social network traffic is lower than it.

Finally, in the testing data set, if current number of Weibo grows higher than the threshold. This time will be regarded as the start of an event. Then the algorithm automatically highlights the irregular outliers and records the details of these outliers, including the time period, subarea, count, threshold, and term frequency. The term frequency refers to the five most-appeared key works in Weibo.

Here, we applied this event-detection algorithm on the Weibo dataset and put forward the analysis of two typical events. In social media, the discussions should show up with reasons, hot events are the source of discussion. In this part of the analysis, we will select time nodes with an anomaly value (the number of traffic-related Weibo) from the temporal distribution as shown in the following figure. It is worth mentioning that due to the complexity of Weibo content, not all peaks in the figure are traffic-related events. We will conduct specific event analysis after verifying whether the specific event is traffic-related or not.

8.6.2018-11.6.2018 Guangzhou rainstorm: During the period from June 8 to June 11, 2018, Guangzhou experienced unusually heavy rainfall. Water accumulation and water logging in many areas severely affected transportation. Its temporal signal on Weibo is shown in the following Figure 18:

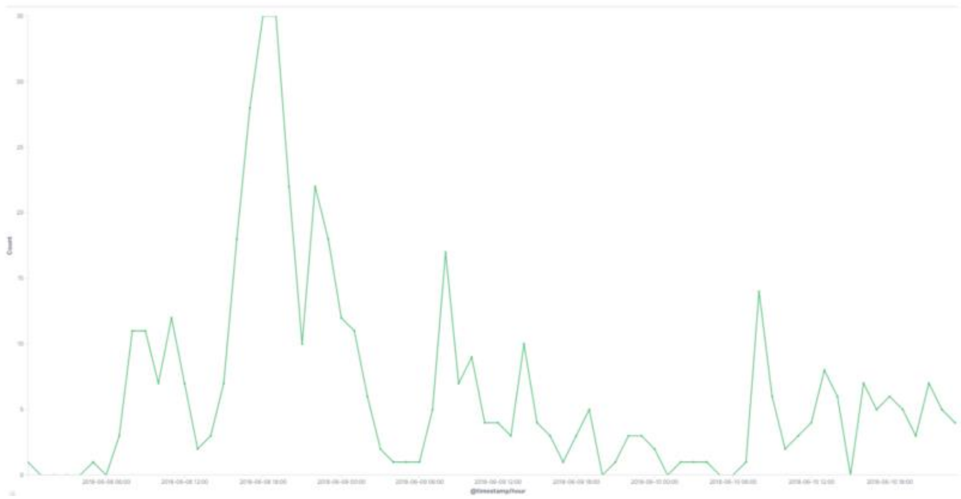


Figure 18. In terms of specific event heat time changes, the peak of the information release of the Guangzhou rainstorm was at 18:00 and 19:00 on June 8, and there were still slight fluctuations in the subsequent period.

Based on the NLP analysis, this Weibo with key word 'transportation' becomes an alert to travellers: 'In the past 21 hours, there have been heavy rainstorms in Guangzhou. The maximum accumulated precipitation reached 388.8 mm. Water accumulation and waterlogging occurred in many places, which seriously affected transportation. People who leave work must pay attention to safety. Don't forcefully pass through the section of stagnant water. Please make a detour or wait for a few hours.'

After the event detected, the information sources are checked, there are many official accounts that publish information, and from the proportion of Weibo account types as shown in the following figure, we can also see that the blue VIP users (the blue VIP account in Weibo is the official company account.) still occupy a large proportion (27.23%). Moreover, many users have participated in the information dissemination process in Weibo (49.78%), joined the discussion, and the proportion of information posted has exceeded the official users.

Figure 19. Weibo account/user analysis's screenshot of Weibo big-data platform.

From the following Sankey diagram, we can also see that the three official accounts of Guangzhou Weather, Guangzhou Traffic Police, and Guangdong Weather are the main information sources to alert the traffic problems influenced by the bad weather. Most of the forwarding starts from these three accounts.

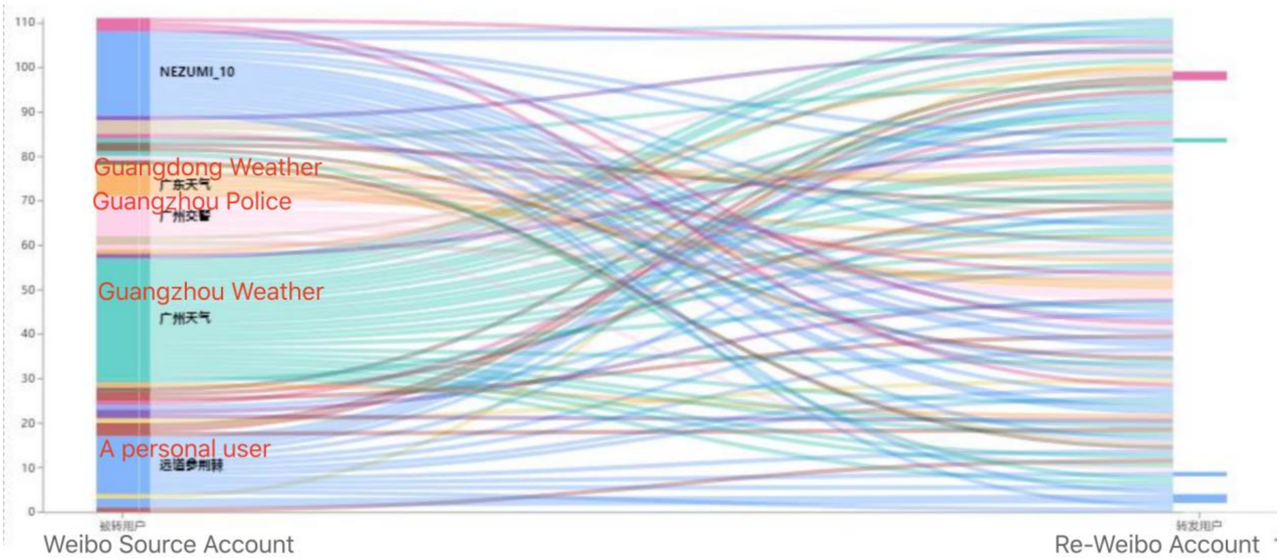


Figure 20. Sankey diagram of the Weibo about Guangzhou rainstorm.

From this phenomenon, we can find that in the information dissemination of some major events, official sources and private sources often attract public attention in different ways and serve certain social functions. In this case, the official account mainly plays the role of an information publisher, directly reminding the public of traffic inconvenience and safety by informing the authoritative information; the non-official account plays the role of an accident communicator, through the description of the specific accident attaching corresponding pictures to attract public attention. The non-official accounts have also served as an indirect reminder.

Summary: This event is different from usual and worthy of attention: (1) This hot event was caused by weather problems (natural disasters) (2) There were more individual/normal users in this event than in daily situations participated in dissemination and discussion (3) In terms of the proportion of information sources, official accounts and unofficial accounts each account for half of the country, and use different communication methods to achieve the function of notification and reminder.

21.5.2019 Serious traffic accident on Linhe Middle Road: On May 21, 2019, a serious traffic accident occurred near CITIC Plaza at Linle Intersection (Linhe Middle Road). The vehicle ran a red light and hit a pedestrian and two cars. The accident caused 13 people to be injured, 2 of whom were seriously injured. The event's Weibo signal is indicated as follow:

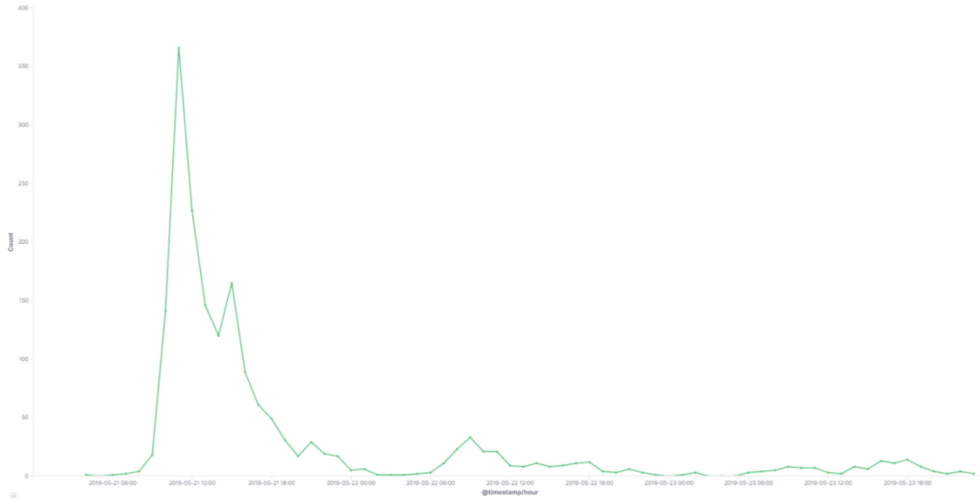


Figure 21. The Weibo popularity of the incident reached its peak at 11 o'clock on the 21st. Due to the sudden traffic accident and the seriousness of the injuries caused by the accident, the total number of discussions on that day was over 1,700.

By comparing the discussions caused by accidents and those caused by natural disasters, we found that the former's heat changes rapidly, and the heat dissipates very quickly, and it is hardly mentioned in the follow-up. Although the latter is not as popular as the former, there are multiple peaks during the entire event, and they continue to be paid attention to and mentioned. In that case, the alert to travelers should have different level according to the power of Weibo signals.

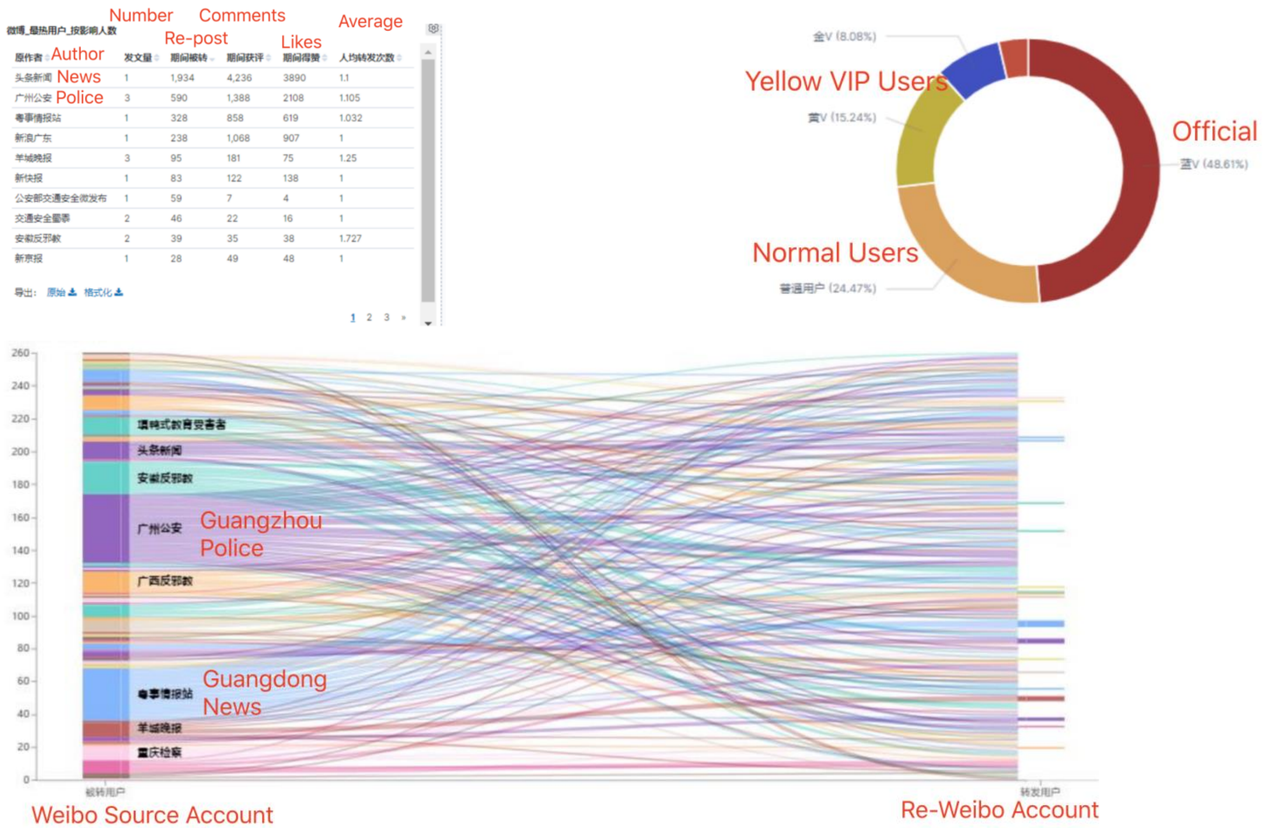


Figure 22 The stochastic information of Weibo and Sankey diagram.

Summary: (1) Accident's signal is different from the heat distribution of the previous event (2) in terms of information dissemination mode and the proportion of participating account types, it is very close to the normal of Weibo traffic information dissemination. (3) Everyone's concerns and topics are relatively concentrated. Using the heterogeneous data sets, we will detect events such as conferences, sports, concerts, etc.

This part uses Natural Language Processing (NLP)) on Weibo stream data and intelligently highlights the geographical key words, such as Tianhe Passenger Station and Zhujiang New Urban Area. This becomes the spatial dimension to characterize different types of traffic. Another is the temporal dimension to indicate when such a type of traffic emerges. Note that, the geolocation information comes from NLP not GPS, so the geographical accuracy should be influenced. An effective way to improve this aspect is to utilise heterogeneous data. We will classify the detected events according to time, location, and their key words.

Overview: Four Types of Traffic Based on Weibo Analysis: It can be seen in the following Figure 23 that different road sections are mentioned in Weibo with different temporal rules. According to the distribution of hotspots (red blocks), we further provide four types of different types of condition. They will directly reflect the traffic condition as well as users' experience according to only Weibo data.

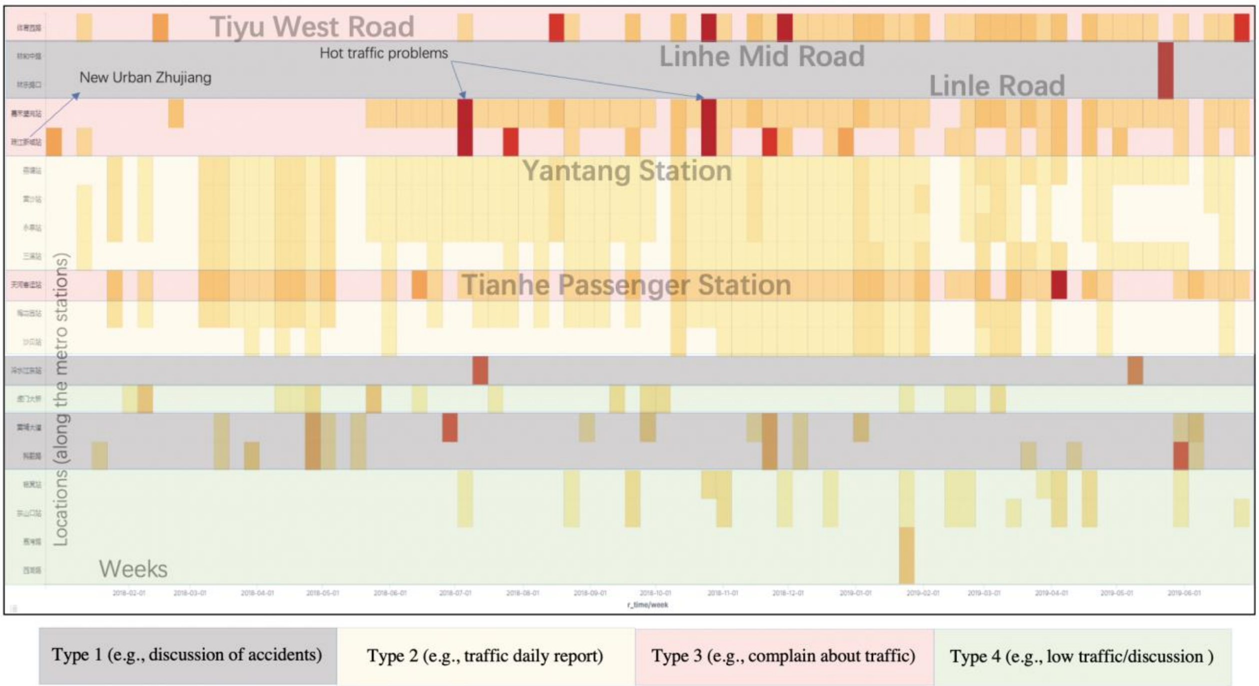


Figure 23. Four types of traffic based on Weibo discussion.

The four types are:

(1) Type 1 (Color Grey): A car accident is detected based on the concentrated discussion of Weibo users about the Linhe Mid Road and Linle Road. A typical characteristic of this traffic is that the heat being discussed is high and the time nodes are concentrated (can be indicated in the above figure). The red block represents a concentrated discussion of a particular traffic event at a particular area. And other blocks for light-colored or even colorless, because these locations are far from city centers and seldom appear on the heat distribution overview. In that case, the probability of an occurrence of accident becomes higher when a concentrated discussion appears.

(2) Type 2 (Color Yellow): The heat level of most locations is low (it is yellow in the above figure), and the time nodes are widely distributed along with the temporal dimensions. They appear in the visualization chart

that the area is covered by most light-coloured blocks, such as Yantang Station. In the previous report, a hypothesis is provided that this phenomenon is caused by daily discussion or mention about these locations. In that case, the core questions are to answer who participated in the discussion and what they were d iscussing. In the next part, we shall try to answer these questions.

(3) Type 3 (Color Red): This type is different from the type 2 because it contains red blocks which indicate that there might be events attracting high interests of Weibo users. In detail, there are both high-heated locations and low-heated dispersion locations. In the above figure, the light-coloured time nodes (yellow blocks) have higher coverage, and the dark-time time nodes (red blocks) appear around high-traffic regions, such as Tianhe Passenger Station, West Sports Road, Zhujiang New Town Station, Jiahe Wanggang Station, etc. These places are well-known for their high-load traffic because many citizens transfer to other transportations here. This report also aims to investigate the traffic of this place.

(4) Type 4 (Color Green): This type owns low daily discussion and no red blocks, so it is not the interested parts. Based on the experience of citizens, the places in this type own low probability to get high-load traffic. Besides, it becomes the type 1 if there is accidents happening.

In our previous report, we provided two deductions or hypothesis:

Hypothesis for Type 1: Some serious traffic incidents that occurred during a certain period of time on a section with a high degree of traffic heat and a relatively concentrated distribution attracted public attention;

This has been verified that we can use the irregularly increasing Weibo discussion to find the car accident.

Hypothesis for Type 2: The broader sections are the objects that are often mentioned in daily discussions.

Here, one more hypothesis is added:

These places are discussed in the daily life for the high traffic volumes, and the accident or traffic jam in Type 3 will be alerted by the red blocks.

6. Deep learning in road traffic prediction

As aforementioned, population distribution traffic has strong correlation in spatial-temporal. Therefore, a deep learning prediction architecture based on spatial-temporal features modelling, is introduced and the diagram is displayed in Figure . Convolutional Long Short-Term Memory (ConvLSTM) can capture the spatiotemporal nature of the data by replacing the inner dense connections of LSTM with convolution operations, which helps to reduce a giant number of parameters in LSTM caused by modelling spatial dependency. Thus, our model is mainly based on ConvLSTM. The crowd distribution data is given as a three-dimensional matrix which is a sequence of distribution grid heatmap in hourly timestamps denoted $D = \{D_1, D_2, \ldots, D_T\}$. The operations of a single ConvLSTM is specified in [3]. $\sigma(.)$ denotes the activation function, which is the sigmoid function, $\odot$ denotes the Hadamard product and $*$ represents the convolution operation.

$$i_t = \sigma(W_{xi} * D_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i),$$

$$f_t = \sigma(W_{xf} * D_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f),$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tanh(W_{xc} * D_t + W_{hc} * H_{t-1} + b_c),$$

$$o_t = \sigma(W_{xo} * D_t + W_{ho} * H_{t-1} + W_{co} \odot C_t + b_o),$$

$$H_t = o_t \odot \tanh(C_t).$$

In the above, $W(\cdot\cdot)$ and $b(\cdot)$ denote the weights and biases which are obtained through model training. Note that besides of the inputs D_t, the gates C_t, H_t, i_t, f_t, o_t in the ConvLSTM's inner structure are all 3-D tensors. With configuring these gates, the model is capable to 'learn to forget' spatiotemporal features, which is beneficial in traffic forecasting.

In addition, 3D-Conv layer, by extending 2D-Conv with few previous sequence data [4] to observe the minor fluctuations of the sequences, e.g. motion capturing [5], is added into ConvLSTM layer. Since the geographical distribution of the crowds is periodic and the pattern turning time is relatively fixed, 3D-Conv can help to strengthen the detection of fluctuation points by slicing the traffic sequence.

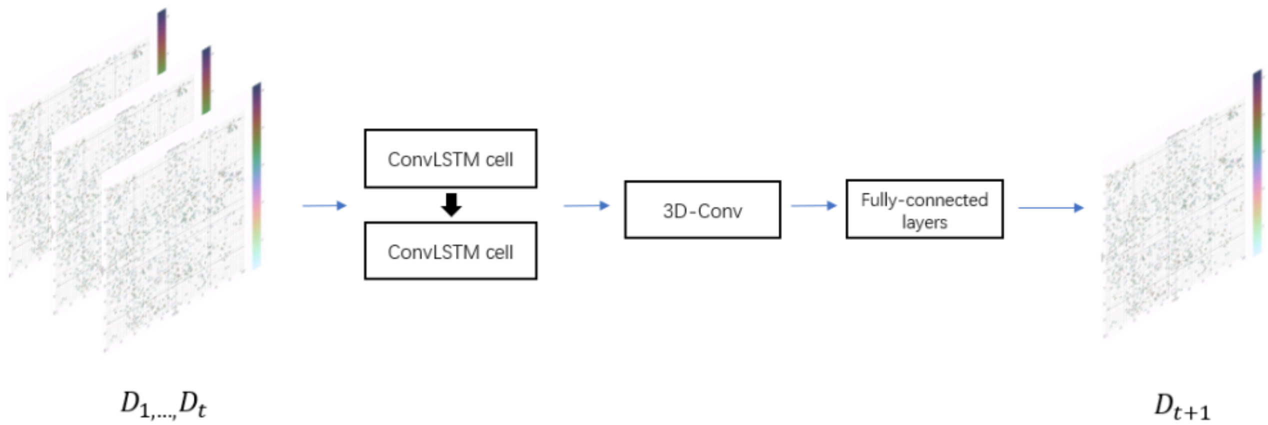


Figure 24 The architecture of proposed model in crowd traffic forecasting.

Through the observation of the entropy function, it finds that UEs traffic data shows a strong predictability, which helps the proposed model to easily achieve ideal results. On the other hand, combining with other algorithms to improve the performance might not be worth its cost.

6.2 Deep stack learning framework in IM traffic prediction

Our proposed framework named as deep Stack Learning-based framework for IM Traffic Prediction in cellular mobile network (SLIM-TP), see Figure 24. There are three components: basic learner module, which consists of three basic learners to forecast the traffic from three perspectives including spatial, temporal dependencies and external information (crowd traffic), evaluation module which provides the relative accurate possibility of results generated by three former learners, and meta-learner module which derives the final results by modelling the correlation between ground truth and the previous results.

Two three-dimensional data metrics are involved as input into this framework, which is WeChat data traffic in a small area (containing multiple grids) in two weeks and the same as the range of UEs traffic. At first, the basic learner module is proposed for training the metrics of WeChat and population traffic distribution snapshots and the traffic flow of WeChat traffic in target grid as inputs respectively to forecast its future series of WeChat traffic. Secondly, the evaluation module is cited to obtain the expected value and the confidence interval to quantify the relative likelihood for the predicted results generated by the previous three learners, which is treated as the auxiliary feature. Then, three predictions, three likelihood values and ground truth data at each timestamp are concatenated as a rebuilt training dataset step by step. Finally, the meta-learner module handles these rebuilt training arrays to provide the improved forecasting results.

In conclusion, SLIM-TP consists of two layers and three modules. The basic learner module in the first layer aims to explore more dependencies and forecast the traffic respectively. Meanwhile, an evaluation module based on GPR algorithm is to measure the reliability of previous results as an auxiliary feature. Then, as a stacking architecture, a MLP based meta-learner module in the second layer is fed an aggregated training dataset as input to establish a mapping relationship between the results of previous modules and the ground truth value to improve the system performance by modelling the internal correlations.

6.3 Experimental results and analysis

6.3.1 Road traffic prediction results

We will introduce normal ConvLSTM model as a comparison method to demonstrate the performance of the proposed model in geographical distribution forecasting. Each predicting snapshot will be transformed as a one-dimensional matrix and then compared to the ground truth. We choose Root Mean Square Error (RMSE), which quantified the errors between predictions and the ground truth, as the evaluation method. Furthermore, the prediction accuracy of the user traffic in each grid is considered as well by introducing LSTM as the comparison algorithm. Because in the previous report, we have proved that LSTM generated more proper results in aggregate SMS traffic than other methods like Linear Regression (LR), Auto-Regressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR).

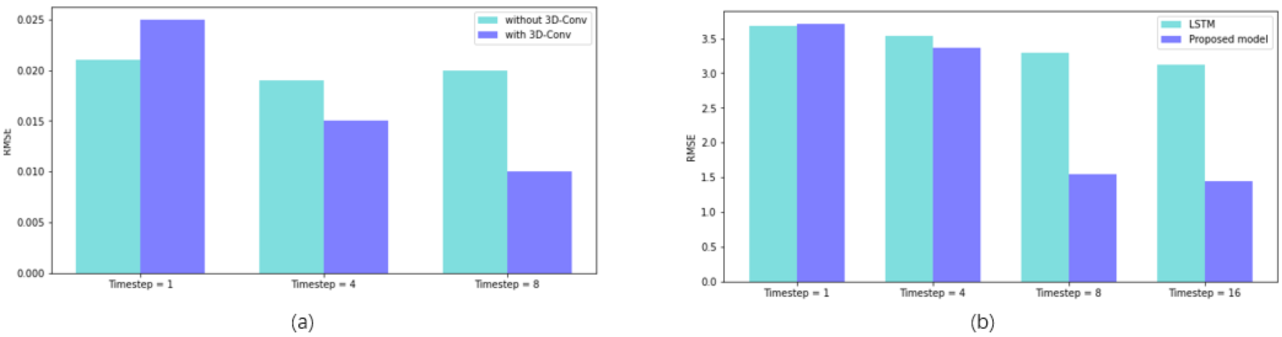


Figure 25 (a)Average RMSE between ConvLSTM only and proposed model in 480 timestamps (b) Average RMSE between LSTM and proposed model in six grids in 480 timestamps.

As aforementioned about the characteristic of 3D-Conv, our proposed model is supposed to enhance the capability in the minor fluctuation observation, but it is based on the multiple previous data consideration. Therefore, a key hyperparameter needs to be mentioned as the variable of environment setting which is the timestep. Time step is the sequence length that how many past data points will be considered in addition to the current data point to make the prediction. As can be seen in Figure 25 (a), when the timestep is 1, predicting by only current data, normal ConvLSTM model has a better result, however, the situation is reversed when the time step increases. The graph clearly explains that compared to normal ConLSTM model, the more historical data are involved, the more sensitive detection our proposed model has. Besides, the time-series traffic prediction in grids is shown in Figure 25 (b). We average the RMSEs of six grids traffic prediction results. The performance of our proposed model is depended on the timesteps as well, the similar capability with LSTM under one timestep and the better capability when timesteps increasing.

Another factor that needs to be introduced which is training time consumption. In Figure 26, it is clear that LSTM has less time cost, however, this is counted for only one grid training. The latter two models have longer training time but are handling with 441 (21*21) grids. Although the standard ConvLSTM has least time consumption, it is acceptable considering that our proposed model has much more improved performance than that in time cost.

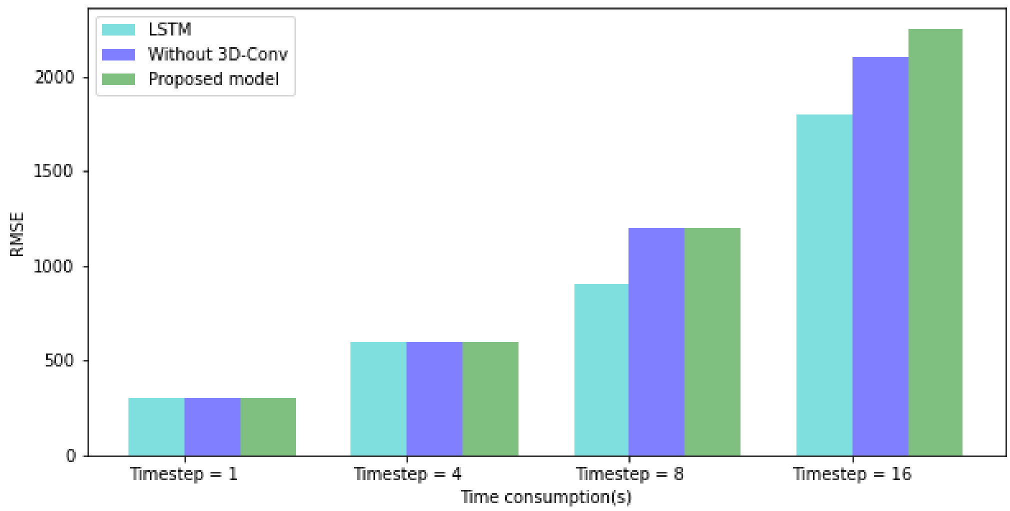


Figure 26 Training time consumption of three models

6.3.2 Communication application traffic prediction results

To evaluate the performance of our proposed model, not only do we compare to the three basic learners in our framework, but also other to the widely used methods in time-series prediction field. We randomly select ten non-zero target grids as targets for traffic prediction. Normalized Root Mean Square Error (NRMSE) is adopted to provide a comprehensive evaluation of different prediction algorithms which presents a better performance in a smaller value. In our proposed model, the basic learners for exploring spatiotemporal characteristics is composed of two layers of ConvLSTM and 3D-conv and a fully connected neural network.

Three basic learners: As mentioned above, We have provided the reason for choosing these learners, which aims to deeply explore and model various dependencies, so as to further optimize the accuracy in WeChat (IM) data forecasting through stack learning architecture. Therefore, these three basic learners are the prior candidates for evaluating our proposed model.

LSTM: LSTM is currently the most widely used neural network for time-series prediction.

Linear Regression: LR is a representative of shallow machine learning methods.

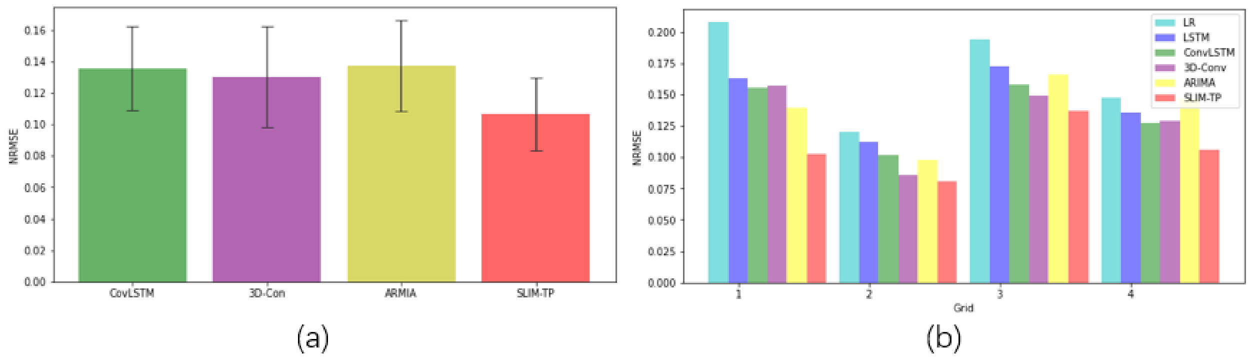


Figure 27 (a) Overall performance on selected grids in NRMSE (b) Performance comparison between SLIM-TP and baselines on four grids

As shown in Figure 27 (b), the prediction result of the classical method LR has the largest value in terms of RMSE in each target grid. Specifically, due to the burstiness in the sequence of the WeChat data traffic, LR has poor performance in the modelling. For the LSTM algorithm, compared to the LR, it generally reduces the result by about 10%, which proves its effectiveness in predicting complicated sequences. However, since only the temporal feature is explored, it has the limitation in further increasing the accuracy of highly bursty traffic forecasting. The ConvLSTM-based model can further decrease the RMSE value by around 5%-7%, which proves the effectiveness of spatial feature in optimizing the prediction. Since the correlation between two types of data (WeChat and crowd traffic) is not stable in time-series, the results are generally slightly higher or lower than the results generated by the previous algorithm. However, the overall result shows that among these ten testing grids, using crowd traffic to predict WeChat traffic is usually more efficient than using WeChat data as input around 7%. ARIMA's forecasting accuracy rank fluctuates, which may be related to the traffic characteristics. Besides, due to the sliding window method applied, the computational complexity of ARIMA is reduced. Therefore, in our system, ARIMA is used in the next step prediction instead of the multi-step prediction commonly used in other articles, which is profitable in the prediction accuracy enhancement. Among the evaluation results of all the target grids, our model performs better than other methods which proves the high effectiveness of high volatility traffic modelling. On the other hand, compared with the basic learners, SLIM-TP has been indeed proved the capability to further improve the accuracy of prediction.

In conclusion, by introducing the consideration of spatiotemporal dependence, user traffic as external information and the confidence probability of basic learner module results, our proposed model SLIM-TP outperforms the traditional methods (for example, LR over 40%, LSTM over 30 %) and the results generated by the basic learner module in the framework (over 20%).

7. Conclusions

In this report, we presented the main research works conducted towards WP5 tasks of our project COSAFE for the development of V2X Traffic Modelling and Prediction model by utilising the machine learning and deep learning. We analyzed the statistical correlation on the road traffic and communication traffic for V2X networks, which were shaped by both road network traffic and communication network traffic. Moreover, we modelled road traffic for various CIV applications under urban scenarios, taking into account the space-time dependence. In the last part, we have developed deep learning algorithms to predict the complex long term spatial-temporal distribution of the road traffic, communication traffic and short-term CIV application traffic.

In conclusion, the WP5 has achieved good progresses with the extensive collaborative research activities. The proposed algorithm and mechanisms provide optimal performance in the desired scenario. The tuning time of the proposed algorithm is significant reduced to that can greatly improve the adaption in different scenario. They have been smoothly deployed in the real scenario, and remarkable performance is obtained. The successful research and development on Task 5.1 provide a strong foundation that will benefit the remaining tasks within WP5.

Reference:

[1] X. Zhou, Z. Zhao, R. Li, Y. Zhou, and H. Zhang, “The predictability of cellular networks traffic,” in 2012 International Symposium on Communications and Information Technologies, ISCIT 2012, 2012, pp. 973–978, doi: 10.1109/ISCIT.2012.6381046.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," in Neural Computation, vol. 9, no. 8, pp. 1735-1780, 15 Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[3] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” Adv. Neural Inf. Process. Syst., vol. 2015-Janua, pp. 802–810, 2015

[4] C. Zhang and P. Patras, “Long-term mobile traffic forecasting using deep Spatio-Temporal neural networks,” Proc. Int. Symp. Mob. Ad Hoc Netw. Comput., pp. 231–240, 2018.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.