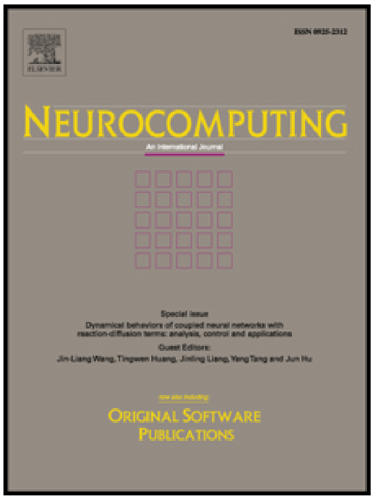
Communicated by Weiming Xiang

Accepted Manuscript



Efficient Traffic Congestion Estimation using Multiple Spatio-temporal Properties

Yongjian Yang, Yuanbo Xu, Jiayu Han, En Wang, Weitong Chen, Lin Yue

PII: S0925-2312(17)31072-X
DOI: 10.1016/j.neucom.2017.06.017
Reference: NEUCOM 18573

To appear in: *Neurocomputing*

Received date: 16 November 2016
Revised date: 3 June 2017
Accepted date: 8 June 2017

Please cite this article as: Yongjian Yang, Yuanbo Xu, Jiayu Han, En Wang, Weitong Chen, Lin Yue, Efficient Traffic Congestion Estimation using Multiple Spatio-temporal Properties, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.06.017

Efficient Traffic Congestion Estimation using Multiple Spatio-temporal Properties

Yongjian Yang[1], Yuanbo Xu[1], Jiayu Han[1], En Wang[1,*], Weitong Chen[2], Lin Yue[3]

[1]School of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia
[3]School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China

***Abstract*—Traffic estimation is an important issue to analyze the traffic congestion in large-scale urban traffic situations. Recently, many researchers have used GPS data to estimate traffic congestion. However, how to fuse the multiple data reasonably and guarantee the accuracy and efficiency of these methods are still challenging problems. In this paper, we propose a novel method Multiple Data Estimation (MDE) to estimate the congestion status in urban environment with GPS trajectory data efficiently, where we estimate the congestion status of the area through utilizing multiple properties, including density, velocity, inflow and previous status. Among them, traffic inflow and previous status (combination of time and space factors) are not both used in other existing methods. In order to ensure the accuracy and efficiency, we apply dynamic weights of data and parameters in MDE method. To evaluate our methods, we apply it on large-scale taxi GPS data of Beijing and Shanghai. Extensive experiments on these two real-world datasets demonstrate the significant improvements of our method over several state-of-the-art methods.**

Keywords:Traffic congestion estimation, Large-scale road networks, Multiple spatio-temporal properties, Dynamic weight calculation, GPS data.

I. Introduction

Currently, advances in terms of Wireless Sensor Networks (WSN) enrich the variety of human mobility information [1, 2]. However, how to use various kinds of massive data (e.g. taxicab GPS trajectory data, road segments, etc.) to solve the real problems is still challenging. Urban traffic congestion is a kind of the real problems, and it has become a critical problem because it not only affects the inhabitants' daily life, but also damages the social and economic development of the city [1]. Nevertheless, urban traffic situation is complex and constantly changing with time and space. The information of traffic usually includes multiple spatio-temporal data, so it is quite difficult for the inhabitants to obtain the current and future traffic condition at a certain road section in time. Estimating the status of urban traffic congestion effectively is the first step to solve the urban traffic congestion problem.

There are three major challenging problems in urban traffic **congestion estimation (CE)**. Firstly, which kinds of data in massive and various information should be analyzed in large-scale urban sections for CE? The sensors equipped in taxicabs, which can collect the Global Position System (GPS) data, are effective facilities [3]. They can offer traffic researchers with massive and detailed real-time traffic data in large-scale urban sections. Ubiquitous mobile sensors can probe a city's rhythm and pulse [4]. GPS data can provide detailed taxicab's information of traffic flow such as longitude, latitude, directions, speed, etc. Using GPS data of taxicabs, we can deduce the whole city's traffic situations because 1) the data can be collected easily and 2) the taxicab has the ability to spread over all road networks in the city than other public traffic [1][5] (shown in Fig.1) (note that every red point is a taxt GPS point). Secondly, because sensors collect the information of traffic periodically and frequently, it is a common view that we have an information explosion. For example, the data collected by 1000 taxicabs in Beijing one day is nearly 10GB text-file. It contains huge information about the congestion in the historical trajectory data. However, tackling all historical data is expensive and meaningless for congestion estimation. Therefore, we should choose the most useful and significant properties in traffic flow data to estimate congestion. Finally, how to improve the accuracy and efficiency of traffic congestion estimation? The existing methods for CE have not addressed the three problems systematically.

This paper proposes a **Multiple Data Estimation (MDE)** method, which estimates congestions in urban sections using traffic flow data (GPS data collected by taxicabs) and road segments. The existing methods for CE usually use one or two properties, while MDE uses traffic density, average velocity, inflow, and previous status as spatio-temporal properties to evaluate a section's congestion situation. These properties are multiple for traffic congestion and difficult to calculate directly. In addition, the weights of properties in MDE are varying in consonance with time and space according to the traffic networks. We use dynamical determination for weights of properties, which can ensure the accuracy and efficiency in

*Corresponding author

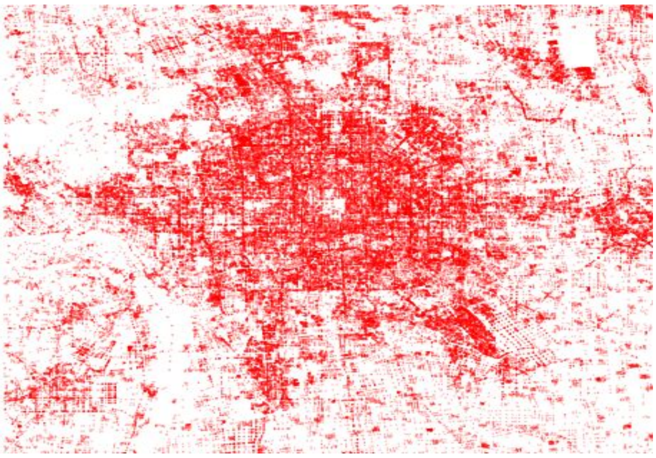


Fig. 1: GPS data collected by taxicabs in Beijing.

comparison with the fixed weight methods.

The main contributions of this paper are as follows:

- We propose a novel method for estimating the traffic state of a section by using four multiple spatio-temporal properties: traffic density, average velocity, inflow and previous status. Inflow represents not only the changing rate of the congestion status but also the influence of surroundings, and previous state represents congestion status of the section from spatio-temporal aspects. With these properties, we propose a framework of MDE method to evaluate the traffic congestion.
- In MDE, the weights of properties are dynamically determined. We use optimization method to determine the weights. The traffic situation is changing periodically with time and space. Therefore, we determine the weights in accordance with different time in one day and different day in one week, which improves the accuracy and efficiency of estimation. Extensive experiments showed that the urban congestion states could be described better in accuracy and efficiency using MDE method.

The rest of the paper is organized as follows. In the next section, we give a brief introduction about the related work. In Section III, we present some basic definitions. We introduce our method framework in IV. MDE is introduced in Section V. We conduct the experiments to evaluate our proposed methods using real taxicab GPS data, describe the procedure and results of the experiments and give an analysis on the experimental results in Section VI. Finally, we conclude with a discussion on our work in Section VII.

II. Related Work

Most evaluation methods used one or more specific characteristics as inputs into their estimation system. Younes et al. [6] proposed an Efficient Congestion Detection protocol to evaluate the traffic characteristics of each road segment. H. Yue et al. [7] apply Linear Constraint Databases to forecast speed of traffic, Pattara-atikom et al. [8] divided the traffic congestion into three levels, and estimated the road traffic congestion using the weighted averages of measured GPS speed. Furthermore, in the paper [9], they studied the relationship between the value of Cell Dwell Time (CDT) to describe road traffic congestion. Instead of using an equal weight for all records [10], Zhang et al. [11] proposed a novel method to reasonably process GPS data by increasing weights of recent records and high velocity to estimate the traffic states. Kong et al. [5] made use of a new fuzzy comprehensive evaluation method in which the weights of multi-indexes are assigned according to the traffic flow. However, these methods don't take spatio-temporal factors into account,which will make estimation more accurately.

Based on vehicle tracking method, Chen et al. [12] presented a method for urban traffic state estimation using the data of GPS and GIS. This method can keep the continuity of vehicle travels. Zhao et al. [13] proposed an improved method based on the Probe-Vehicle-Tracking to estimate the state of urban traffic flow. Based on GPS probe vehicles, Kong et al. [14] presented the curve-fitting-based method and the vehicle-tracking-based method for the traffic state estimation. H. Yue et al. [15][16] use video data to archiving and retrieval. Shi et al. [17] proposed a method whose location-amended GPS data are dynamically fitted with the adaptive traffic flow, and it can estimate the state of the traffic flow along rolling time periods. All these methods just use one or two factors in traffic (density,velocity or traffic flow), but not fuse them together to evaluate traffic estimation like our proposed method MDE.

Data fusion is another effective means to describe the traffic flow states in a comprehensive way. Kong et al. [18] presents an information-fusion-based approach to estimate the states of the urban traffic, the approach fused online data from underground loop detectors and GPS. In another paper [19], the same authors introduced a fusion-based system composed of real-time traffic state surveillance. Zhao et al. [20] built a traffic state estimation system based on fusion of the information of the Geographic Information Systems for Transportation (GIS-T) and GPS. These data-fusion method use GIS and GPS data, which improves the accurancy but has the high expense relatively, while MDE use only GPS data as the source entry of method which has lower expense but good evaluation performance.

Herring et al. [21] proposed a probabilistic modeling framework for arterial traffic estimation with a Coupled Hidden Markov Model (CHMM). Hadachi et al. [22] used the Sequential Monte Carlo to increase the accuracy of travel time estimations. Tabibiazar et al. [23] utilized a kernel-based density estimation method and developed a probabilistic framework to model the traffic data with generalized Gaussian density. Li et al. [24] presents an ensemble learning framework to appropriately combine estimation results from multiple macroscopic traffic flow models.Different from above methods, Pongpaibool et al. [25] used two fuzzy systems which are manually tuned fuzzy logic and adaptive neuro-fuzzy techniques respectively.

III. Preliminaries

In order to make a clear statement of MDE, we will clarify some basic definitions in this section, including GPS trajectory, GST and several important properties in urban road networks. With these definitions, we make a proper definition for congestion estimation which will be solved in MDE.

A. Basic Definitions

Definition 1: GPS trajectory: GPS trajectory *traj* is a data set which describes the objects' movement, including a sequence of time-location points: $traj = \{p_1, p_2, p_3...p_k\}$, where a GPS point $p_i = (id_i, x_i, y_i, t_i, dir_i, spe_i)$, $1 \leq i < k$ with id_i as a unique code for objections (people with mobile phone, city monitor sensor and taxicabs), t_i as a time-stamp ($t_i < t_{i+1}$), (x_i, y_i)as the two-dimension coordinates, dir_i as the moving direction and spe_i as the moving speed.

Definition 2: GST (graph of spatio-time): *GST G (V, E)* is a directed graph in time *t*, which consists of grids(divided as requirement of different applications), where *V* and *E* denote complete set of vertices and edges in GST respectively. Each $v \in G.V$ has a two-dimension location (x_v, y_v) and each $e \in G.E$ has a starting point v_s and ending point v_e. We define e belongs to a grid $g(e \in g)$ in the situation that $e.v_s, e.v_e$ both in g. There are other vertices denoting the GPS trajectory in t. In a word, GST is an abstraction of real world which data comes from GPS device. In GST, we define that all GPS trajectory has been collected right.

Definition 3: Density/Velocity: Density/Velocity of a grid (den/vel) is used to represent different aspects of congestion status. The den describes how many taxicabs in the grid g in t time, while the vel describes the traffic flow status by the average speed of each taxi in g. Both two properties can be calculated using GSTs.

Definition 4: inflow: Inflow of an edge $e(inflow)$ in time interval t is as Eq.(1). $p(e', e, t)$ means cars' flows from e' to e in time interval t. This can be calculated with spe_i,dir_i in GPS data. Likewise, $inflow$ of a grid g in time t is as Eq.(2), which means we should calculate all the edges cutting through the boundary of g and ignore the cars' flow from the road inside g.

$$f_{in}(e,t) = \sum_{e' \in G.E - \{e\}} p(e', e, t) \tag{1}$$

$$f_{g,t} = \sum_{\substack{(e.v_s \in g \cap e.v_e \notin g) \cup \\ (e.v_e \in g \cap e.v_s \notin g)}} \left(f_{in}(e,t) - \sum\nolimits_{e' \in g} p(e', e, t) \right) \tag{2}$$

Definition 5: Detailed GST: $DetailedGST$ has the same construction as GST, but add some detailed information such as density, velocity and inflow of each grid to enrich the expression of the GST.

Definition 6: CQS (Congestion Quality Status): In $GST\ G$, CQS describes the congestion level. CQS of a grid $g.CQS$ is divided into five states *(quite rapid, rapid, normal, crowed, blocked)* as basic evaluation sets. A subgraph S consists of a series of grids in G. The CQS of a subgraph $S.CQS$ is decided by $G.CQS$ of grids in this subgraph. And in this paper, we use 0,1,2,3,4 to present the relative basic states to simply our description. A CQS matrix is corresponding to a GST, with the values (0,1,2,3,4) signed for each grid in it.

B. Problem Definition

Congestion Estimation(CE): Given a series of GST $G_t\{G_{t1}, G_{t2}, G_{t3}...G_{ti}\}$, the congestion estimation (CE) is to assign the grid $g\{g_1, g_2, g_3...g_n\}$ in G_t into different status (CQS), such that:

1) $\forall g \in G_t$ satisfies **Definition 2**

2) After CE, each g should have a status, which satisfies **Definition 6**.

3) If a subgraph S consists of many grids $\{g_1, g_2, g_3...g_m\}$, and each grid has a CQS, after CE, $S.CQS$ should be provided. $S.CQS$ also satisfies **Definition 6**.

Using CE, the traffic researcher can know the traffic congestion better, traffic supervisors can plan or design the traffic network more efficiently and drivers can decide their own traffic path more conveniently.

Because the information in G_t is various, including the GPS data and the road network data, and various properties in traffic networks influence each other and may reinforce mutually, it is difficult for the traffic researchers to estimate the congestion status with all the original information. The volume of the data in G_t is massive for traditional estimation method to calculate. In MDE, we use multiple spatio-temporal data in traffic networks and the neighbors' congestion state as the standard of the section, and thus it is more accurate to describe the congestion state.

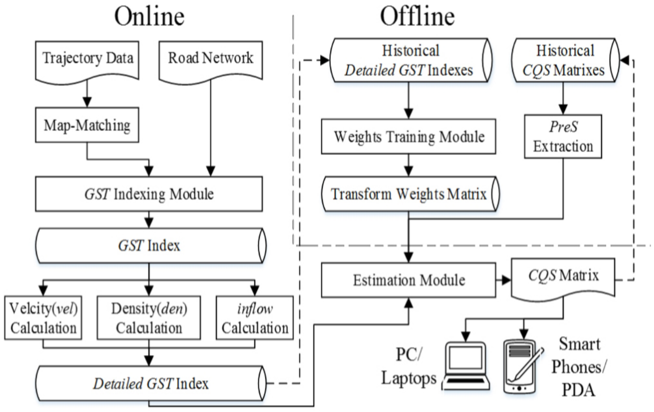


Fig. 2: Framework of Multiple Data Estimation method.

IV. Framework

In this section, we demonstrate our method's framework in Fig.2. Our method consists with two parts: Online and Offline. Users can access the system using PC/Laptops or Smartphones/PDA. Trajectory data and Road network as the real-time input are tackled online and transformed into GST index. Then the system uses GST to calculate the multiple spatio-temporal properties needed in MDE, including den, vel and $inflow$. Meanwhile, offline part calculates other properties and trains the weight matrix with historical Detailed GSTs and CQS matrices. Finally, our estimation module uses the properties and the weight matrix as input, and outputs the CQS matrix to the users.

Off-Line: Off-Line part should tackle with the massive historical data collected by taxicabs. As the historical data are comprised of various kinds of data, such as GPS trajectory data (Definition 1), GST data and Congestion Quality Status (Definition 6), we use Hadoop to store and filter the historical data and train the weight matrix in the distributed system,which can accelerate the calculation and improve efficiency of MDE.

On-Line: On-Line part consists of six sub-modules, including Map-matching, GST Indexing, Density Calculation, Velocity Calculation, Inflow Calculation and Estimation Module.

Map-matching and GST Indexing: In order to ensure the data quality, we normalize the GPS data collected by taxicabs. Firstly, we use Map-matching [26] to adjust the trajectory data into the real map. The map will be transformed into GST in GST Indexing Module. This module use Hash Mapping to map the grid matrix into a list, GST index and every set of the list contains these grids' attributes. With these two modules, we can normalize the trajectory data and reduce the space to store GSTs.

V. Urban Congestion Estimation

In this section, we first introduce the MDE CE method with appropriate spatio-temporal multiple properties to evaluate the congestion. In addition, the method to decide the weights of properties dynamically are also proposed. Finally, we described the whole procedure of our proposed method MDE.

A. Determination of Multiple Properties in Congestion Estimation

First, the number of cars in a section (density of a section: den) is the most significant property to estimate urban congestion [4-6]. The cars' quantity in the road can reflect the congestion status directly. Otherwise, the average velocity (vel) is another important property. The den and the vel can be calculated with GPS data collected by taxicabs. Beyond these two usual properties, there are two spatio-temporal properties we consider as well as important in CE:

Inflow ($inflow$): Inflow of a section means the increment of den. Moreover, inflow can be computed with GPS data. The estimation for urban congestion should not only describe the status of the moment, but also make a guidance for a period in the future. In addition, $inflow$ is also considering the influence of neighbors. With the help of $inflow$, the estimation can be more instructive.

Previous Status ($PreS$): GPS data collected by sensors equipped in taxicabs are massive and frequent uploading. Therefore, we can get quite a lot information about the section we want to evaluate in previous time. The status of a section usually does not change sharply among continuous time intervals except for some special reasons (traffic accidents, or other emergencies). $PreS$ means the previous status of the grid we want to evaluate. In this paper, $PreS$ is a sequence of CQSs. In MDE, we use two different kinds of $PreS$: a sequence of CQSs just before the time t we want to evaluate the congestion status in this section ($T.PreS$), and another sequence of CQSs for the same time t in this section in history ($S.PreS$). $T.PreS$ is corresponding to the temporal continuity while $S.PreS$ is corresponding to the space continuity.

The existing methods to evaluate the urban congestion often only use traffic volume, traffic speed, or mixture of both. Honestly, den and vel are leading properties to characterize the traffic congestion. However, we find quite many congestions can be described more in details with $inflow$ and $PreS$ through traffic investigations. In this paper, we use four properties: den, vel, $inflow$, $PreS$ to evaluate a section's traffic congestion status.

In MDE, den and vel represent the situation of the section we want to evaluate, $inflow$ represents the neighbors' contribution to the section, the changing rate of den, and $PreS$ represents the historical status and congestion pattern. Four properties focus on the different aspects to describe the congestion status, and reinforce mutually. So CE problem transforms into a multiple spatio-temporal data-learning problem. MDE can tackle the multiple data-learning problem as described in next section.

Meanwhile, noting that the section's congestion status plays an important role for traffic researchers, drivers and guidance systems. MDE uses *quite rapid-0, rapid-1, normal-2, crowed-3, blocked-4* as its basic evaluation description sets.

B. MDE Congestion Estimation Method

1) Section Estimation: As we proposed above, we have five basic estimation levels: *quite rapid, rapid, normal, crowed, blocked* and four multiple properties in MDE: traffic density den, average velocity vel, inflow f and previous status $PreS$. The previous three properties should be calculated by real time GPS data collected by sensors in taxicabs. As Definition 1, original GPS data are massive and hard to calculate directly. Therefore, we use GST Indexing to transform GPS trajectory data and road network into GST Index. Through GST, we can calculate den with Eq.(3), vel with Eq.(4) and f with Eq.(2) for each grid g in time t.

$$den_{g,t} = \sum_{p_i \in g} p_i \tag{3}$$

$$vel_{g,t} = \frac{\sum_{p_i \in g} spe_i}{den_{g,t}} \tag{4}$$

den, vel, f are added to GST and transform GST to $DetailedGST$. As introduced before, the previous status $PreS_{g,t}$ can be extracted from historical estimation data. We use FAC_t (Matrix of Factors) to present the properties of each grid in time t:

$$FAC_t = \begin{pmatrix} den_{0,t} & vel_{0,t} & f_{0,t} & PreS_{0,t} \\ \vdots & \vdots & \vdots & \vdots \\ den_{i,t} & vel_{i,t} & f_{i,t} & PreS_{i,t} \end{pmatrix} \tag{5}$$

Original FAC_t should be transformed into a weighted matrix $TFAC_t$ by multiply a transform matrix $Trans$ in order to satisfy the threshold ∂ :

$$\begin{aligned} TFAC_t &= FAC_t \bullet Trans \\ &= \begin{pmatrix} Tden_{0,t} & Tvel_{0,t} & Tf_{0,t} & TPreS_{0,t} \\ \vdots & \vdots & \vdots & \vdots \\ Tden_{i,t} & Tvel_{i,t} & Tf_{i,t} & TPreS_{i,t} \end{pmatrix} \end{aligned} \tag{6}$$

So right now we can evaluate each gird in GST to describe the traffic congestion in t time with the matrix $C(t)$:

$$C(t) = \left(\sum_{j=0}^{i} TFAC_t \bullet \left(W(t)^T \bullet A_j \right) \right)^T \tag{7}$$

In Eq.7, W is the weight matrix of properties, A_j is used to pick every col in $W(t)^T$ as a $i*1$ matrix in which j col is set to 1 and other cols are 0. We can use $C_{g,t}$ in $C(t)$ and ∂ to evaluate a single grid g:

$$CQS_{g,t} = \begin{cases} 0, quiterapid\,(C_{g,t} \le \partial_1) \\ 1, rapid\,(\partial_1 \le C_{g,t} \le \partial_2) \\ 2, normal\,(\partial_2 \le C_{g,t} \le \partial_3) \\ 3, crowed\,(\partial_3 \le C_{g,t} \le \partial_4) \\ 4, blocked\,(\partial_4 \le C_{g,t}) \end{cases} \tag{8}$$

2) Determination of Weights in Congestion Estimation: After determining the properties, basic evaluation description sets and method framework to evaluate the grid, we need to determine the transform matrix $Trans$, the weight matrix of properties W and the threshold ∂ for MDE.

While different cities and sections have different situations about the congestion, it is not reasonable to determine fixed weights. Therefore, we employ a mixed-determination method, which uses experts' experience and the real situation about the city and every section we want to evaluate. Experts' experience is usually subjective and various because different experts have different opinions on the traffic situation. The real situation about the city is objective because it comes from the real historical data. Nevertheless, it has not covered such many properties, especially human properties as experts have.

To avoid the shortage of both two methods, we determine to use mixed-determination method. We use historical $DetailedGST$ indexes and historical CQS matrices. Historical CQS is the conclusion of experts' experience and historical estimation results. Moreover, historical $DetailedGST$ represents the historical real situation of the traffic.

First, we determine the transform matrix $Trans$, which is used to normalize the multiple spatio-temporal properties in FAC to satisfy the requirement of MDE. The four properties in MDE, den, vel, $inflow$ are values while $PreS$ consists of two sequences of CQSs: $S.PreS$ and $T.PreS$. Therefore, we use $Trans$ to normalize $PreS$. In MDE, we use 0,1,2,3,4 to represent five CQS status, so $S.PreS$ and $T.PreS$ consist of values such as:

$$\begin{aligned} &S.PreS_{g,t} = \{CQS_{g,t_{s_1}}, CQS_{g,t_{s2}}...CQS_{g,t_{si}}\} \\ &T.PreS_{g,t} = \{CQS_{g,t-\Delta t}, CQS_{g,t_i-2\Delta t}, \\ &CQS_{g,t_i-3\Delta t}...CQS_{g,t-j\Delta t}\} \end{aligned} \tag{9}$$

In $S.PreS$, $CQS_{g,t_{s_1}}$ means congestion status the same time t a period before. For example, if we want to evaluate the grid g at 13:00 on Monday, $\{CQS_{g,t_{s_1}}, CQS_{g,t_{s2}}...CQS_{g,t_{si}}\}$ means the CQSs of g on Monday before one week, two weeks...i weeks. In $T.PreS$,$\{CQS_{g,t-\Delta t}, CQS_{g,t_i-2\Delta t}...CQS_{g,t-i\Delta t}\}$ means the status of g at the nearest time interval Δt, 2Δt...$j\Delta$t. We should normalize the two sequence into a value to present the spatio-temporal relationships, such as:

$$\begin{aligned} PreS_{g,t} = \lambda \sum_{n=1}^{i} P(t_n) S.PreS_{g,t}(t_n) \\ +(1-\lambda) \sum_{m=1}^{j} Q(t_n) T.PreS_{g,t}(t_m) \end{aligned} \tag{10}$$

$$\begin{gathered} P(t_a) \geq P(t_b), 1 \leq a \leq b \leq i \\ P(t_1) + P(t_2) + ...P(t_i) = 1, P(t) \geq 0 \\ Q(t_a) \geq Q(t_b), 1 \leq a \leq b \leq j \\ Q(t_1) + Q(t_2) + ...Q(t_j) = 1, Q(t) \geq 0 \end{gathered} \tag{11}$$

In Eq.(10)and(11), $P(t)$ and $Q(t)$ are the functions to determine the importance of each CQS in $S.PreS$ and $T.PreS$. The traffic flow is usually related to the status in the nearest time. So $P(t)$ and $Q(t)$ are monotonic decreasing functions. λ is used to decide the importance of $S.PreS$ and $T.PreS$. We can adjust λ to balance the importance of time and space. In this paper, λ is set as 0.5 (explained in Section VI). It means they have equal importance. With this transform function, we determine the transform matrix $Trans$.

Then, we use the optimization method to determine the weight matrix of properties W. We divide a day into 48 time intervals, which means we need to calculate 48 different W for every time intervals to ensure the accuracy and efficiency.

$W(t)$ can be presented as:

$$W(t) = \begin{pmatrix} W_1 \\ \vdots \\ W_i \end{pmatrix} = \begin{pmatrix} w_{10} & \cdots & w_{13} \\ \vdots & \ddots & \vdots \\ w_{i0} & \cdots & w_{i3} \end{pmatrix} \tag{12}$$

In $W(t)$, k row is for a grid g_k in GST index, 1≤k≤i, $w_{k0}+w_{k1}+w_{k2}+w_{k3}$=1, 0≤$w_k$≤1, where weights are normalized between 0 and 1. Weights matrices should be updated with the GPS data increasing or road network changing. Historical CQS matrices can be presented as:

$$C_h(t_m) = (CQS_{g_1,t_m}, CQS_{g_2,t_m}...CQS_{g_i,t_m}) \tag{13}$$

Otherwise, we use historical $DetailedGST$. With the Eq.(5) and (6), we can calculate $hTFAC$ (historical $TFAC$). Considering the characteristics of urban traffic, we choose $hTFAC$ on the same day in a week as the train set of $W(t)$. For example, if we want to calculate W for 13:00 on Monday, choose the other $hTFAC$s at 13:00 on Monday in the former weeks as the train set. So the determination of W at time t can be transformed into a linear optimization problem, as shown in formula (14).

$$\begin{aligned} \arg\min_{W_k} \{\lambda_C (C_h(t) - (\sum_{j=0}^{i} hTFAC_t \bullet (W^T(t) \bullet A_j))^T) \\ -(\tfrac{\lambda_W}{2} \|W\|_F^2 + \tfrac{\lambda_P}{2} \|P\|_F^2 + \tfrac{\lambda_Q}{2} \|Q\|_F^2)\} \end{aligned} \tag{14}$$

subject to constraints (6) (7) (10) (11) and (13);

In (14), $\lambda_C, \lambda_W, \lambda_P, \lambda_Q$ are penalty coefficients, and $\|W\|_F^2, \|P\|_F^2, \|Q\|_F^2$ are 2-Norm of W, P, Q as penalty terms to avoid over-fitting. Because of the huge calculation and complex computation, we use stochastic gradient descent (SGD) as our optimization method, and ε as the condition of convergence. SGD is sensitive to the initialized value [15]. Therefore, we set the initialized W following the investigation of urban traffic and the experts' experience, to make sure the W can present the importance of each properties. To improve the accuracy and efficiency of SGD, we set the learning rate $\eta_0 = 0.05$ at first. In t update process, set η_t as $\eta_t = \frac{\eta_0}{1+\kappa\times t}, \kappa = 10^{-3}$ to update smoothly. And P, Q are initialized descending random sequence of value in range 0 to 1, which satisfy Eq.(11).

With dynamic weights, our method improves the accuracy and the efficiency of CE compared to the methods that used fixed weight determination.

Finally, we set the threshold ∂ for MDE. In this paper, we use five integer (0,1,2,3,4) to present all the statuses of the congestion. So we define (0.5,1.5,2.5,3.5) as the threshold. The details of MDE method are given in Algorithm 1.

C. Complexity Analysis

In MDE (shown in Algorithm 1), the loop from $line: 6$ to $line: 14$ needs to calculate $(k*4)$ times for every grid weights W in GST. Outside the loop, in $line: 1$, $line: 2$, den, vel, f and $PreS$ need to be calculated $(n+n+n^2+4*t^2)$ times for each grid (Eq.(1)(2)(3)(4)(10)), n means the number of taxicabs in a grid and t means how many previous status have been considered in $PreS$. Therefore, the time complexity of our proposed method MDE is $(i*k*4+i*2n+i*n^2+i*4*t^2)$<$(i^2+i^2+i^2*n+i^2)$=$(3i^2+i^2*n)$ (i is the number of grids in GST and in congestion estimation, $n \ll i$, $t^2 \ll i$). So the time complexity is $O(n*i^2)$. And for whole GST, we should only store historical CQS $(t*i)$ and historcial GST $(4*t*i)$, MDE's space complexity is $O(i^2)$. Both complexities are acceptable for computing ability of a distributed computing system.

Algorithm 1 MDE algorithm

Input: GST G,Historical $hGST$,historical $hCQS$,threshold ∂,initialized W_0,
tolerance ε,time t,maximum iteration count k_{max},$Trans$,
grid number i, learning rate η_0
Output: Congestion matrices CQS_t

Compute den,vel,and f using G for each grid;
Compute $PreS$ using $hCQS$ for each grid;
$FAC_t \Leftarrow den, vel, f, PreS$;
$TFAC_t = FAC_t \bullet Trans$;
Bulid $hTFAC_t, C_h(t)$ using $hCQS, hGST$;
$k = 0$;
while $k \leq k_{max}$ **do**
  $F(W_k)$=Eq.(14);
  if $F(W_k) \leq \varepsilon$ **then**
    $W = W_k$, **break**;
  else
    $W_{k+1} = W_k - \eta_k grad(F(W_k))$;
    k++;
    update $\eta_{k+1} = \frac{\eta_0}{1+\kappa\times k}$;
$C(t) = \left(\sum_{j=0}^{i} TFAC_t \bullet \left(W(t)^T \bullet A_j\right)\right)^T$;
Compute CQS_t using $C(t)$ and ∂;
Update $hGST, hCQS$;

VI. Experimental Results

In this section, we first describe the GPS data, the pretreatment of traffic data and performance indicators. Then, we conduct the experiment to evaluate our proposed method. In our experiments, we use GPS data from Beijing and Shanghai to ensure the authenticity and reliability.

A. Data Description

In the experiment, we use *Beijing Taxicab Data* and *Shanghai Taxicab Data* to evaluate our method.

Beijing Taxicab Data: The Beijing Taxicab Data is collected from more than 30,000 taxis, in Nov, 2012. It contains over 140,000 road nodes and 90,000 road segments. The data are recorded and uploaded once per minute.

Shanghai Taxicab Data: The Shanghai Taxicab Data is collected from more than 14,000 taxis, in Apr 2015. It contains over 120,000 road nodes and 120,000 road segments. We record and upload the data once per ten seconds.

In our experiment, we use [116.40, 39.90]-[116.46, 39.94] of longitude, latitude in Beijing taxicab data as GST's diagonal point. Moreover, we divided it into 30×20 grids, which range of a grid is 200m×200m. And we use [31.34, 121.62]-[31.09, 121.24] of longitude, latitude in Shanghai taxicab data as GST's diagonal point. Moreover, we divided it into 70×105 grids, which range of a grid is 400m×400m.

B. Performance Indicators

We consider the performance indicators from different aspects. These indicators are divided into two aspects: accuracy indicators and efficiency indicators.

1) Accuracy Indicators: The accuracy indicators include mean absolute error (MAE), normalized mean absolute error (NMAE), mean square error (MSE), normalized mean square error (NMSE) and hamming loss (HLoss), as shown in following equations:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{m}\sum_{g=1}^{m}\left|CQS_{g,t}^{real} - C_{g,t}\right| \quad (15)$$

$$NMAE = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{m}\sum_{g=1}^{m}\frac{\left|CQS_{g,t}^{real} - C_{g,t}\right|}{CQS_{\max}^{real} - CQS_{\min}^{real}} \quad (16)$$

$$MSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{1}{m}\sum_{g=1}^{m}\left(CQS_{g,t}^{real} - C_{g,t}\right)^2} \quad (17)$$

$$NMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\frac{1}{m}\sum_{g=1}^{m}\left(\frac{CQS_{g,t}^{real} - C_{g,t}}{CQS_{\max}^{real} - CQS_{\min}^{real}}\right)^2} \quad (18)$$

$$HLoss = \frac{1}{n}\sum_{t=1}^{n}\frac{1}{m}\sum_{g=1}^{m}\left|CQS_{g,t}^{real} - CQS_{g,t}^{predict}\right| \quad (19)$$

In this equation, n donates the time range, m donates the space range, $C_{g,t}$ denotes the MDE calculation results for grid g in t time, CQS^{real} denotes the real congestion situation, and $CQS^{predict}$ denotes the estimation status in MDE.

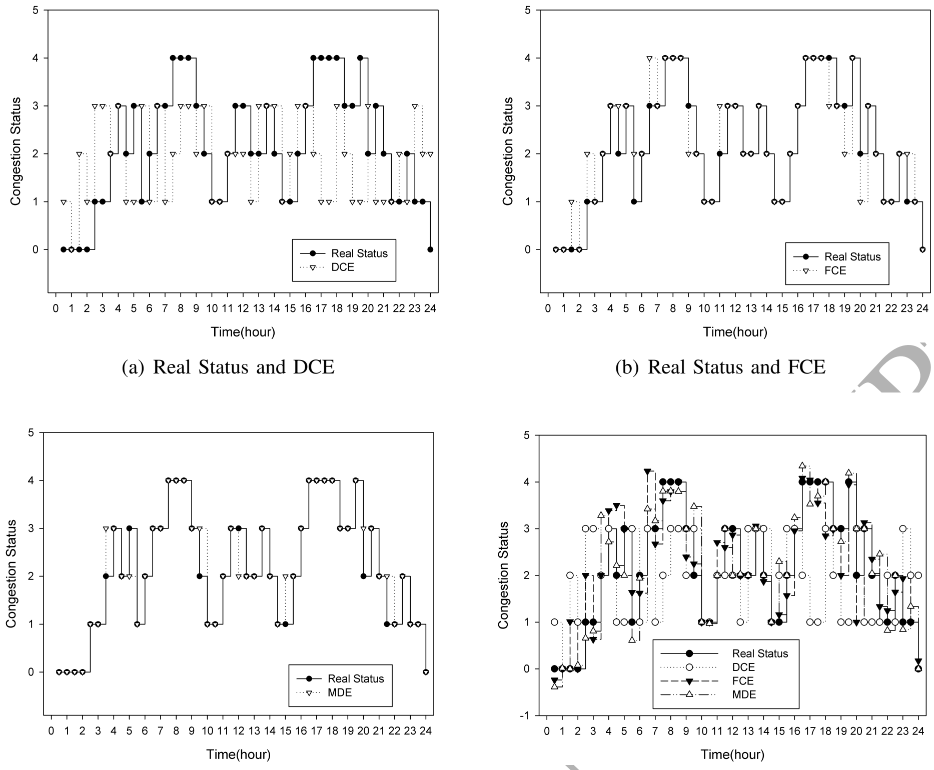
We use the former four indicators to evaluate the ability for calculating the status. Meanwhile, HLoss is the most valuable indicators for us to consider because the best estimation method should have the best HLoss performance.

2) Efficiency Indicators: The efficiency indicators consist of two parts: estimation time (*es-time*), updating time (*up-time*). *up-time* donates to the feasibility of changing the attributes in estimation method. *es-time* is estimation time,which is spended to calculate CQS.

C. Experiment and Discussion

We choose two baseline methods for traffic CE:

Density Clustering Estimation (DCE) [9]: DCE only uses density to evaluate a section's congestion. It is the most common estimation method in real application. DCE usually uses K-means clustering method to find out the different clusters in maps. With these clusters, DCE can describe the congestion status. In DCE, once a new grid we want to



(a) Real Status and DCE (b) Real Status and FCE

(c) Real Status and MDE (d) Real Status and three methods

Fig. 3: CE results Comparison between different methods and real status at Weekends

TABLE I: Comparison Among Characters of Different Estimation Methods

	DCE	FCE	MDE
Implement	Online	Online-Offline	Online-Offline
Data Input	GPS data	GPS data, Road Network	GPS data, Road Network
Chosen Properties	*den*	*den,vel, road segment*	*den,vel, inflow,PreS*
Weights	Fixed	Fixed	Dynamical
Threshold	Fixed	Fixed	Fixed
Historical Data	Used	Used	Used

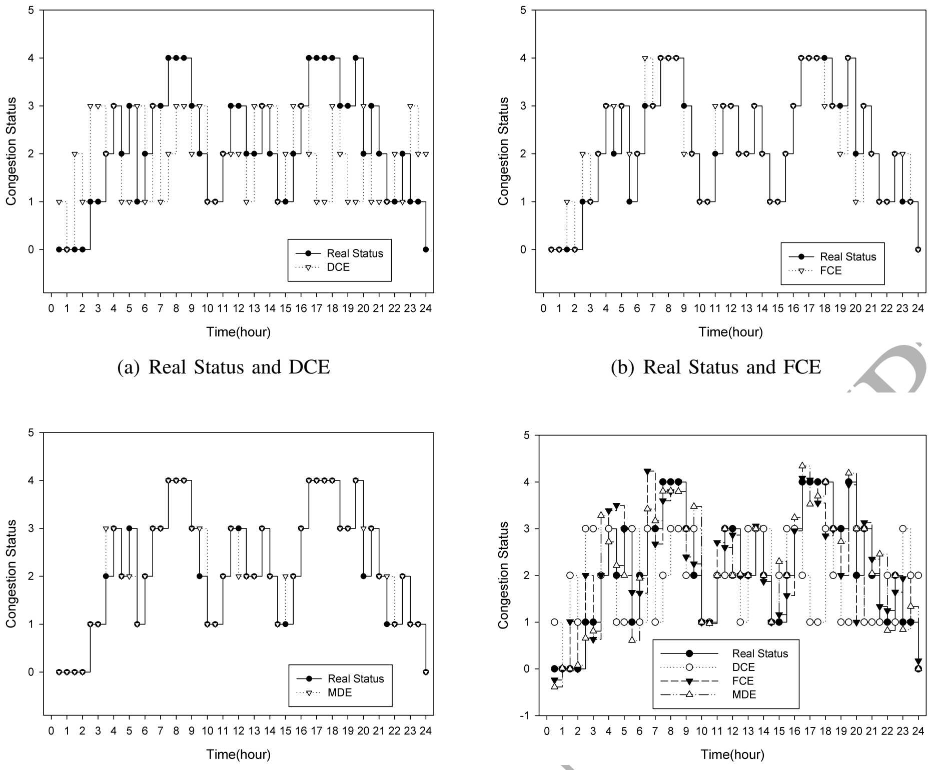
evaluate comes, it needs to do the clustering to find out the congestion level online.

Fuzzy Comprehensive Estimation (FCE) [5]: FCE is an estimation method to generate appropriate properties and evaluation sets for traffic congestion. FCE uses density, average speed and road segment as the properties to evaluate congestion. Like our method, FCE uses subfunctions to calculate a level rank for urban section, and thresholds to decide the congestion status. The threshold and weights of property are fixed in FCE, and divided into three situations: morning peak, usual time and evening peak. Table I shows the comparison of the methods.

TABLE II: Congestion Estimation Performance on Weekends with Different Methods

Indicators	DCE	FCE	MDE	MDE-ran
MAE	0.687	0.165	0.044	0.078
NMAE	0.171	0.052	0.016	0.031
MCE	0.957	0.211	0.106	0.153
NMCE	0.239	0.044	0.021	0.041
HLoss	0.687	0.279	0.104	0.177
es-time	4.301s	0.069s	0.070s	0.081s
up-time	0s	0.801s	0.803s	0s

1) Initial Parameters Determination: DCE needs to set the cluster number and cluster thresholds. In order to compare with other methods, we set five clusters corresponding to CQS



(a) Real Status and DCE (b) Real Status and FCE

(c) Real Status and MDE (d) Real Status and three methods

Fig. 4: CE results Comparison between different methods and real status on Weekdays

statuses in MDE method and the thresholds are 5,8,10,15 in Beijing and 10,20,30,40 in Shanghai for the density of GPS points.

FCE needs to set the congestion levels, weights of properties for each period, and the thresholds. Similarly, we set five congestion levels for FCE, and the weights for density, velocity, road segment are set as W_p=[0.43,0.27,0.3] for morning peak and evening peak, W_u=[0.23,0.17,0.6] for other time. In addition, we set four thresholds to classify the congestion status.

MDE needs to set the congestion levels CQS (5 level is set according to historical dataset), threshold ∂ , initial weights matrices of properties W, initial transform matrices $Trans$, and some attributes in training procedure (learning rate, etc.). CQS , ∂ are set as described before. W is set randomly as w_m=[0.3,0.3,0.2,0.2] as initial value, 1≤m≤4. We decide to use four periods of historical data to constitute $PreS$, which is set in $Trans$. When training W, we set ε to 0.001.

2) Estimation with Baseline Methods: We choose the congestion at the last week in Nov, Beijing and in Apr, Shanghai. Then we divide a week into two kinds: Weekdays and Weekends. In our experiment, one day is consists of three periods:

TABLE III: Congestion Estimation Performance on Weekdays with Different Methods

Indicators	DCE	FCE	MDE	MDE-ran
MAE	1.145	0.170	0.101	0.133
NMAE	0.286	0.042	0.027	0.033
MCE	1.414	0.235	0.118	0.162
NMCE	0.353	0.058	0.024	0.038
HLoss	1.145	0.229	0.145	0.202
es-time	4.741s	0.052s	0.055s	0.061s
up-time	0s	0.792s	0.810s	0s

0:00-7:00, 9:00-16:00, 19:00-24:00 (usual time), 7:00-9:00 (morning peak), and 16:00-19:00 (evening peak). In addition, to verify different initial values' effect on MDE method, we randomly pick w_m for 100 times, and calculate the average MAE, NMAE, MCE, NMCE, Hloss and *es-time* for MDE-ran.

The experiment results of traffic congestion estimation for grid A [116.43,39.92] in Beijing are shown in Fig.3 and Table.II for weekends and in Fig.4 and Table.III for weekdays.

From the results, we find that weekdays and weekends have the different congestion status pattern. In Table.II and Table.III, we can see that DCE has the worst performance in three methods. Moreover, in Fig.3(a) and Fig.4(a), DCE shows its obvious shortage of accuracy and efficiency. With huge calculation complexity and low accuracy, DCE is not a good method for traffic estimation.

We compare FCE and MDE through the results of experiment. From Table.II and Table.III, we can see that MDE has a better performance in the first five indicators, especially HLoss (even MDE-ran has overmatched FCE, which means MDE is still better wiht a random initial value). The HLoss of MDE is improved nearly 50% by MDE comparing with FCE on either weekdays or weekends, which means the wrong congestion status of grid estimated by MDE is half of the results estimated by FCE. That is a significant improvement. Although the *es-time* and *up-time* of MDE is a little longer than FCE, it can be remedied by DS-compute-scheme or the optimation of matrix calculation because most time is used to determine the weights matrix W.

The statistical analysis comparison between Fig.3,Fig.4 shows that evaluations for bounds (*blocked* or *quite rapid*) is relatively estimated with high accuracy for both FCE and MDE. But MDE has a better performance than FCE in the middle congestion status estimation. In order to see the difference between different situation of traffic flow, we compare FCE and MDE for three time periods–morning peak(MP), usual time(UT) and evening peak(EP) to estimation section B in Shanghai. We use the average HLoss for seven days as the results of experiment, as shown in Table.IV.

The results show that in the morning peak and the evening peak, FCE and MDE have the relatively little difference for estimation (MDE is still better than FCE). But for the usual time, MDE improves the accurancy (HLoss) twice of FCE. Besides, FCE and MDE have the same performance in NMCE, which means MDE is better than FCE in urban congestion estimation.

TABLE IV: Performance Comparison among FCE and MDE Estimation Results in Different Periods

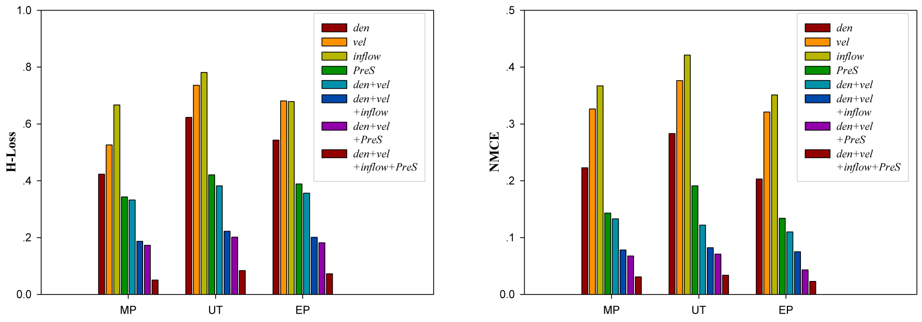
Indicators	FCE	MDE
HLoss MP	0.11	0.051
HLoss UT	0.22	0.104
HLoss EP	0.156	0.094
NMCE MP	0.044	0.031
NMCE UT	0.105	0.035
NMCE EP	0.032	0.023

3) Effect of Different Factors on MDE: There are four factors in MDE: den, vel, $inflow$ and $PreS$. These four factors greatly affect the performance of MDE. To verify the effect of factors, therefore, we conduct the combination of different factors. Fig.5 shows the effect on HLoss and NMCE. From Fig.5, we got some conclusions about MDE: 1) Different factors have different effects on performance of MDE, 2) if using only one factor in MDE, $PreS$ gets the best performance, rather than any other three factors, 3) the performance of combination with two or three factors is better than using only one factor and 4) the combination of four factors that we have applied in MDE leads to the best estimation performance.

As mentioned in Section V, $PreS$ is a novel notion in traffic congestion estimation which consist of space factor ($S.PreS$) and time factor ($T.PreS$). In Fig.5, we find that only using $PreS$ as input of MDE can get a much better performance than using den,vel, and $inflow$ respectively. So $PreS$ may make the greatest effect on MDE. In order to explore the effect of $PreS$, we do some experiment about parameters λ, i, j in Eq.10. i presents the number of CQS in $S.PreS$, while j presents the number of CQS in $T.PreS$, and λ decides the importance of $S.PreS$ and $T.PreS$. In MDE, we define $i = j$ to simplify calculation. In order to understand the relationships between them, we change λ and i respectively with other factors fixed. The effect on MDE by λ and i is shown in Fig.6.

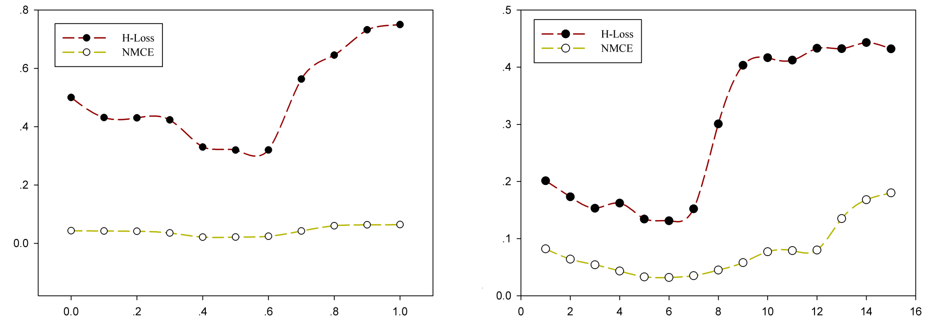
In Fig.6(a), we can see that the HLoss has a great change with λ, while NMCE doesn't. if λ is close to 0, it means we use $S.PreS$ (space factors) as the important part of $PreS$. If λ is close to 1, it means $T.PreS$ (time factor) is more important. Note that HLoss is smallest in the middle of curve. We conclude that the balance between space and time can achieve a better performance. So we define λ to 0.5 in our proposed method. In Fig.6(b), the effect on MDE by the CQS number i in $PreS$ is shown. When i is small, HLoss and NMCE are close to using only three factors (den,vel and $inflow$) in MDE. While in the middle of curve, $i = 5, 6, 7$, it achieves the best performance. When i is larger than 10, the result is close to that only uses $PreS$, which means MDE overfits the $PreS$. In order to simplify the calculation, avoid the overfitting and get the best performance, we choose $i = 5$ as the number of CQS in $PreS$.

From the experiment, we have seen MDE and FCE methods outperform DCE in all the performance indicators, which means methods with multiple spatio-temporal properties are better than the method with single property. Moreover, MDE considers the historical data and traffic flow, and has a better performance than FCE. Finally, MDE has the best comprehensive performances, which can satisfy the requirements of



(a) Effect on HLoss (b) Effect on NMCE

Fig. 5: Effect on MDE by different combination of factors



(a) λ's Effect on MDE (b) CQS number i's Effect on MDE

Fig. 6: Effect on MDE by λ and CQS number i in $PreS$

accuracy and efficiency.

VII. Conclusion

In this paper, we investigated how to evaluate traffic congestion accurately and efficiently using multiple spatio-temporal properties. A novel urban traffic congestion estimation method MDE is used to evaluate the congestion status of a section in the city. This method uses GPS trajectory data and road segment and calculates four significant multiple properties in urban traffic: trajectory density, average velocity, traffic inflow and previous status. We formulate the traffic data into a multiple data set which covers all important traffic spatialtemporal information, and gives full play on the multiple correlations with traffic congestion. In order to ensure the congestion estimation performance in terms of accuracy and efficiency, the weights of properties and parameters are determined dynamically. MDE can tackle different situations in traffic status estimation like data missing, low quality and low accurate better than some existing methods. Through the experiments with real traffic data in Beijing and Shanghai, the results showed that our proposed method is a better traffic estimation method outperforming baseline methods.

In the future, we plan to optimize the matrix calculation in MDE. Moreover, we will consider much more complex problems in traffic situation, such as data missing, transport delay, and traffic prediction. Our estimation has a certain delay about the real situation. Therefore, we plan to improve the estimation speed as much as we can in the case of ensuring the estimation accuracy and efficiency.

Acknowlegement

This paper is supported by the Project 2016184 Supported by Graduate Innovation Fund of Jilin University and Project 20160204021GX Supported by Key Projects of Science and Technology Development Plan of Jilin Province.

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *Acm Transactions on Intelligent Systems & Technology*, vol. 5, no. 3, pp. 222–235, 2014.

[2] A. Kaklauskas, *Intelligent Decision Support Systems*. Springer International Publishing, 2015.

[3] J. Liu, O. Wolfson, and H. Yin, “Extracting semantic location from outdoor positioning systems,” in *International Conference on Mobile Data Management*, 2006, pp. 73–73.

[4] Y. Zheng, “Trajectory data mining: An overview,” *Acm Transactions on Intelligent Systems & Technology*, vol. 6, no. 3, pp. 1–41, 2015.

[5] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, “Urban traffic congestion estimation and prediction based on floating car trajectory data,” *Future Generation Computer Systems*, vol. 61, pp. 97–107, 2015.

[6] M. B. Younes and A. Boukerche, “A performance evaluation of an efficient traffic congestion detection protocol (ecode) for intelligent transportation systems,” *Ad Hoc Networks*, vol. 24, no. 2, pp. 317–336, 2014.

[7] H. Yue, E. G. Jones, and P. Revesz, “Local polynomial regression models for average traffic speed estimation and forecasting in linear constraint databases,” in *International Symposium on Temporal Representation and Reasoning*, 2010, pp. 154–161.

[8] W. Pattara-Atikom, P. Pongpaibool, and S. Thajchayapong, “Estimating road traffic congestion using vehicle velocity,” in *International Conference on ITS Telecommunications Proceedings*, 2006, pp. 1001–1004.

[9] W. Pattara-Atikom, R. Peachavanish, and R. Luckana, “Estimating road traffic congestion using cell dwell time with simple threshold and fuzzy logic techniques,” in *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*, 2007, pp. 956–961.

[10] C. H. Lo, W. C. Peng, C. W. Chen, T. Y. Lin, and C. S. Lin, “Carweb: A traffic data collection platform,” in *International Conference on Mobile Data Management*, 2008, pp. 221–222.

[11] J. D. Zhang, J. Xu, and S. S. Liao, “Aggregating and sampling methods for processing gps data streams for traffic state estimation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1629–1641, 2013.

[12] Y. Chen and Y. Liu, “A new method for gps-based urban vehicle tracking using pareto frontier and fuzzy comprehensive judgment,” in *IEEE International Geoscience & Remote Sensing Symposium, IGARSS 2007, July 23-28, 2007, Barcelona, Spain, Proceedings*, 2007, pp. 683–686.

[13] Q. Zhao, Q. Kong, Y. Xia, Y. Liu, Q. Zhao, Q. Kong, Y. Xia, and Y. Liu, “An improved method for estimating urban traffic state via probe vehicle tracking,” in *łLD*, 2011, pp. 5586–5590.

[14] Q. J. Kong, Q. Zhao, C. Wei, and Y. Liu, “Efficient traffic state estimation for large-scale urban road networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 398–407, 2013.

[15] H. Yue, L. R. Rilett, and P. Z. Revesz, “Spatio-temporal traffic video data archiving and retrieval system,” *GeoInformatica*, vol. 20, no. 1, pp. 59–94, 2016.

[16] H. Yue and P. Z. Revesz, “Tvics: An efficient traffic video information converting system,” in *International Symposium on Temporal Representation and Reasoning*, 2012, pp. 141–148.

[17] W. Shi, Q. J. Kong, and Y. Liu, “A gps/gis integrated system for urban traffic flow analysis,” in *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, 2008, pp. 844–849.

[18] Q. J. Kong, Z. Li, Y. Chen, and Y. Liu, “An approach to urban traffic state estimation by fusing multisource information,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 499–511, 2009.

[19] Q. J. Kong, Y. Chen, and Y. Liu, “A fusion-based system for road-network traffic state surveillance: A case study of shanghai,” *IEEE Intelligent Transportation Systems Magazine*, vol. 1, no. 1, pp. 37–42, 2009.

[20] Q. Zhao, Q. J. Kong, Y. Shen, and Y. Xia, “A traffic state estimation system using accurate digital map information,” in *Chinese Control and Decision Conference(ccdc*, 2011, pp. 1921 – 1926.

[21] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, “Estimating arterial traffic conditions using sparse probe data,” in *International IEEE Conference on Intelligent Transportation Systems*, 2010, pp. 929–936.

[22] A. Hadachi, C. Lecomte, S. Mousset, and A. Bensrhair, “An application of the sequential monte carlo to increase the accuracy of travel time estimation in urban areas,” 2011, pp. 157–162.

[23] A. Tabibiazar and O. Basir, “Kernel-based modeling and optimization for density estimation in transportation systems using floating car data,” in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, 2011, pp. 576–581.

[24] L. Li, X. Chen, and L. Zhang, “Multimodel ensemble for freeway traffic state estimations,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1323–1336, 2014.

[25] P. Pongpaibool, P. Tangamchit, and K. Noodwong, “Evaluation of road traffic congestion using fuzzy techniques,” in *TENCON 2007 - 2007 IEEE Region 10 Conference*, 2007, pp. 1–4.

[26] J. Yuan, Y. Zheng, C. Zhang, X. Xie, and G. Z. Sun, “An interactive-voting based map matching algorithm,” in *Eleventh International Conference on Mobile Data Management, MDM 2010, Kanas City, Missouri, USA, 23-26 May 2010*, 2010, pp. 43–52.

Yongjian Yang received his B.E. degree in automatization from Jilin University of Technology, Changchun, Jilin, China, in 1983; and M.E. degree in Computer Communication from Beijing University of Post and Telecommunications, Beijing, China, in 1991; and his Ph.D. in Software and theory of Computer from Jilin University, Changchun, Jilin, China, in 2005. He is currently a professor and a PhD supervisor at Jilin University, the Vice Dean of Software College of Jilin University, also Director of Key lab under the Ministry of Information Industry, Standing Director of Communication Academy, member of the Computer Science Academy of Jilin Province. His research interests include: Theory and software technology of network intelligence management; Key technology research of wireless mobile communication and services; research and exploiture for next generation services foundation and key productions on wireless mobile communication. He participated 3 projects of NSFC, 863 and funded by National Education Ministry for Doctoral Base Foundation. He has authored 12 projects of NSFC, key projects of Ministry of Information Industry, Middle and Young Science and Technology Developing Funds, Jilin provincial programs, ShenZhen, ZhuHai, and Changchun.

Yuanbo Xu received his B.E. degree and M.E. degree in the College of Computer Science and Technology, Jilin University, Changchun, China. He is pursuing his Ph.D. degree in Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun, China. His research interests include applications of data mining, recommender systems, and mobile computing.

Jiayu Han received the B.E. degree and M.E. degree in the College of Computer Science and Technology, Jilin University, Changchun, China. She is pursuing the Ph.D. degree in Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun, China. Her research interests include data mining, data fusion, recommender systems and machine learning.

En Wang received his B.E. degree in software engineering from Jilin University, Changchun, in 2011, his M.E. degree in computer science and technology from Jilin University, Changchun, in 2013, and his Ph.D. in computer science and technology from Jilin University, Changchun, in 2016. He is currently a lecturer in the Department of Computer Science and Technology at Jilin University, Changchun. He is also a visiting scholar in the Department of Computer and Information Sciences at Temple University in Philadelphia. His current research focuses on the efficient utilization of network resources, scheduling and drop strategy in terms of buffer-management, energy-efficient communication between human-carried devices, and mobile crowdsensing.

Weitong Chen received the B.E. degree in Information System form Griffith University, M.E. degree in Computer Scientist from the University of Queensland, Australia. He is pursuing his Ph.D. degree in The University of Queensland with scholarship fund in Australia. He currently is now a research officer in The University of Queensland in Brisbane, Queensland, Australia. His major areas of research interests and expertise include: Biomedical applications and Big data analytic Data Mining, Social Computing.

Lin Yue received the B.E. degree and M.E. degree in data mining from the College of Computer Science and Technology, Northeast Normal University, Changchun, China. She is pursuing the Ph.D. degree in Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, Jilin University, Changchun, China. She has been awarded a scholarship under the State Scholarship Fund to pursue her study in the University of Queensland, Australia, as a joint PhD. Student from 2014 to 2016. She is now a lecture at Northeast Normal University. Her research interests include data mining, natural language processing and machine learning..