Detecting and Analyzing Urban Regions with High Impact of Weather Change on Transport

Ye Ding, *Member, IEEE*, Yanhua Li, *Senior Member, IEEE*, Ke Deng, *Member, IEEE*, Haoyu Tan, *Member, IEEE*, Mingxuan Yuan, *Member, IEEE*, and Lionel M. Ni, *Fellow, IEEE*

Abstract—In this work, we focus on two fundamental questions that are unprecedentedly important to urban planners to understand the functional characteristics of various urban regions throughout a city, namely, (i) how to identify regional weather-traffic sensitivity index throughout a city, that indicates the degree to which the region traffic in a city is impacted by weather changes; (ii) among complex regional features, such as road structure and population density, how to dissect the most influential regional features that drive the urban region traffic to be more vulnerable to weather changes. However, these two questions are nontrivial to answer, because urban traffic changes dynamically over time and is essentially affected by many other factors, which may dominate the overall impact. We make the first study on these questions, by developing a weather-traffic index (WTI) system. The system includes two main components: weather-traffic index establishment and key factor analysis. Using the proposed system, we conducted comprehensive empirical study in Shanghai, and the weather-traffic indices extracted have been validated to be surprisingly consistent with real world observations. Further regional key factor analysis yields interesting results. For example, house age has significant impact on the weather-traffic index, which sheds light on future urban planning and reconstruction.

Index Terms—Trajectory analysis, weather-traffic index, traffic prediction, urban computing

1 Introduction

Urban computing connects urban sensing, data management, data analytic and service providing into a recurrent process for an unobtrusive and continuous improvement of people's lives, city operation systems and the environment [2]. The aim is to solve a variety of emerging city problems, such as traffic congestion, energy consumption, and pollution, based on the data of traffic flow, human mobility, and geographical data, etc. In particular, many works have been done to investigate the impact of inclement weather to traffic [3], [4], [5]. For example, a heavy rain may slow down the traffic and cause congestions due to low visibility and high demand of vehicles; the decreasing temperature in very cold days will freeze the roads and influence the transport performance, etc. Table 1 describes the general relevance of the impact of weather change to transport in US.

In July 21st, 2012, Beijing faces its largest rainstorm since 1951, with an average rainfall of 164 millimeters. According to the news report [7], there are 77 people died in this catastrophic natural disaster. The transport of Beijing suffered from various contingencies due to the serious flood, as shown in Fig. 1. During that time, a variety of photos titled "see the sea in Beijing" widespread on the Internet. This disaster not only shows the serious problems of the urban transport system of Beijing, but also inspires our research interest: how can we identify those regions being highly influenced by weather change on transport?

The early works often focus on the correlation of weather and traffic in some particular roads where devices have been deployed to continuously collect traffic data. By analyzing the traffic change in different weather conditions, the traffic prediction can be better preformed considering the weather forecast. However, the weather-traffic correlation covering most roads throughout a city (known as *regional weather-traffic sensitivity index* or for simplicity *weather-traffic index* (WTI)) is still an open problem vain in spite of the practical value in our daily life. One essential reason is the lacking of effective traffic monitoring system in city-wide scale. Another open problem is how to disclose the key factors behind the weather-traffic index, to explain the reason why some regions in a city are more vulnerable to inclement weather and others are not. These factors are the regional features including the density of roads, the number of road intersections, the number of points of interest (POIs), the traffic volume, the average age of the household, the density of buildings and more in the surrounding regions. The weather-traffic index throughout a city and the knowledge of key factors behind the correlation provides effective support to help government agent to understand the functional character of districts throughout a city, to improve traffic performance and to learn the key factors in urban planning,

- *Y. Ding and H. Tan are with Guangzhou HKUST Fok Ying Tung, Research Institute, Hong Kong University of Science and Technology, Kowloon, Hong Kong. E-mail: {yeding, haoyutan}@ust.hk.*
- *Y. Li is with the Computer Science Department, Worcester Polytechnic Institute (WPI), Worcester, MA 01609. E-mail: yli15@wpi.edu.*
- *K. Deng is with the School of Computer Science and Information Technology, RMIT University, Melbourne, Vic 3000, Australia. E-mail: ke.deng@rmit.edu.au.*
- *M. Yuan is with Noah's Ark Lab, Huawei Technologies Co., Ltd., Kowloon, Hong Kong. E-mail: yuan.mingxuan@huawei.com.*
- *L.M. Ni is with University of Macau, Taipa, Macau. E-mail: ni@umac.mo.*

Manuscript received 25 Nov. 2015; revised 20 July 2016; accepted 18 Oct. 2016. Date of publication 31 Oct. 2016; date of current version 7 June 2017. Recommended for acceptance by Y. Zheng, C.T. Silva, R. Ghani, and C. Mascolo. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TBDATA.2016.2623320

TABLE 1
Changes in Climate and Weather Relevant on US Transport [6]

Change in Climate or Weather	Likelihood
Decreases in very cold days	Virtually certain
Increases in Arctic temperatures	Virtually certain
Later onset of seasonal freeze and earlier onset of seasonal thaw	Virtually certain
Sea level rise	Virtually certain
Increases in very hot days and heat waves	Very likely
Increase in intense precipitation events	Very likely
Increases in drought conditions for some regions	Likely
Changes in seasonal precipitation and flooding patterns	Likely
Increases in hurricane intensity	Likely
Increased intensity of cold-season storms, with increases in winds and in waves and storm surges	Likely

etc. For example, if the traffic of a region is highly affected by heavy rains, and the key factors include the sewer system, then it is important for the government to examine and improve the sewer system of the region in first priority.

To enable weather-traffic index throughout a city and factor analysis, the effective traffic monitoring in city-wide scale is a must. Nowadays, with the widely commercial use of taxi tracking system, the most feasible means probably is to extract traffic information from numerous taxis driving on roads due to its availability, wide-coverage and low-cost. A taxi tracking system combines the use of automatic vehicle location in individual vehicles with software that collects these fleet data. Typically, taxi data continuously record the information including location, speed, occupancy status, and orientation of the taxis. The traffic parameters (e.g., traffic speed) extracted from taxi data are practically sparse since the number of taxis in a city is typically limited. Therefore, we partition the city by Voronoi diagram where the seeds are the road intersections. Compared to the region-oriented city map partition approach such as equal-sized rectangles [8] where the roads in some cells are highly dense and in others are highly sparse, the advantage of our road-intersection-oriented partition makes sure every cell include at least one road intersection and a number of roads connected to this intersection. Given a period of time, the average parameters of driving taxis in each Voronoi cell (or called *cell* for simplicity in the rest of this work) is extracted as the average traffic parameter of the cell. In addition to traffic data, weather data and complicated regional features in the same period of time are also required to perform the study.

This work has developed a *weather-traffic index system* which mainly aim to fulfill two tasks. The first is to set up a weather-traffic index throughout a city, which indicates the impact of weather to traffic from light to heavy. The second is to reveal the key factors behind the weather-traffic index throughout the city and their relative weights. Although there are many existing traffic prediction and measurement works as introduces in related works, they mainly focus on the analysis of road segments; on the contrary, this paper is the first study on local traffic-weather sensitivity throughout a city and the investigation to reveal the key factors behind the sensitivity.

Fig. 1. The rainstorm of Beijing in the year of 2012.

We have addressed a series of techniques challenges in this work, and the central contributions are summarized as follows:

- A systematic approach has been proposed for establishing weather-traffic index throughout a city. The challenge is to separate the impact of weather to traffic from many other reasons. The other reasons include the traffic in peak-hour differing from that in non-peak hours, the traffic, for example, 5 minutes ago in the nearby road networks, and the road works slowing down the average speed, etc. A novel method has been proposed to successfully address this challenge.
- A supervised learning method have been proposed to disclose the key factors and their weights contributing to weather-traffic index throughout the city. It is a challenging task because many factors have composite and delicate influence concurrently.
- Using the proposed system, we conduct empirical study in Shanghai, the largest city in China, using 115.2 GB traffic data (extracted from more than 4,000 taxi trajectories) for two years, the weather data of the same period of time, the road networks and complicated regional features. The established weather-traffic index and the discovered key factors have been extensively verified against the observations in the real world.

In the rest of this paper, we outline the related works in Section 2, and show the framework of the proposed system in Section 3. Then, the data preparation component of the system is introduced in Section 4, the weather-traffic index establishment component is presented in Section 5, and the factor analysis component is detailed in Section 6. We conduct empirical study using the proposed system in Section 7. Finally, this work is concluded in Section 8.

2 Related Work

Urban computing works often focus on a particular city problem, such as traffic congestion, energy consumption, and pollution, based on the data of traffic flow, human mobility, and geographical data, etc. For example, in [8],

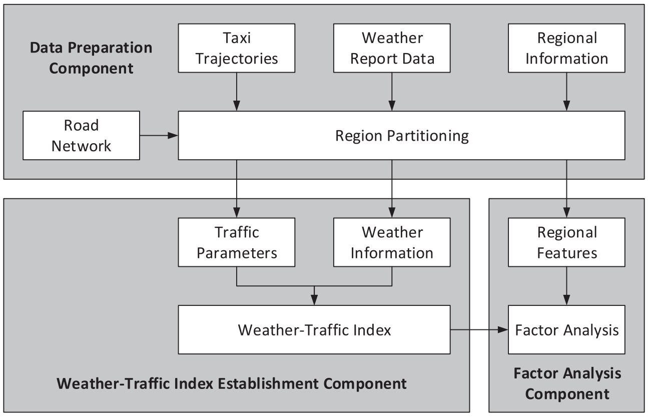


Fig. 2. Framework of weather-traffic index (WTI) system.

they inferred the real-time and fine-grained air quality information throughout a city, based on the air quality data reported by existing monitor stations and a variety of data sources observed in the city. In [9], they tried to identify the hot spots of moving vehicles in an urban area via a novel, non-density-based approach, called mobility-based clustering. In [10], they proposed a framework, called DRoF, to discover regions of different functions in a city using both human mobility among regions and points of interests located in a region. In [11], the authors tried to sense the refueling behavior and citywide petrol consumption in real-time, based on the trajectories of vehicles. In [12] and [13], they tried to discover the traveling companions and gathering patterns of vehicles, respectively.

Being an important topic in urban computing and cross-domain data analytics [14], the early research on the relation between weather and traffic is mainly based on quantitative analysis and statistical methods. For example, in [3], they presented an algorithm for forecasting physical road surface conditions based on weather and road surface data they have collected, and aim to identify icy roads during a cold weather in advance in order to predict the impact to traffic. In [4], they proposed a crash-likelihood prediction model based on both real-time traffic flow variables measured through series of underground sensors and the rain data collected at weather stations in order to alarm potential crash occurrence in advance. In [5], they developed a neuro-wavelet prediction algorithm to forecast hourly traffic flow considering the effect of rainfall. The experiments show that the rainfall data successfully augments the traffic flow data as an exogenous variable in periods of inclement weather. The early works focus on some particular roads where devices have been deployed to continuously collect traffic data. None of them investigated the weather-traffic correlation throughout a city and conduct analysis of the key factors behind the regions whose traffics are highly influenced by inclement weather.

3 Overview

This work aims to develop a *weather-traffic index system* which performs two tasks: establishment of weather-traffic index throughout a city and analysis of key factors behind the index. The framework of the proposed system is shown in Fig. 2, which consists of three functional components:

- *Data preparation:* The road networks in the city of interest is partitioned into cells via Voronoi diagram where the seeds are road intersections. For each cell, the traffic parameters are extracted from taxi trajectories and the regional features are collected. The weather information of the same period of time is also collected. The details are presented in Section 4.
- *Weather-traffic index establishment:* The weather-traffic index is established for each cell by analyzing traffic data and weather data. In specific, given a cell g, the weather-traffic index $\rho(g)$ is a value indicating the extent to which the traffic parameter in g is affected by weather. This component is discussed in Section 5.
- *Factor analysis:* The input includes the established weather-traffic index and the regional features. The aim is to identify which regional features make traffic in cells vulnerable to inclement weather. In particular, the weights of regional features are quantitatively measured. The methodology is provided in Section 6.

4 Data Preparation

In this section, we introduce the data preparation component which partitions the city into fairly distributed regions, and collects relevant source data for each region.

4.1 Region Partitioning

A straightforward region partitioning method is region-oriented partitioning such as in [8] where the city region is split into equal size grids. However, this partitioning method is improper if the traffics of road networks in grids are concerned. The reason is that the road networks in a city are often distributed unevenly. For example, the road networks are typically much denser in the urban areas than that in the rural areas. As a consequence, the road networks in some grids are highly dense and in some grids are highly sparse. This situation motivates us to apply a different region partitioning method.

Our method is to partition the city region using Voronoi diagram [15]. A Voronoi diagram is a partitioning of a plane into regions (or *cells*) based on the distance to points (or *seeds*) in a specific subset of the plane, and the shapes and sizes of the cells differ from each other. In this paper, we choose road intersections as the seeds. We call such partitioning method as *road-intersection-oriented partitioning*. In particular, if several road intersections are very close to each other, for example within 50 meters, they are grouped together as a complex intersection. So, each cell includes at least one road intersection and the road segments connected with this intersection. The indices of all cells are obtained following the equal procedure no matter they are in dense and non-dense areas.

The road-intersection-oriented partitions in Shanghai is shown in Fig. 3 where the seeds are the intersections of major roads. We observe that the cells are relative small in the urban areas while the cell tends to be large in rural areas. This partitioning method has two desirable properties. The first is the relatively even distribution of road networks in all cells. The second is that traffic jam, in particular in extreme weather condition like a thunderstorm or a heavy rain, often happens in the road intersections

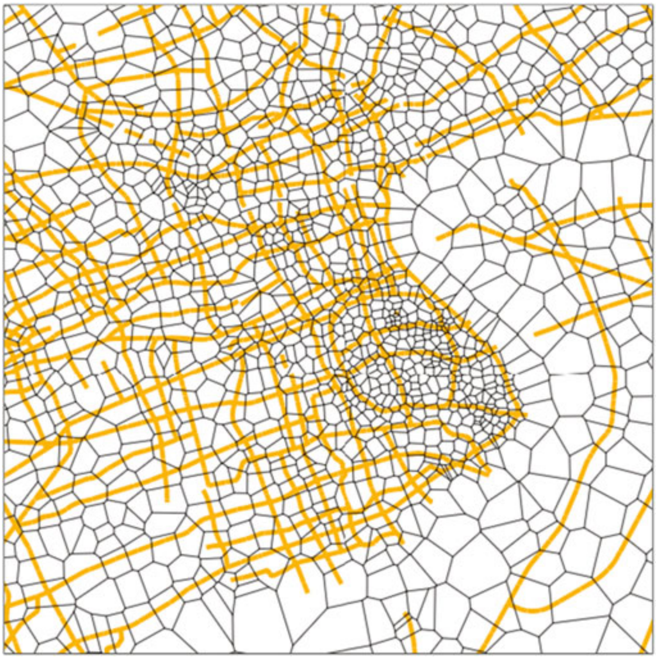


Fig. 3. The Voronoi diagrams partitions in Shanghai. The under layer represents the road networks.

according to our experience in daily life. In other words, the road-intersection-oriented partitioning method helps to portray the relation of weather and traffic investigated in this work.

In this paper, Voronoi cell is the unit region of weather-traffic index. For each cell, traffic and weather of a long period of time are analyzed to decide the weather-traffic index.

4.2 Source Data

The input of the system includes the road networks, traffic data, and regional features in the city of interest, and the weather data in the same period of time. A road network $G(V, E)$ consists of a set of road segments E and a set of road intersections V. A road segment in E is associated with its type, length, speed limit, two end points and other meta information. A road intersection in V is associated with its location (i.e., latitude and longitude) and type. The carriageway between two road intersections in E may consist of multiple edges in E connected in sequence.

From traffic data, a certain traffic parameter of interest, such as average speed, can be extracted. Traffic parameter can be classified in terms of one of the following: quantity measures, e.g., "how much or at what rate is traffic moving or waiting to move?"; quality assessment measures, e.g., "how well is traffic moving?"; movement measures, e.g., "where is traffic coming from and going to?"; and composition / classification measures, e.g., "what kind of traffic is moving?". While all kinds of traffic parameters can be applied in our weather-traffic index, this work use average speed as an example. Speed expresses the rate at which traffic is moving and, therefore, is a natural measure of the quality of the flow.

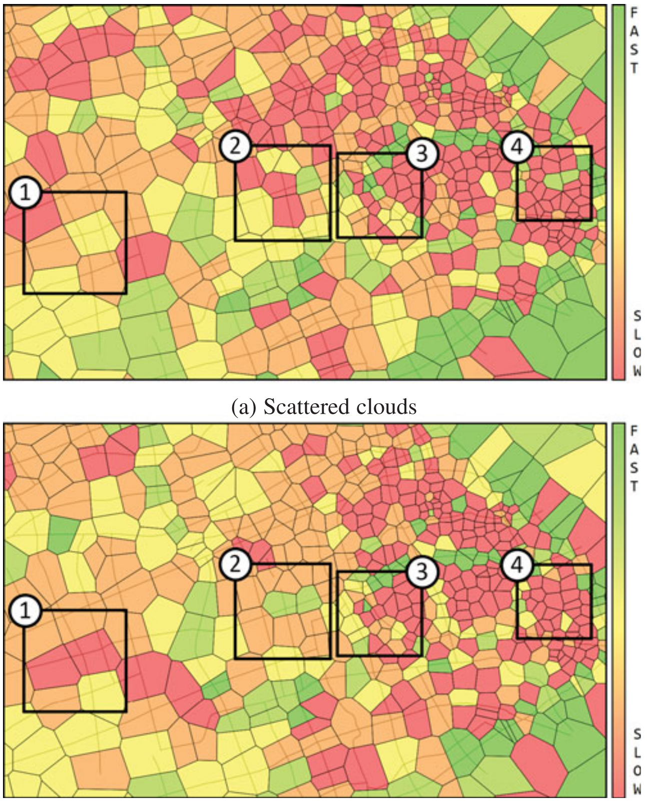
In this work, *time mean speed* (also called *average speed*) is used as the traffic parameter, which is defined as the arithmetic mean of individual spot speeds that are recorded over a selected time period. An adequately sized sample of spot speeds is needed to ensure that the time mean speed approximates the population mean to within the desired accuracy. The traffic parameters are extracted from large volume of taxi trajectory data collected. In our study, the average speed of the driving taxis in each cell are calculated. In particular, the average speed is split into seven classes: less than 10, 10-30, 30-50, 50-70, 70-90, 90-110 km/h, and more than 110 km/h. Since traffic parameters are categorical results and our objective is to establish index, we split the continuous variables because 1) it reduces the complexity of the problem, and 2) it well supports our objective. In particular, if continuous values are used, the main ideas proposed in this work are still applicable with trivial modification. The average speed of one road segment is subject to the traffic parameter of that road segment only, which is not comparable with other road segments. For example, the average speed of 30 km/h reflects drastically different traffic condition on a small local street and a highway. Hence, in this paper, we only compare the changes of the average speed on each road segment separately.

TABLE 2
Specifications of Weather Report Data

Attribute	Description
Time	Time of the weather report.
Temperature	Temperature in Celsius degrees.
Dew Point	The temperature at which the air must be cooled for water vapor to condense, forming water droplets, fog, or clouds.
Humidity	The relative amount of water vapor in the air.
Wind Speed	Speed of wind shown in km/h.
Wind Gust	The maximum wind speed in km/h.
Wind Direction	The direction of wind in degrees.
Visibility	The ability to see an object in the atmosphere in km.
Pressure	The Atmospheric air pressure in millibars.
Wind Chill	The perceived decrease in air temperature felt by the body on exposed skin due to the flow of air.
Heat Index	An index that combines air temperature and relative humidity to estimate the human-perceived equivalent temperature.
Precipitation	The condensation in mm of atmospheric water vapor that falls under gravity, including drizzle, rain, sleet, snow, etc.
Condition	Weather condition, e.g., clear, rainy, and cloudy.
Extreme Weather	Indicator of a fog, rain, snow, hail, thunder, or tornado.

Weather is the state of the atmosphere, to the degree that it is hot or cold, wet or dry, calm or stormy, clear or cloudy. The details of the weather data used in this paper are described in Table 2 in Section 7.1. If a particular weather condition is interested, such as rain, the weather-traffic index can be specialized as rain-traffic index and accordingly the factor analysis is specialized to rain as well.

For each Voronoi cell, the complicated regional features are collected including four categories in the surround regions. The details of the regional features used in this paper are described in Table 4 in Section 7.1.



(a) Scattered clouds

(b) Thunderstorms and rain

Fig. 4. The average speed at 14:00 on two days in summer in Shanghai, where the weather is scatter cloudy (top) and rainy (bottom).

5 Weather-Traffic Index Establishment

The weather data and traffic data from data preparation component is the input of weather-traffic index establishment component. The intuition that the traffic is influenced by weather can be proven by the example shown in Fig. 4. It illustrates the average speeds in different cells in Shanghai at the same time slots in two different weather conditions: cloudy and rainy. It is clear that the average speed in rainy days is generally lower than that in cloudy days. At the same time, it also demonstrates that the average speeds in some cells are unchanged in the rainy days and in cloudy days. Therefore, weather-traffic index is necessary to indicate the impact of weather to traffic in different cells.

Given a cell g, its value in weather-traffic index is the correlation between traffic and weather, denoted as $\rho(g)$. $\rho(g)$ takes value from a discrete range, such as $[1, 2, 3, 4, 5]$. The following section will discuss how to detect such correlation.

5.1 Correlation Detection

In a cell, for detecting the correlation between the traffic speed, denoted as F_t, and weather, denoted as F_w, a simple method is to train a classifier which infers directly from F_w to F_t, as shown in Fig. 5. The input is the weather represented as a feature vector and the output is one of the seven speed classes. The trained classifier is tested. If the inference accuracy is high, it means the correlation between the traffic and weather is high in this cell; otherwise, the correlation is low. This method is commonly used in statistics to measure the correlation between two random variables.

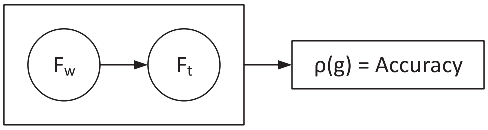


Fig. 5. A simple weather-traffic correlation detection method where traffic speed is directly inferred from weather.

However, we observed critical weakness of this method in correlation detection between traffic and weather. This is because there are many other reasons which impact traffic. For example, the traffic in peak-hour differs from that in non-peak hours, the traffic accident in one road segment will influence the traffic in nearby road networks, and the road works slow down the average speed, etc. Compared to weather, these reasons are dominant in most cases. Therefore, the main challenge in weather-traffic index establishment is to separate the impact of weather to traffic in each cell from other reasons.

To address this challenge, we propose a novel method inspired by the Granger causality test [16]. The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. A time series X is said to *Granger-cause* Y if it can be shown that those X values provide statistically significant information about future values of Y. Hence, we say that a variable X that evolves over time *Granger-causes* another evolving variable Y if predictions of the value of Y based on its own past values and on the past values of X are better than predictions of Y based only on its own past values.

In this paper, the initiative is to train a traffic prediction model which considers all other reasons besides weather, and then train a traffic prediction model which considers all other reasons and weather. We observe the difference between the inference accuracies of the two models. If the accuracy is improved after considering weather, it indicates that the weather does impact the traffic in this cell in general; otherwise, the impact of weather is uncertain in this cell. The overview of our method is shown in Fig. 6. The traffic prediction models are trained separately in different time slots. The reason is that the traffic regularity in time

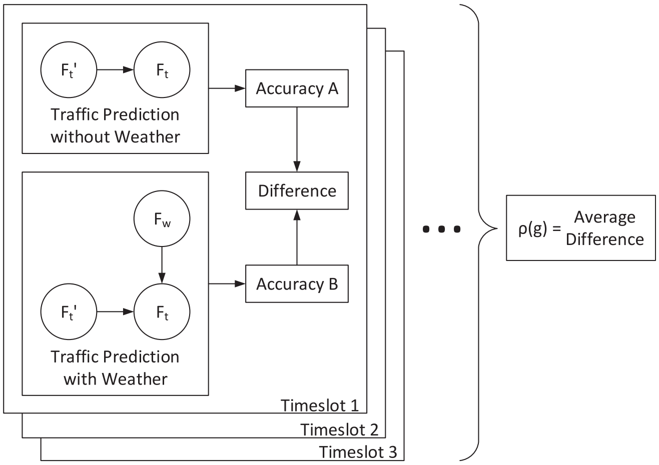


Fig. 6. The weather-traffic correlation detection method used in this paper.

slot, for example, 7:00-9:00 am can be very different from another time slot, for example 9:00-11:00 am. As shown in Fig. 6, the average of the traffic prediction accuracies in different time slots is used.

The weather-traffic index value $\rho(g)$ is assigned to each cell to indicate the extent to which the traffic prediction accuracy is impacted by weather as discussed above. After considering weather, in some cells the traffic prediction is strongly improved and in some cells the traffic prediction is weakly improved. The cells are organized in ascending order of the traffic prediction accuracy improvement, and then they are divided by k-quantiles, i.e., dividing the ordered cells into k equal-sized subsets. Thus, the k-quantiles show the correlation between traffic and weather from weak to strong. The motivation of quantiles is because the cells are essentially normally distributed and a large percentage of cells are close to the mean. By using k-quantiles, the number of cells in each subset is about equal.

Due to the requirement of a traffic prediction model in weather-traffic index establishment, the following section will discuss traffic prediction in details.

5.2 Traffic Prediction

Traffic prediction is a well studied problem. Since early 1980s, univariate time series models, mainly Box-Jenkins Auto-Regressive Integrated Moving Average (ARIMA) [17] and Holt-Winters Exponential Smoothing (ES) [18], have been widely used in traffic prediction. In the last decade, neural network models have also been used in forecasting travel time [19]. In [20], spatial-temporal characteristics of traffic events are considered in training traffic prediction models. In [21], authors use AQ21, a natural induction system that learns and applies attributional rules, to predict traffic by autonomous agents within a vehicle route planning system. In [22], they estimate the traffic flow of a road segment by analyzing taxi trajectories. A recent study successfully uses the weather situations as supplementary information in traffic prediction model to enhance the prediction accuracy [4]. In this work, any traffic prediction model can be used.

In this work, the traffic parameter of interest is consisted of discrete classes, thus we treat traffic prediction as a classification problem. To be robust, we use three different linear inference methods, including support vector machine (SVM) [23], logistic regression (a.k.a. logit) [24], and perceptron [25]. The average accuracy of a 10-fold cross-validation is used to compute the accuracy difference as shown in Fig. 6. Our framework is compatible to various inference models to infer WTI. In this paper, we use support vector machine as an example because SVM is both suitable for time series prediction [26] and adopted in some works on weather-traffic inference [27]. We use logistic regression and perceptron, both of them are popular linear models, to verify the output of support vector machine. We conclude the weather-traffic index for one cell only when all three models indicate the similar results. In the rare case that the results of three models are not consistent, a special value is assigned to the cell in the weather-traffic index to indicate the uncertainty of the correlation between weather and traffic.

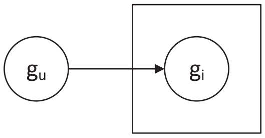


Fig. 7. Weather-traffic index inference from adjacent cells.

6 Factor Analysis

In this section, our discussion is based on the assumption that the weather-traffic indices of all cells have been certainly assigned.

The weather-traffic index indicates which cells are correlated with weather in terms of traffic. It provides the possibility for us to investigate the key factors behind the correlation. The factors are the *regional features*, denoted as F_r, as shown in Table 4 in Section 7.1. The factor analysis identifies the key factors and their weights contributing to the weather-traffic indices of cells. In other words, it discloses what regional features make the traffic in some cells vulnerable to inclement weather.

6.1 Key Factor Verification by Index Inference (KFVII)

Given a set of regional features, our methodology verifies they are the key factors based on the following intuition. The weather-traffic index of one region can be inferred from the indices of its closely located (or *adjacent*) cells. The intuition is feasible because all the regions are connected by the road network, which can directly show the sensitivities of regions against weather. Based on the intuition, give a set of regional features F_r, if the inference accuracy is satisfactory using F_r as input, it indicates that such set of regional features are the key factors.

The intuition leads to the model as shown in Fig. 7. In Fig. 7, the parent node g_u specifies the source cell, and the child node g_i is a set of observed cells which are closely located to g_u. This model is not symmetric since $g_i \rightarrow g_u$ may have a different probability comparing with $g_u \rightarrow g_i$. The inference model can be any graphical classifier but we propose to use naïve Bayes classifier [28], because the location closeness can be naturally considered by naïve Bayes classifier. Since different cells have different numbers of neighboring cells, it is hard to use other classifiers such as logistic regression, SVM, neural network, and random forest where the number of input features is fixed. In this situation, Naïve Bayes classifier is a reasonable choice.

The following sections describe the details of the Naïve Bayes classifier, from constructing the marginal distribution to the detailed index-index inference method.

6.1.1 Marginal Distribution

The marginal distribution used in this paper is shown in Fig. 8. A marginal distribution describes the probability distribution of the regions contained in a similarity subset [28]. Specifically in this paper, it describes the probability of one region being the index of i given one of its adjacent regions with index j, if the two regions have a certain similarity. The similarities are split into subsets because the probability distributions may vary upon different similarities. In this paper, we use cosine similarity in terms of regional features

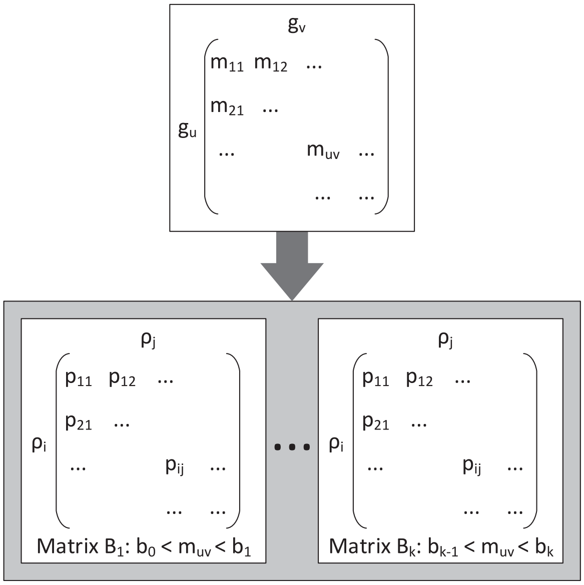


Fig. 8. Converting from similarity matrix to marginal distribution.

as shown in Equation (1) to describe the similarity m_{uv} between two regions g_u and g_v

$$\begin{aligned} m_{uv} &= \frac{F_r^u \cdot F_r^v}{\|F_r^u\| \|F_r^v\|} \\ &= \frac{\sum_{i=1}^{n} F_r^u(i) \times F_r^v(i)}{\sqrt{\sum_{i=1}^{n} (F_r^u(i))^2} \times \sqrt{\sum_{i=1}^{n} (F_r^v(i))^2}}. \end{aligned} \tag{1}$$

According to the similarity of regional features, all pairs of adjacent cells are clustered into k groups. Suppose b_0 is the minimum similarity and b_k is the maximum similarity. The similarity ranges of the k groups are $[b_0, b_1], \ldots, [b_{k-1}, b_k]$ as shown in Fig. 8. The group of $[b_{i-1}, b_i]$ only contains the pairs whose similarities are in between b_{i-1} and b_i. So, the pairs in the same group have the similar similarity. In the group of $[b_{i-1}, b_i]$, the pairs of cells are summarized to marginal distribution matrix B_i. The rows of B_i are the weather-traffic indices of g_u and the columns of B_i are weather-traffic indices of g_v. Specifically, when the weather-traffic index of g_u is ρ_i, the probability that the weather-traffic index of g_v is ρ_j is recorded in p_{ij}. For example, there are 500 pairs of cells in group of $[b_{i-1}, b_i]$. Suppose, in 200 pairs of them, one cell has index 2 and the number of another cells whose index is 1 is 50. Then p_{21} in matrix B_i is 0.4. It indicates that, if two cells have similarity in terms of regional features in between b_{i-1} and b_i, and the weather-traffic index of one cell is 1, the probability that the weather-traffic index of the other cell is 2 is 0.4. Formally,

$$\begin{aligned} p_{ij} &= Pr(\rho(g_u) = \rho_i | \rho(g_v) = \rho_j) \\ &= \frac{|\rho(g_u) = \rho_i, \rho(g_v) = \rho_j|}{|\rho(g_v) = \rho_j|}. \end{aligned} \tag{2}$$

6.1.2 Index-Index Inference

Once the marginal distribution is obtained, the weather-traffic index for a particular cell can be inferred from its adjacent cells using naïve Bayes classifier, which follows Bayes rule

$$\begin{aligned} & Pr(\rho(g_u) = \rho_u | \rho(g_1) = \rho_1, \rho(g_2) = \rho_2, \ldots) \\ &= \frac{Pr(\rho(g_1) = \rho_1, \ldots | \rho(g_u) = \rho_u) * Pr(\rho(g_u) = \rho_u)}{\sum_{i=1}^{k} Pr(\rho(g_1) = \rho_1, \ldots | \rho(g_u) = \rho_i) * Pr(\rho(g_u) = \rho_i)} \\ &= \frac{Pr(\rho(g_1) = \rho_1 | \rho(g_u) = \rho_u) * \cdots * Pr(\rho(g_u) = \rho_u)}{\sum_{i=1}^{k} Pr(\rho(g_1) = \rho_1 | \rho(g_u) = \rho_i) * \cdots * Pr(\rho(g_u) = \rho_i)}. \end{aligned} \tag{3}$$

Given a cell g_u, the marginal distribution allows naïve Bayes classifier to infer which value the weather-traffic index of g_u is most likely to be, based on the weather-traffic indices of its adjacent cells $\rho(g_1)$, $\rho(g_2)$, ... The inference accuracy of 10-fold cross validation is then obtained.

6.2 Weight Estimation of Regional Features

Given a set of regional features, some of them may have trivial impact to weather-traffic index, or are just noise. This requires us to test the weight of each regional feature through a feature selection [29] process.

There are many feature selection methods could be used in this paper. For example, Fisher score [30], where features are scored by considering that features with high quality should assign similar values to instances in the same class and different values to instances from different classes; and ReliefF [31], [32], which selects features to separate instance from different classes. In this paper, we uses the following method similar as mutual information based methods [33], [34], [35].

Suppose a regional feature has nontrivial impact to weather-traffic index. Let us remove this regional feature from the set of regional features. We can use the KFVII method in Section 6.1 to test whether the remaining set of regional features is still the set of key factors which results in high overall accuracy. If not, it is a strong signal that the removed regional feature is very important; otherwise, it is less important. We use $\delta(F_r^i)$ to denote the weight of the regional feature F_r^i. Look closely, the similarity of every two adjacent cells are recomputed in terms of the remained regional features, as well as the marginal distribution. If the inference accuracy is increased more, the removed regional feature has more weight. The idea is presented in Fig. 9.

7 Empirical Study

In this section, we conduct empirical study using the proposed weather-traffic index system in Shanghai. We first introduce our data sources in Section 7.1, and then present the results of regional traffic-weather index obtained in Section 7.2, and finally present the regional features which are the key factors behind the weather-traffic index in Section 7.3.

7.1 Datasets

The input of our weather-traffic index system includes (i) road network data of Shanghai, (ii) taxi trajectory data collected in Shanghai; (iii) weather report data of the same period of time; and (iv) regional information data.

7.1.1 Road Network Data

The road network data of Shanghai is provided by the government, where a *road* (or precisely a *road segment*) is defined as the carriageway between two intersections. An expressway or a large avenue may have two different road

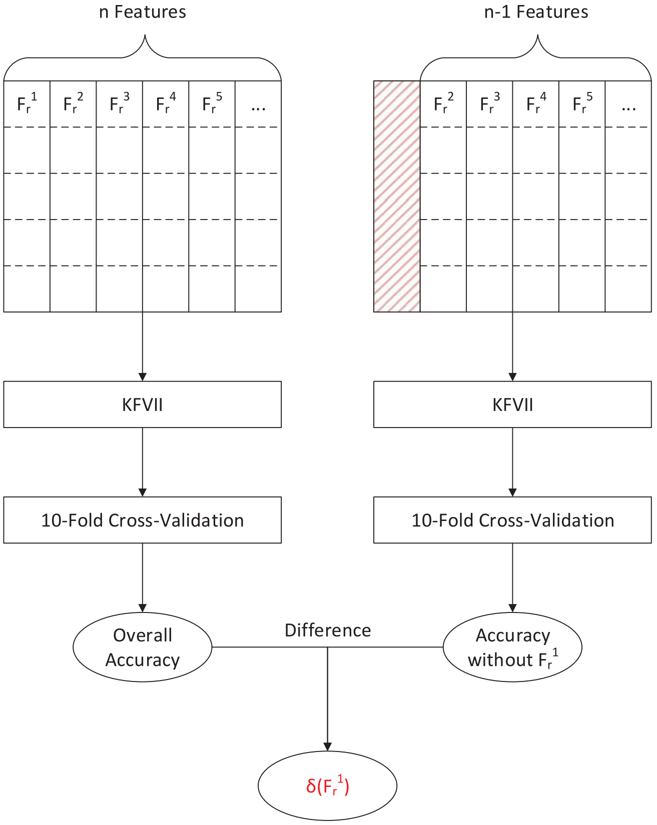


Fig. 9. Weight estimation of regional feature F_r^1.

segments between two intersections, because they are different directions with limited-access. A road network is consisted of a set of roads. There are in total seven levels of roads: national expressway, city expressway, regular highway, large avenue, primary way, secondary way, and regular road [36]. We consider the first four levels of roads as *major roads* and the other three levels as *minor roads*. In this study, only the major roads are used in region partitioning and the minor roads are ignored. By conducting the road-intersection-orientated partitioning method introduced in Section 4.1, the city area of Shanghai is partitioned into 3,207 Voronoi cells.

7.1.2 Weather Report Data

Weather report data are collected from *Weather Underground* (wunderground.com), which is a leading website on commercial weather service providing weather forecast and historical weather information. The weather data contain rich information covered by 14 weather features, including temperature, wind speed, precipitation, etc. In Table 2, we summarize all 14 weather features used in this paper, and they are processed all together. For data alignment, the collected weather reports in Shanghai cover the same period of time as that of taxi trajectory data, i.e., January 2006 to November 2007. The weather is reported on hourly basis. Accordingly, we split day time into time slots by hours.

7.1.3 Trajectory Data

A trajectory is represented as a series of spatial-temporal points [37], where each point is associated with additional information including the driving speed. Our trajectory data of 115.2 GB are collected from 4,529 taxis in Shanghai from January 2006 to November 2007. The average sampling rate of the dataset is about 20 seconds. Table 3 lists the fields recorded in the trajectory data. By extracting the driving speeds of all taxis in each Voronoi cell at each time slot, the average speed is obtained.

TABLE 3
Specifications of Trajectory Data

Attribute	Description
Taxi ID	Taxi registration plate number.
Time	Timestamp of the sample record.
GPS Location	Spatial location of the record in latitude and longitude.
Speed	Instant speed of the taxi (in km per hour).

7.1.4 Regional Information Data

Complicated regional information data have been collected including real estate data and POI data (i.e., points of interests). The real estate data is crawled from *soufun.com*, which is a real estate website providing marketing, e-commerce, listing, and other value-added services for China's real estate and home-related sectors. The real estate data provides a wide range of information including location, price, and age of residential communities. The POI data is collected from *dianping.com*, a website in China providing local life information and third-party consumer service ratings. The data includes detailed merchant information, where the merchants are labeled by categories, such as tourism attractions, hotels, restaurants, leisure facilities, etc.

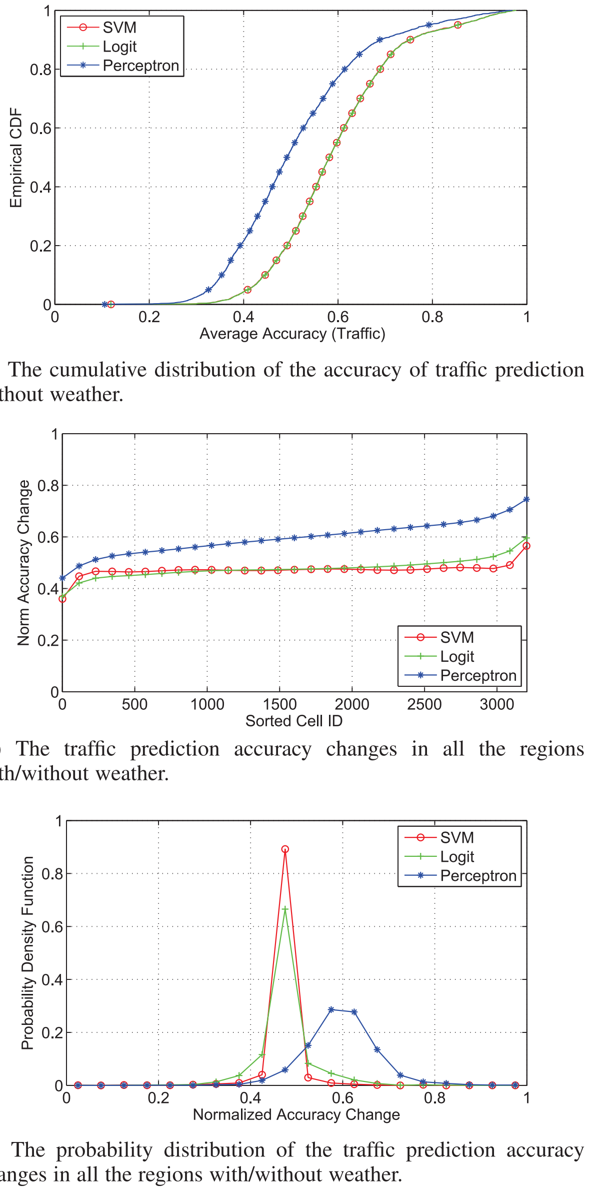
For each Voronoi cell, the regional features can be extracted from real estate data, POI data and road networks. The extracted regional features can be generally clustered into four categories: POI, structure, density, and community. The details are listed in Table 4.

7.2 Weather-Traffic Index

Weather-traffic index in Shanghai is constructed using the weather-traffic index establishment method introduced in Section 5. Briefly, for each cell in each time slot, the average

TABLE 4
Specifications of Regional Features

ID	Feature Detail	ID	Feature Detail
	POI		**Density**
1	# of attractions	13	# of attractions per m^2
2	# of restaurants	14	# of restaurants per m^2
3	# of hotels	15	# of hotels per m^2
4	# of leisures	16	# of leisures per m^2
	Structure	17	Major road length per m^2
5	# of major roads	18	Minor road length per m^2
6	# of minor roads	19	Total road length per m^2
7	# of intersections	20	# of intersections per m^2
8	Ratio of major / minor roads		**Community**
9	Total road length	21	# of residential communities
10	Average road length	22	Average house age
11	Geographical cell size (m^2)	23	Average house unit price
12	# of neighboring cells		



(a) The cumulative distribution of the accuracy of traffic prediction without weather.

(b) The traffic prediction accuracy changes in all the regions with/without weather.

(c) The probability distribution of the traffic prediction accuracy changes in all the regions with/without weather.

Fig. 10. Evaluation of weather-traffic index.

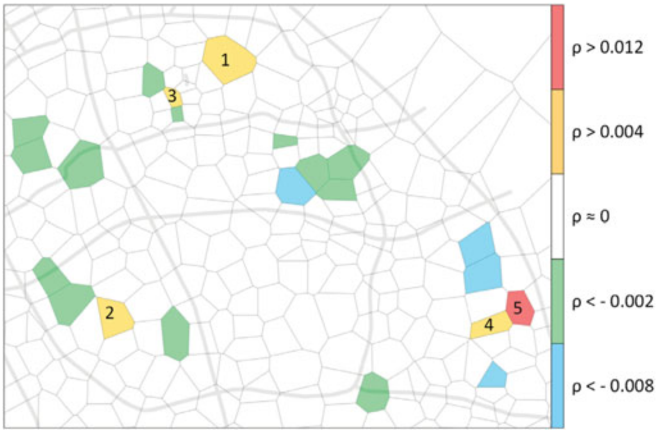


Fig. 11. The weather-traffic index of the regions in the urban areas of Shanghai. The details of the labeled regions are shown in Table 5.

speed is inferred using traffic prediction model with/without weather. The inference accuracy difference indicates the sensitivity of this cell at this time slot. The average of the inference accuracy differences at all time slots indicate the sensitivity of the cell, depending on the fraction of the time that the cell was experiencing abnormal weather. In particular, the traffic prediction model without weather uses the average speeds of previous days of the same cell at the same time slot as the input. The purpose is to predict the current average speed with minimal weather impact since the previous days are in “good” and “bad” weather conditions and as a result the weather impact is trade-off. In the traffic prediction model with weather, the weather features of the current day is used as the additional input features in the traffic prediction.

7.2.1 Robustness

First, we show the accuracy of the traffic prediction without weather in Fig. 10a. The relative high average traffic prediction is necessary. If the traffic prediction is poor, it typically means the input features contain much noise. As a result, we have less confidence to the impact of weather detected. Fig. 10a shows that the traffic prediction accuracy in most cells is 0.5 using support vector machine. This result is promising since it is comparable to the current state-of-the-art in traffic prediction [20], [21]. The other two traffic prediction methods logistic regression and perceptron show low traffic prediction accuracy, but they will be used to test whether the traffic prediction accuracy change with/without weather is independent of traffic prediction models, in other words, to test the robustness of our weather-traffic index establishment method.

In Fig. 10b, the traffic prediction accuracy changes with/without weather are presented using support vector machine, logistic regression and perceptron. For better presentation, we sort the cells by the predication accuracy changes. The probability distribution of the traffic prediction accuracy changes with/without weather is illustrated in Fig. 10c. We observe that the accuracy changes are generally normally distributed with small variance, i.e., the accuracy changes of most cells are close to the mean. In particular, the bias of the distribution using perceptron is because the accuracy of perceptron has greater variance compared to that of support vector machine and logistic regression. Nevertheless, the similar distributions of the three models clearly show the robustness of our method.

7.2.2 Validation

The effectiveness of weather-traffic index established have been verified against the observations in the real world. Fig. 11 shows the resulting weather-traffic index of the regions in the urban areas of Shanghai. In Fig. 11, a positive weather-traffic index describes a region with high impact of weather change on transport, and a negative weather-traffic index describes a region with low impact of weather change on transport. There are five regions with very high weather-traffic index as labeled 1-5 in the figure, and the details of these regions are shown in Table 5. In practical, there are often many people walking near a tourism attraction, and when there is a rain, the tourists may have rush home. Thus, it may cause transport problems and reduce the traffic efficiency. Regions 1-3 in Fig. 11 illustrate such situation.

TABLE 5
Details of Regions with High Weather-Traffic Index

ID	Description
1	Yu Garden, a tourism attraction.
2	Shanghai Confucian Temple, an old temple with many restaurants around.
3	Shanghai Town God's Temple, a tourism attraction.
4	An area with many old buildings and construction sites.
5	Construction sites (Bund House, a high-end residential community is built several years later).

Notably, not all the tourism attractions have a high impact of weather on transport, some even more popular ones, such as Xintiandi (historical location of the first Congress of the Chinese Communist Party), is not highly influenced by weather on transport. We have noticed that it may because the district of Xintiandi is constructed later than regions 1-3, and the public transport is more efficient. Besides tourism attractions, there are many other reasons may cause the region vulnerable to extreme weather conditions, such as contraction sites (regions 4 and 5) and old districts (regions 2 and 4). They both prove that the regions distinguished by our weather-traffic index make sense.

Another validation is shown in Fig. 12, with the four labeled areas in Fig. 4 presented as examples. The third column is about the average speeds in the cloudy day and the fourth column is about the average speeds in the rainy day. The weather-traffic indices of the labeled areas are shown in the first column and the second column shows the rain-traffic indices of the corresponding areas. In the weather-traffic index, all available 14 features in the weather report data are applied. Since rain has own impact to traffic, the hypothesis is that the composite impact of all features provides a general description of the weather impact to traffic, and the rain-traffic index should be better to present the impact of rain to traffic in cells. Interestingly, the observations in the four labeled areas give strong support to this hypothesis.

Look closely, the circled cell in the area labeled 1 show low average speed in rainy day and faster average speed in cloudy day. So, the weather-traffic index and rain-traffic index shows the impact of weather/rain to traffic is relatively high. In the circled cell in the area labeled 3 and 4, the average speed is significantly slowed by rain compared to that in cloudy day. So, the weather-traffic index and rain-traffic index shows the impact of the weather/rain to traffic is significantly high. If we observe other cells in each labeled area, we found that the rain-traffic index generally shows more accurate description of the impact of rain to traffic than the weather-traffic index. In the area labeled 2, it is interesting to observe that the average speed in the circled cell is high in rainy day and is low in cloudy day. After deep investigation, we found in such cell the roads are usually crowded with pedestrians, such as the regions around Shanghai Town God's Temple, a hot tourist spot. The average speed of taxis are slow in normal days. In rainy days, the number of pedestrians are reduced such that the speeds of taxis tend to increase. However, such region is much less than the regions where the average speed slows down in rainy days compared to cloudy days. In the factor analysis, we are only interested in the regional features of the cells where the average speed slows down in rainy days.

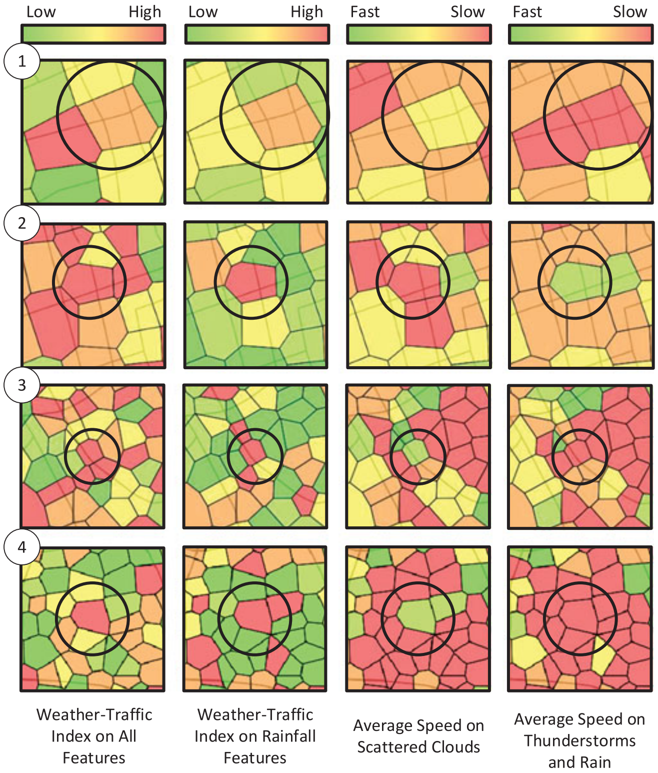

Low
High
Low
High
Fast
Slow
Fast
Slow
1
2
3
4
Weather-Traffic Index on All Features
Weather-Traffic Index on Rainfall Features
Average Speed on Scattered Clouds
Average Speed on Thunderstorms and Rain

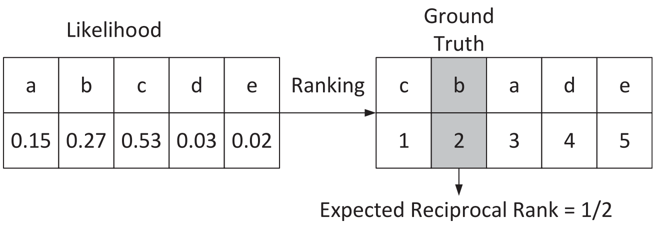
Fig. 12. The weather-traffic index validation using the four labeled areas in Fig. 4.

7.3 Effectiveness of KFVII

In the weather-traffic index system, the factor analysis includes two functions. The first function is to verify a set of regional features are key factors behind the vulnerability of traffic in cells to inclement weather, and the second function is to estimate the weight of each regional feature. Both of them are based on the inference of weather-traffic index of each cell by the weather-traffic indices of other cells. The inference method must be accurate. If the inference method is inaccurate, the noise may dominate the impact of regional features. We propose to use naïve Bayes classifier because the location closeness of cells can be naturally considered by it. So, the empirical study evaluates the inference method in KFVII first, by comparing two straightforward methods, through a fine-grained evaluation metric.

7.3.1 Evaluation Metric

The inference accuracy is evaluated through *expected reciprocal rank* (ERR) [38]. Given a cell, the expected reciprocal rank evaluates each inference result which indicates the likelihood for this cell to take each index value. An example is shown in Fig. 13. Let us define five index values a, b, c, d and e, where the likelihoods returned by the inference method are 0.15, 0.27, 0.53, 0.03, 0.02, respectively. Thus, the sorting orders of the likelihoods are 3, 2, 1, 4, 5, respectively. If the true index value of this cell is, for example, b, the expected reciprocal rank is the reciprocal of 2 (i.e., $\frac{1}{2}$), the position of b in the sorted list. Similarly, if the true index value of this cell is d, the expected reciprocal rank is the



a	b	c	d	e
0.15	0.27	0.53	0.03	0.02

c	b	a	d	e
1	2	3	4	5

Fig. 13. An example of expected reciprocal rank.

reciprocal of 4 (i.e., $\frac{1}{4}$). For the expected reciprocal rank of the inference method, the average value of the reciprocals of all cells are used.

Expected reciprocal rank can be considered as a fairer metric comparing with maximum-likelihood accuracy. For example, if the likelihoods of the categories of weather-traffic index of a region are $\langle a: 0.49, b: 0.48, c: 0.03\rangle$, it is actually difficult to determine whether the category is a or b, but it is clear that the type is not c. However, using the metric of maximum likelihood cannot give the bonus of such observation, since no matter the ground truth is a or b, the accuracy for this region may result in 0 in a 50 percent chance. If we evaluate the likelihoods using expected reciprocal rank, it will give us a comprehensive distribution of the accuracy (either 1 or $\frac{1}{2}$). Hence, the evaluation of expected reciprocal rank is fairer, which widens the gap between different likelihoods.

7.3.2 Straightforward Methods

There are two straightforward methods implemented in our empirical study. One is random guess and the other is artificial neural network (ANN).

In random guess, we assume the probability of guessing any weather-traffic index value of a cell is $1/l$. Based on the expected reciprocal rank metric, the expectation of the random guess is

$$\frac{1*\frac{1}{l}+\frac{1}{2}*\frac{1}{l}+\frac{1}{3}*\frac{1}{l}+\cdots+\frac{1}{l}*\frac{1}{l}}{l*\frac{1}{l}}=\frac{1}{l}*\sum_{i=1}^{l}\frac{1}{i}. \tag{4}$$

In this paper, since we have five categories of the weather-traffic index, the expectation is around 0.4567.

We use an artificial neural network with one hidden layer for the inference on weather-traffic index directly from regional features as another baseline method. Artificial neural networks have been used to solve a wide variety of tasks that are difficult to solve using ordinary classification models, including urban comping problems [8].

In ANN, we observed that the cells have different number of adjacent cells. Thus, it is difficult to train an ANN where the input layer is related to adjacent cells. Instead, the input of the ANN are the observations of all Voronoi cells. Because the size of hidden layer affects the results, we conduct configurations with different sizes of the hidden layer. In the output layer, we use softmax regression [39] as the classification model. The softmax regression model generalizes logistic regression to classification problems where the class label can take on more than two possible values. For the activation function, we use sigmoid function [40], which refers to a special case of the logistic function.

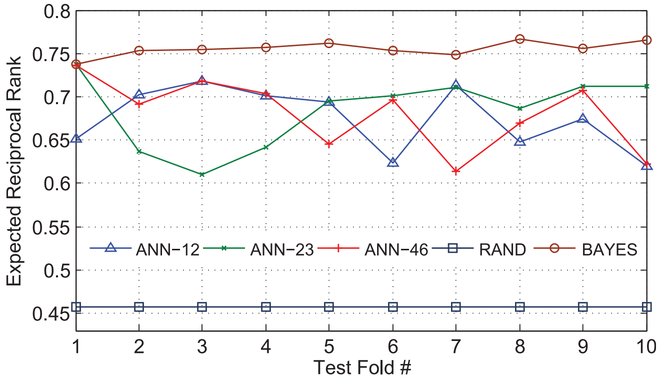


Fig. 14. Comparison of the expected reciprocal ranks of the inference of the categories of traffic-weather index.

7.3.3 Comparison of Methods

We test the inference accuracy of the naïve Bayes classifier, random guess and ANN, and the expected reciprocal ranks of the results are presented in Fig. 14. In particular,

- BAYES: the naïve Bayes classifier where the initial probabilities of $Pr(\rho(g)=\rho_k)$ is the statistical distribution of the categories. That is, if there are m regions with category k among a total of n regions, we set $Pr(\rho(g)=\rho_k)=m/n$. The pairs of adjacent cells are clustered into five equal groups for the marginal distribution based on their similarities: $[0,0.2]$, $(0.2,0.4]$, $(0.4,0.6]$, $(0.6,0.8]$, and $(0.8,1]$.
- RAND: the random guess method.
- ANN-12/23/46: the artificial neural network classifier with 12/23/46 neurons in the hidden layer.

In Fig. 14, it is clear that BAYES has the best performance compared to ANN in all settings and RAND. Moreover, the expected reciprocal rank of BAYES reaches around 0.75. On the other hand, ANNs with different settings lead to an average expected reciprocal rank of roughly 0.65. The random guess method performs the worst, with a consistent expected reciprocal rank of 0.4567.

Next, we first verify the key factors via KFVII using different sets of regional features, and find out some are key factors and some are not. Then, we assume all regional features are key factors, and estimate the weight of each regional feature.

7.4 Factor Analysis

Given any set of regional features, we verify they are key factors to weather-traffic index or not by KFVII introduced in Section 6.1, and then generalize the weight of the set of regional features by the method introduced in Section 6.2.

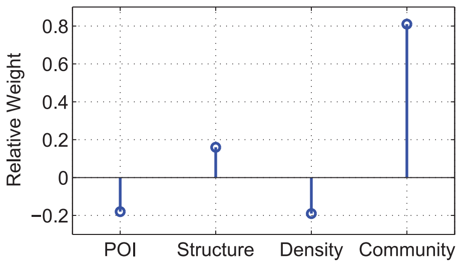


Fig. 15. The weights of the regional feature sets in Table 4.

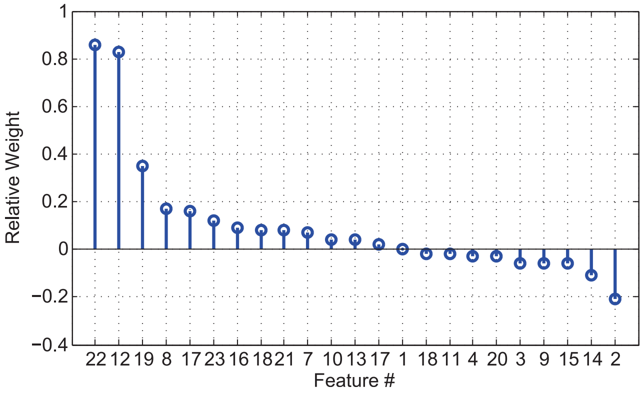


Fig. 16. The weights of the regional features in Table 4. From left to right: 1) house age, 2) # of neighbors, 3) road length / area size, 4) ratio of major / minor roads, 5) density of major roads, 6) house price, 7) density of leisures, 8) density of minor roads, 9) # of communities, 10) # of intersections, 11) average road length, 12) density of attractions, 13) density of major roads, 14) # of attractions, 15) # of minor roads, 16) area size, 17) # of leisures, 18) density of intersections, 19) # of hotels, 20) road length, 21) density of hotels, 22) density of restaurants, 23) # of restaurants.

We test four sets of regional features, each corresponding to one of the four regional feature categories, i.e., POI, structure, density, and community listed in Table 4. Fig. 15 shows the weight for each set. It is clear that the set of regional features in community is the relatively most important key factor set, and the regional features in structure are relatively less important key factors. The regional features in POI and density are not key factors. This conclusion is verified by the observations in Fig. 16.

The weight estimation of each feature is shown in Fig. 16. It demonstrates some surprising phenomena, for example, *house age* and *the number of neighboring cells* have the highest impact to the weather-traffic index. After reviewing all factors and checking up the information of the cells in the real world, we conclude some explanations to the unexpected outcomes. The older house age usually reflects that the cell is typically quite mature with old traffic facilities, more business outlets, narrow roads and more populations. As a results, when the weather changes, e.g., heavy rain, those cells may cause serious traffic problems. Moreover, we observe that the second most weighted regional feature is the number of neighboring cells. If a cell has many neighboring cells, it indicates the region has a more complicated road structure, i.e., more intersections and in turn it typically is a mature region with high density of population.

After the weights of each regional features being estimated, we found that the regional features with the highest weights are from the community category as listed in Table 4. This is consistent with the results in Fig. 15 which indicates the regional features in community category together are the key factors. Moreover, the regional features in structure have smaller weights, and most regional features in POI and density have the least weights.

The effectiveness of estimated weights of regional features have been verified against the observations in the real world. In Fig. 17, the four labeled areas in Fig. 4 are presented as examples. The first column is the weather-traffic index, the second column is the average house age, the third column is the number of neighboring cells, and the fourth column is the geographical cell size. The hypothesis is that, if a regional feature has more weight, it should be more correlated to the weather-traffic index; otherwise, it is not. Among the three regional features, the weights of average house age and the number of neighboring cells are similar and they are significantly higher than that of geographical cell size. The observations in the four labeled areas shows consistent with this hypothesis. In the second column, some cells are with white color to indicate the relevant information is missing in the cell. The high weight of average house age means that for the cells which have the house age information they have high correlation with the weather-traffic index. The fourth column indicates the lowest correlation of geographical cell size.

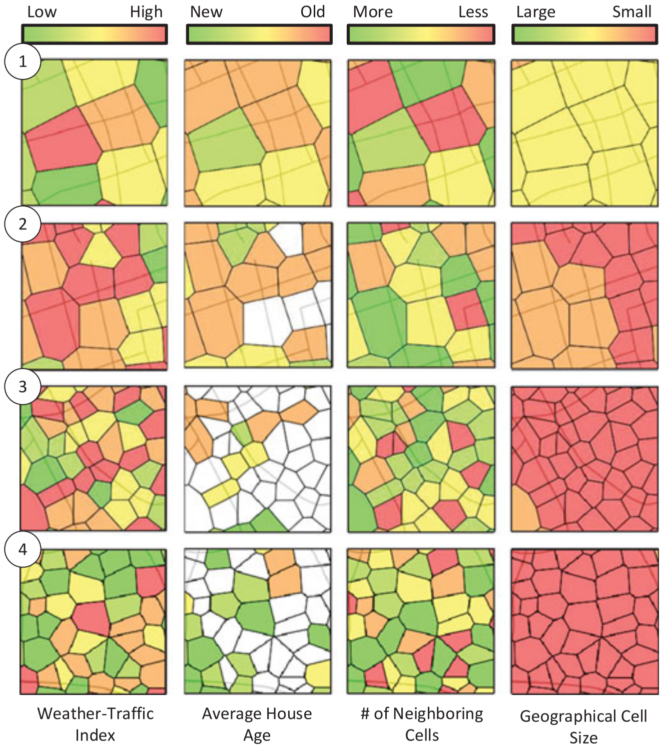


Fig. 17. Validation of regional feature weight estimations using the four labeled areas in Fig. 4.

8 Conclusion

This work fills the gap in the study on the impact of weather to traffic from few locations to all road networks throughout a city, more importantly, the regional features leading to the vulnerability of traffic in local areas to inclement weather are systematically revealed for the first time. The empirical study in Shanghai demonstrates the effectiveness of the proposed system. The regional weather-traffic indices extracted have been validated to be surprisingly consistent with real world observations. Further regional key factor analysis yields interesting results. For example, the regional house age has significant impact on the region's weather-traffic index. The achievement in this work will benefit government agent to understand the functional character of districts throughout a city, to improve traffic prediction and to learn the key factors in urban planning, etc. The knowledge of key factors learned from one city is transferable to another city because modern cities often have road networks with similar quantitative density and other features. At last, the investigated problem has important practical value, but the research is still in its early stage. We are working on to incorporating more data sources to continuously improve the results.

Acknowledgments

This paper was supported in part by the National Key Basic Research and Development Program of China (973) Grant 2014CB340303 and NSFC grant 61300030. An earlier version of this work appeared in the Proceedings of IEEE ICDM [1], Nov 2015.

References

[1] Y. Ding, Y. Li, K. Deng, H. Tan, M. Yuan, and L. M. Ni, "Dissecting regional weather-traffic sensitivity throughout a city," in *Proc. 15th IEEE Int. Conf. Data Mining*, 2015, pp. 739–744.

[2] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, 2014, Art. no. 38.

[3] A. Saegusa and Y. Fujiwara, "A study on forecasting road surface conditions based on weather and road surface data," *IEICE Trans. Inf. Syst.*, vol. 90-D, no. 2, pp. 509–516, 2007.

[4] M. A. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time traffic crash-prediction model using archived weather and its traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 2, pp. 167–174, Jun. 2006.

[5] S. Dunne and B. Ghosh, "Weather adaptive traffic prediction using neurowavelet models," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 1, pp. 370–379, Mar. 2013.

[6] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," *Transp. Res. Part D: Transport Environment*, vol. 14, no. 3, pp. 205–221, 2009.

[7] Talking about smart transportation from the rainstorm of Beijing, 2012. [Online]. Available: http://info.secu.hc360.com/2012/07/300822649026.shtml

[8] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1436–1444.

[9] S. Liu, Y. Liu, L. M. Ni, J. Fan, and M. Li, "Towards mobility-based clustering," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 919–928.

[10] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 186–194.

[11] F. Zhang, D. Wilkie, Y. Zheng, and X. Xie, "Sensing the pulse of urban refueling behavior," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 13–22.

[12] K. Zheng, Y. Zheng, N. J. Yuan, and S. Shang, "On discovery of gathering patterns from trajectories," in *Proc. IEEE Int. Conf. Data Eng.*, 2013, pp. 242–253.

[13] L. A. Tang, et al., "On discovery of traveling companions from streaming trajectories," in *Proc. 28th Int. Conf. Data Eng.*, 2012, pp. 186–197.

[14] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Mar. 2015.

[15] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Comput. Surveys*, vol. 23, no. 3, pp. 345–405, 1991.

[16] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, pp. 424–438, 1969.

[17] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. San Francisco, CA, USA: Holden-Day, Incorporated, 1990.

[18] B. Williams, P. Durvasula, and D. Brown, "Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1644, pp. 132–141, 1998.

[19] H. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "State space neural networks for freeway travel time prediction," in *Proc. Int. Conf. Artif. Neural Netw.*, 2002, pp. 1043–1048.

[20] B. Pan, U. Demiryurek, and C. Shahabi, "Utilizing real-world transportation data for accurate traffic prediction," in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 595–604.

[21] J. D. Gehrke and J. Wojtusiak, "Traffic prediction for agent route planning," in *Proc. 8th Int. Conf. Comput. Sci.*, 2008, pp. 692–701.

[22] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 316–324.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[24] H.-F. Yu, F.-L. Huang, and C.-J. Lin, "Dual coordinate descent methods for logistic regression and maximum entropy models," *Mach. Learn.*, vol. 85, no. 1/2, pp. 41–75, 2011.

[25] F. Rosenblatt, "The perceptron—a perceiving and recognizing automaton," Cornell Aeronautical Laboratory, Buffalo, NY, USA, Tech. Rep. 85–460-1, 1957.

[26] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Proc. 7th Int. Conf. Artif. Neural Netw.*, 1997, pp. 999–1004.

[27] C.-H. Wu, J.-M. Ho, and D.-T. Lee, "Travel-time prediction with support vector regression," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 276–281, Dec. 2004.

[28] D. Lowd and P. M. Domingos, "Naive Bayes models for probability estimation," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 529–536.

[29] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2014, pp. 37–64.

[30] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.

[31] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of reliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1/2, pp. 23–69, 2003.

[32] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Workshop Mach. Learn.*, 1992, pp. 249–256.

[33] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, pp. 284–292.

[34] H. Peng, F. Long, and C. H. Q. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[35] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 856–863.

[36] Y. Ding, J. Zheng, H. Tan, W. Luo, and L. M. Ni, "Inferring road type in crowdsourced map services," in *Proc. 19th Int. Conf. Database Syst. Adv. Appl.*, 2014, pp. 392–406.

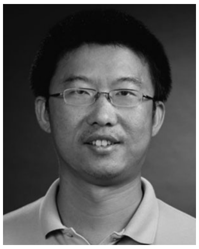
[37] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, 2015, Art. no. 29.

[38] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 621–630.

[39] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

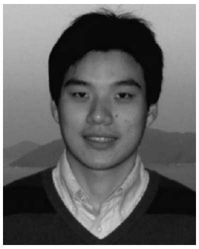
[40] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *Proc. Int. Workshop Artif. Neural Netw.*, 1995, pp. 195–201.

Ye Ding received the PhD degree in computer science and engineering supervised by Prof. Lionel M. Ni from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, in 2014. He is currently a research associate with Fok Ying Tung Graduate School, Hong Kong University of Science and Technology. His research interests include spatial-temporal data analytics and big data. He is a member of the IEEE since 2013.



Yanhua Li (S'09-M'13-SM'16) received the two PhD degrees in electrical engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and in computer science from the University of Minnesota, Twin Cities, in 2013, respectively. He has worked as a researcher with HUAWEI Noahs Ark LAB, Hong Kong, from Aug 2013 to Dec 2014, and has interned with Bell Labs, New Jersey, Microsoft Research Asia, and HUAWEI Research Labs of America from 2011 to 2013. He is currently an assistant professor in the Department of Computer Science, Worcester Polytechnic Institute, Worcester, Massachusetts. His research interests include big data analytics and urban computing in many contexts, including urban network data analytics and management, urban planning, and optimization. He served on TPC of INFOCOM 2015-2016, ICDCS 2014-2016, SIGSPATIAL GIS 2015, and he is the co-chair of SIMPLEX 2015. He is a senior member of the IEEE.

Ke Deng received the PhD degree in computer science from the University of Queensland (UQ), in 2007 with focus on data and knowledge engineering. He had been an acting lecturer and course convener in the School of Information Technology and Electrical Engineering, UQ during 2008-2012. He had been a postdoctoral research fellow with CSIRO ICT Center, in 2007. He had been an ARC Australian postdoctoral fellow working during 2010-2012. He had been a researcher in Huawei Noah Arks Research Lab, Hong Kong, during 2013-2015. Currently, he is a lecturer in the School of Computer Science and Information Technology, RMIT University. He is a member of the IEEE.



Haoyu Tan received the PhD degree in computer science and engineering from Hong Kong University of Science and Technology, in 2013. He has worked as a research assistant professor with HKUST Fok Ying Tung Research Institution from May 2013 to present. His research interests include big data management and processing, machine learning, large-scale data mining, human behaviour analysis, urban computing, and recommender systems. He is a member of the IEEE.

Mingxuan Yuan received the PhD degree from Hong Kong University of Science and Technology. He is a researcher with Noah's Ark Lab, Huawei. His main research interests include spatiotemporal data management/mining, telco (telecommunication) big data management/mining, telco big data privacy, and visualization. He is a member of the IEEE.

Lionel M. Ni is a chair professor in the Department of Computer and Information Science and Vice Rector for Academic Affairs with the University of Macau. He has chaired more than 30 professional conferences and has received eight awards for authoring outstanding papers. He serves on the editorial boards of the *Communications of the ACM*, the *IEEE Transactions on Big Data,* and the *ACM Transactions on Sensor Networks*. He is a fellow of the IEEE and Hong Kong Academy of Engineering Science.