2015 IEEE International Conference on Data Mining

# Dissecting Regional Weather-Traffic Sensitivity throughout a City

Ye Ding [1], Yanhua Li [2], Ke Deng [3], Haoyu Tan [1], Mingxuan Yuan [4], Lionel M. Ni [5]
[1] *Guangzhou HKUST Fok Ying Tung Research Institute, The Hong Kong University of Science and Technology*
[2] *Computer Science Department, Worcester Polytechnic Institute (WPI), Worcester, MA*
[3] *School of Computer Science and Information Technology, RMIT University, Australia*
[4] *Noah's Ark Lab, Huawei Technologies Co., Ltd.* [5] *University of Macau*
yeding@ust.hk, yli15@wpi.edu, ke.deng@rmit.edu.au
haoyutan@ust.hk, yuan.mingxuan@huawei.com, ni@umac.mo

***Abstract*—The impact of inclement weather to urban traffic has been widely observed and studied for many years, with focus primarily on individual road segments by analyzing data from roadside deployed monitors. However, two fundamental questions are still open: (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. Answering these questions is unprecedentedly important for urban planners to understand the functional characteristics of various urban regions throughout a city, and to improve traffic prediction and learn the key factors in urban planning. 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. In this work, we make the first study on these questions, by developing a weather-traffic index (WTI) system. The system includes two main components: WTI establishment and key factor analysis. Using the proposed system, we conducted comprehensive empirical study in Shanghai, and the WTI 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 WTI, which sheds light on future urban planning and reconstruction.**

***Keywords*—*Trajectories, weather-traffic index, city dynamics, urban computing.***

## I. Introduction

Urban computing [1] aims 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 [2][3]. 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.

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 *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 WTI, 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 POIs (points of interest), the traffic volume, the average age of the household, the density of buildings and more in the surrounding regions. The WTI 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 prediction and to learn the key factors in urban planning, etc.

To enable WTI 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 [4] 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

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.25

WTI throughout a city, which indicates the impact of weather to traffic from light to heavy. The second is to reveal the key factors behind WTI 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.

We have addressed a series of techniques challenges in this work, and the central contributions are summarized as follows: 1) A systematic approach as shown in Section II has been proposed for establishing WTI 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. 2) A supervised learning method as shown in Section II-C has been proposed to disclose the key factors and their weights contributing to WTI throughout the city. It is a challenging task because many factors have composite and delicate influence concurrently. 3) Using the proposed system, we conduct empirical study as shown in Section III in Shanghai, the largest city in China, using 115.2 GB traffic data (extracted from more than 4000 taxi trajectories) for two years, the weather data of the same period of time, the road networks and complicated regional features. The established WTI and the discovered key factors have been extensively verified against the observations in the real world.

**Related Work.** The early research on the relation between weather and traffic is mainly based on quantitative analysis and statistical methods. For example, in [2], 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 [3], they developed a neurowavelet 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.

## II. Framework

This work aims to develop a *weather-traffic index system* which performs two tasks: establishment of WTI throughout a city and analysis of key factors behind the index. The framework of the proposed system is shown in Figure 1, which consists of three functional components: 1) data preparation component; 2) WTI establishment component; and 3) factor analysis component.

### A. Data Preparation

In this section, we introduce the data preparation component which partitions the city into fairly distributed regions,

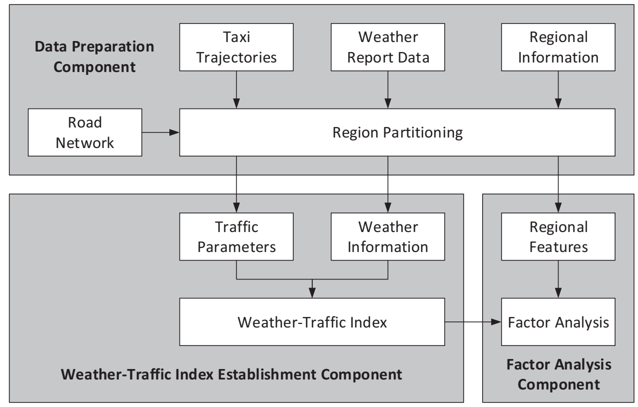


Fig. 1: Framework of the weather-traffic index (WTI) system.

and collects relevant source data for each region.

**Region Partitioning and Data Collections.** We partition the city region into cells using Voronoi diagram [5], where seeds are the road intersections, and the shapes and sizes of the cells differ from each other. We call it *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. Such 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 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 WTI.

For regions partitioned, four types of data sources are collected, including the road network which consists of a set of road segments and a set of road intersections, the traffic data, the regional features in the city of interest, and the weather data in the same period of time. (See Section III-A for more details.)

**Evaluating Traffic Conditions.** We adopt the average speed on a road segment during a time interval as the quantitative measure for the traffic, which indicates the traffic condition. 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, 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 several classes. 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

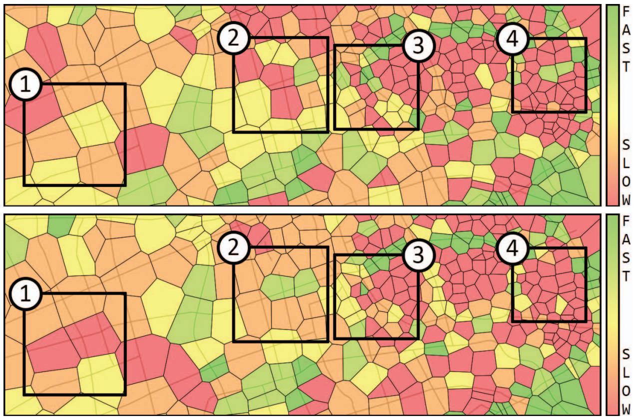


Fig. 2: The average speed at 14:00 on two days in summer in Shanghai, where the weather is scatter cloudy (top) and rainy (bottom).

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. Note that our framework is applicable to more general traffic condition measures, such as the number of passing unique vehicles, etc.

### *B. Weather-Traffic Index Establishment*

The weather data and traffic data from data preparation component are the input of WTI establishment component. The intuition that traffic is influenced by weather can be proven by the example shown in Figure 2. 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, WTI is necessary to indicate the impact of weather to traffic in different cells. In this paper, the WTI value of a cell $g$ is denoted as $\rho(g)$ and takes value from a discrete range.

**Correlation Detection.** Now, we are in a position to show how we evaluate the correlation between the traffic speed, denoted as $F_t$, and weather, denoted as $F_w$. The idea behind our method 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 Figure 3. The traffic prediction models are trained separately in different time slots. The reason is that the traffic regularity in time slot, for example, 7:00 am - 9:00 am can be very different from another time slot, for example 9:00 am - 11:00 am. As shown in Figure 3, the average of the traffic prediction accuracies in different time slots is used.

The WTI 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

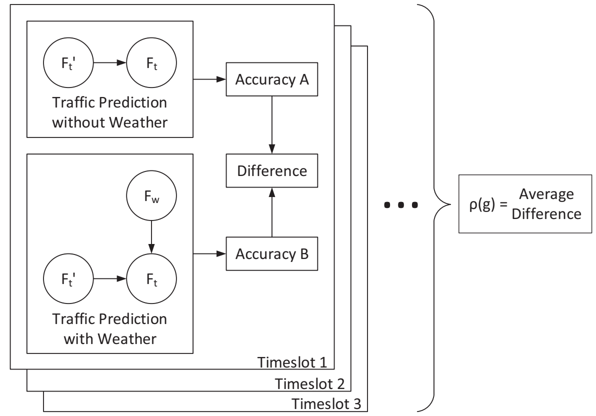


Fig. 3: Weather-traffic correlation detection.

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.

**Traffic Prediction.** Traffic prediction is a well studied problem in recent years [6][7]. In this work, the traffic parameter of interest is consisted of discrete classes, thus we treat traffic prediction as a classification problem. We initially choose support vector machine (SVM) as the inference method because it is suitable for time series prediction [8], and it is a common method in the inference of traffic in transport research [9]. To be fair and robust, we also choose logistic regression and perceptron, both of them are popular linear models like SVM, to verify the output of SVM. The average accuracy of a 10-fold cross-validation is used to compute the accuracy difference as shown in Figure 3. We conclude WTI 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 WTI value is assigned to the cell to indicate the uncertainty of the correlation between weather and traffic.

### *C. Factor Analysis*

In this section, our discussion is based on the assumption that WTI of all cells have been certainly assigned. WTI indicates which cells are correlated with weather in terms of traffic, and it provides the possibility for us to investigate the key factors behind the correlation. The factors are *regional features*, denoted as $F_r$. The factor analysis identifies the key factors and their weights contributing to the WTI of cells. In other words, it discloses what regional features make the traffic in some cells vulnerable to inclement weather.

**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 WTI 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. The

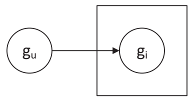


Fig. 4: Weather-traffic index inference from adjacent cells.

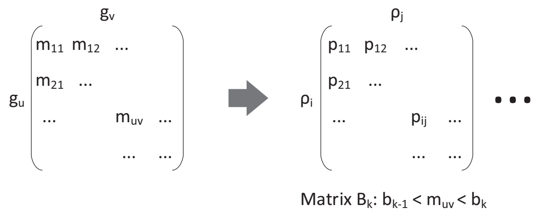


Fig. 5: From similarity matrix to marginal distribution.

intuition leads to the model as shown in Figure 4. In Figure 4, 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, 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.

We first extract the marginal distribution of the model. A marginal distribution describes the probability distribution of the regions contained in a similarity subset. 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 to describe the similarity $m_{uv}$ between two regions $g_u$ and $g_v$. The marginal distribution used in this paper is shown in Figure 5. 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 $[b_0, b_1]$, $\cdots$, $[b_{k_1}, b_k]$ as shown in Figure 5. 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 WTI of $g_u$ and the columns of $B_i$ are the WTI of $g_v$. Specifically, when the WTI of $g_u$ is $\rho_i$, the probability that the WTI of $g_v$ is $\rho_j$ 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 WTI of one cell is 1, the probability that the WTI of the other cell is 2 is 0.4. Formally,

$$
\begin{aligned}
p_{ij} &= Pr(\rho(g_u) = \rho_i | \rho(g_v) = \rho_j) \\
&= \frac{|\rho(g_u) = \rho_i, \rho(g_v) = \rho_j|}{|\rho(g_v) = \rho_j|}
\end{aligned} \tag{1}
$$

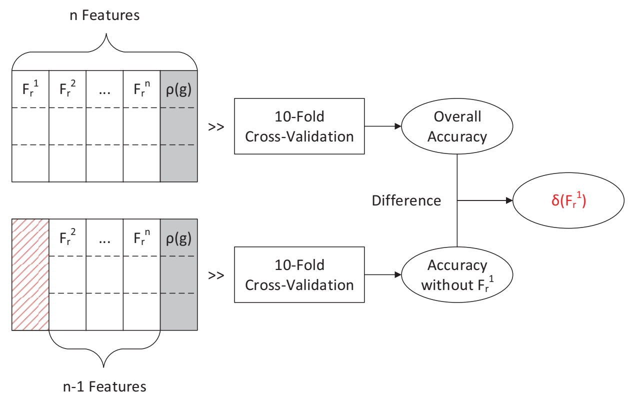


Fig. 6: Weight estimation of regional feature $F_r^1$.

Using the marginal distribution obtained, now we show how to infer the weather-traffic index for a particular cell from its adjacent cells using naïve Bayes classifier, which follows Bayes rule:

$$
\begin{aligned}
&Pr(\rho(g_u) = \rho_u | \rho(g_1) = \rho_1, \rho(g_2) = \rho_2, \cdots) \\
=&\frac{Pr(\rho(g_1) = \rho_1, \cdots | \rho(g_u) = \rho_u) * Pr(\rho(g_u) = \rho_u)}{\sum_{i=1}^{k} Pr(\rho(g_1) = \rho_1, \cdots | \rho(g_u) = \rho_i) * Pr(\rho(g_u) = \rho_i)} \\
=&\frac{Pr(\rho(g_1) = \rho_1 | \rho(g_u) = \rho_u) * \cdots * Pr(\rho(g_u) = \rho_u)}{\sum_{i=1}^{k} Pr(\rho(g_1) = \rho_1 | \rho(g_u) = \rho_i) * \cdots * Pr(\rho(g_u) = \rho_i)}
\end{aligned} \tag{2}
$$

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

**Weight Estimation of Regional Features.** Given a set of regional features, some of them may have trivial impact to WTI, or are just noise. This requires us to test the weight of each regional feature. Suppose a regional feature has nontrivial impact to WTI. Let us remove this regional feature from the set of regional features. We can use the KFVII method 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 Figure 6.

## III. Empirical Study

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

### A. Datasets

The input of our WTI system includes: 1) road network data of Shanghai provided by the government. In this study,

TABLE I: 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 | | |

only the major roads (expressway, large avenue, etc.) are used in region partitioning and the minor roads are ignored. By conducting the road-intersection-orientated partitioning method introduced in Section II-A, the city area of Shanghai is partitioned into $3,207$ Voronoi cells. 2) taxi trajectory data of 115.2 GB collected from 4,529 taxis in Shanghai from January 2006 to November 2007. The average sampling rate of the dataset is about 20 seconds. The fields recorded in the trajectory data include taxi ID, time, GPS location, and driving speed. 3) weather report data of the same period of time collected from *wunderground.com* with $14$ weather features, including time, temperature, humidity, etc. The weather is reported on hourly basis. 4) regional information data, including real estate data from *soufun.com* and POI data from *dianping.com*. For each Voronoi cell, regional features can be extracted from them and clustered into four categories: POI, structure, density, and community. The details are listed in Table I.

### *B. Weather-Traffic Index*

Weather-traffic index in Shanghai is constructed using the WTI establishment method introduced in Section II-B. The effectiveness of WTI established have been verified against the observations in the real world. In Figure 7, the four labeled areas in Figure 2 are 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 WTI 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 WTI, 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

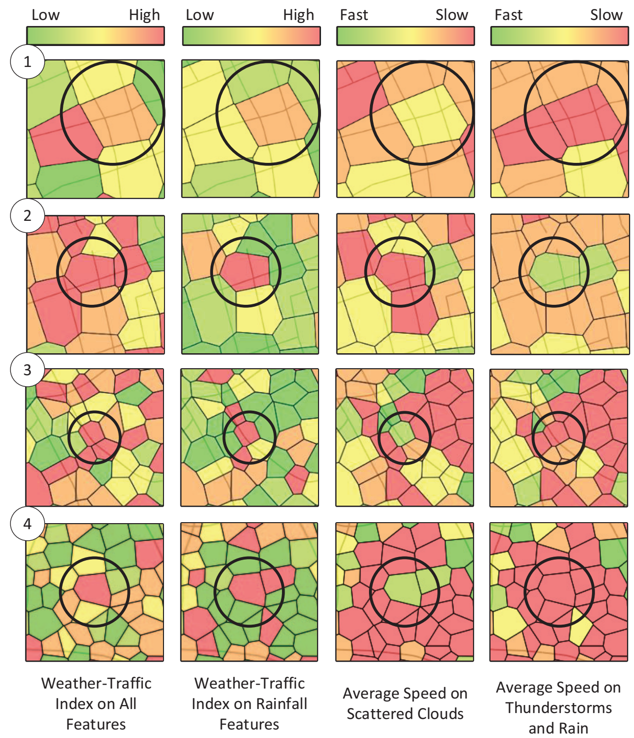


Fig. 7: WTI validation using the four labeled areas in Fig 2.

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.

### *C. Effectiveness of KFVII*

In this section, we evaluate the inference method in KFVII by comparing it with two baseline methods, through a fine-grained evaluation metric.

**Evaluation Metric.** The inference accuracy is evaluated through *expected reciprocal rank* [10]. Given a cell, the expected reciprocal rank evaluates each inference result which indicates the likelihood for this cell to take each index value. For example, 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 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.

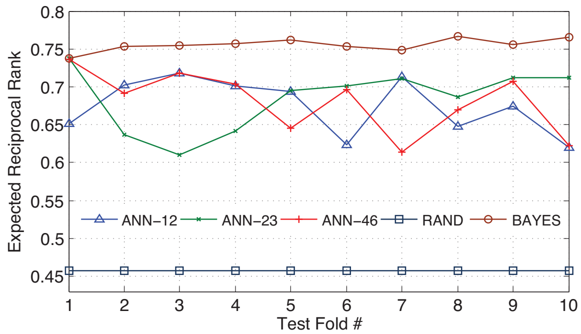


Fig. 8: Comparison of the expected reciprocal ranks of the inference of the categories of traffic-weather index.

**Baseline Methods.** There are two baseline methods implemented in our empirical study. One is *random guess*, where we assume the probability of guessing any WTI value of a cell is equal. Since we have five WTI values, the expected reciprocal rank is around $0.4567$. The other baseline is *artificial neural network (ANN)*, where the input of the ANN are the observations of all Voronoi cells, and it has one hidden layer.

**Results.** In Figure 8, it is clear that our method based on naïve Bayes classifier has the best performance compared to ANN in all settings and random guess, with an expected reciprocal rank around $0.75$. 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 $0.4567$.

### D. Factor Analysis

Given any regional feature in Table I, we verify whether it is a key factor to WTI by generalizing the weight of it using the method introduced in Section II-C. The weight estimation of each feature is shown in Figure 9. It demonstrates some surprising phenomena, for example, *house age* and *the number of neighboring cells* have the highest impact to the WTI. 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 result, 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 I. This 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. These results may help the government to investigate the regions with traffic problems more efficiently.

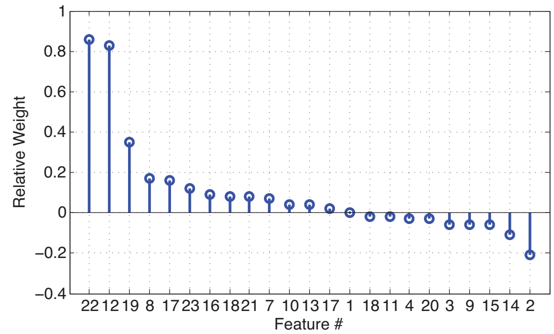


Fig. 9: The weights of the regional features in Table I.

## IV. Take-Away Messages and Conclusion

Through our analysis on the regional weather-traffic indices, we obtains interesting constructive results that sheds light on future urban planning and reconstruction: 1) the regional house age has significant impact on the region's WTI; (2) if a region has a more complex road structure, weather tends to influence the regional traffic more; (3) the regional *community* features such as house age, price have more impact on the WTI, where POI and density features tend to have the least impact. These understandings will benefit government agents to design and manage the functional characteristics of districts throughout a city, to improve traffic prediction, and to learn the key factors in urban planning, etc.

## V. Acknowledgment

This paper was supported in part by National Basic Research Program of China (973 Program) under grant no. 2014CB340303 and NSFC under grant no. 61300030.

## References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM TIST*, vol. 5, no. 3, p. 38, 2014.

[2] M. A. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time traffic crash-prediction model using archived weather and its traffic data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 2, pp. 167–174, 2006.

[3] S. Dunne and B. Ghosh, "Weather adaptive traffic prediction using neurowavelet models," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 370–379, 2013.

[4] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: when urban air quality inference meets big data," in *KDD*, 2013, pp. 1436–1444.

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

[6] B. Pan, U. Demiryurek, and C. Shahabi, "Utilizing real-world transportation data for accurate traffic prediction," in *ICDM*, 2012, pp. 595–604.

[7] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *KDD*, 2011, pp. 316–324.

[8] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *ICANN*, 1997, pp. 999–1004.

[9] C.-H. Wu, J.-M. Ho, and D.-T. Lee, "Travel-time prediction with support vector regression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 276–281, 2004.

[10] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *CIKM*, 2009, pp. 621–630.