Mapping dynamic traffic information with cellular network data by FOSS4G

Kemin Zhu,
zerokam@163.com
Junli Liu,
gisjunli@qq.com
Xianfeng Song
xfsong@ucas.ac.cn
College of Resources and Environment, Chinese Academy of Sciences, Beijing 100049, China.
Jiang Xu
xujiang@homcom.cn
Beijing Homcom Technology Co., Ltd., Beijing 100086, China

Abstract

In recent years, intelligent transportation systems (ITS) have become the central strategy for varies of traffic application in many cities. Providing road-based traffic information such as traffic speed to travelers is one of the primary goals of any ITS. With comprehensive spatial–temporal coverage of the transportation network and certain penetration rate in the population, aggregated geospatial big data such as cellular data have been used to quantitatively estimate traffic speed on urban scale. Nevertheless, it is still a challenge to estimate traffic speed at fine spatiotemporal scales over a large geographical space, mainly due to the spatial inaccuracy and temporal sampling sparsity of cellular phone data. In this article, we propose a method to estimate traffic speed estimation on arterial roads by examining the time-series individual trajectories, which were reconstructed from cellular phone data. We first identified target individuals (bus divers /conductors) from large collections of cellular datasets by applying an LCSS-SVM filter and then reconstructed time-series target individual trajectories by using Compass Search (CS) -- an evolutionary algorithm. Finally, we estimated traffic speed based on reconstructed trajectories and validated results with a grounded GPS dataset. In comparison with manual interpreted samples dataset, and the accuracy of identifying drivers/conductors has a F1 score of 88.5% and 85.1% for bus vehicles. Meanwhile, a ground test shows that estimation error is 144.5m, which is fairly accurate. the target vehicles have an average speed of 33.40km/h, speed accounted for 30.79km/h and 30.60km/h at morning rush-hour and evening rush hour respectively, shows a dynamic patterns of traffic situation in the study area.

Keywords: cellular network data; vehicle trajectory reconstruction; longest common subsequence problem; dynamic programming, kernel density estimator

1. Introduction

In recent years, intelligent transportation systems (ITS) have become the central strategy for varies of traffic application in many cities. One of the primary goals of ITS is to estimate traffic speed accurately, traditional traffic travel speed estimates for road links are based primarily on various types of sensors embedded in the pavement. Nowadays, with increasing prevalence of mobile devices, mobile phone spatiotemporal dataset is becoming a new kind of data source for research in spatiotemporal data mining and population mobility patterns, enabling large-scale analysis and applications. Development of the cellular positioning technology and the popularization of the mobile phone, providing a chance to track the drivers equipped with wireless phone as traffic probes. In this study, we aim to estimate travel speed on road links based on mobile phone dataset. Dataset used in this study consisting of anonymous mobile phone signalling data, which calculating the mobile phone signalling data into approximate locations over time for devices. The dataset contains location estimations for more than ten million devices for 24 hours in Beijing.

The objective of this study is to establish a method to identify bus vehicles from large collections of mobile phone data and use them as probe vehicles to estimate travel speed on road links. Firstly, in order to recognize record of bus driver from dataset, a dynamic programming algorithm was carried out to find a longest common subsequence between individuals' trajectories and fixed bus traveling routes. Absorbing ideas from string matching and taking unique characteristics of movement of buses into account as well, a set of indexes are applied to evaluate the similarity between bus traveling routes and individuals' trajectories and then the devices carried by bus drivers are identified accurately.

Considering that mobile phone dataset used in this study records continuous 24-hours duration of approximate user position, it is necessary to determine the time sections when buses move normally. Status of bus drivers can be classified into 4 types: before departs, after shutdown, moving on road links and parking in bus station. Thus, to determine the status of bus drivers at particular time kernel density estimators were exploited to find dwell time of individuals' trajectories and further determine whether bus driver individuals' moving throughout the route.

Another challenge is that information on individuals' trajectories cannot be obtained directly from dataset, the accurate position of user must be inferred, to eliminate the error between trajectory of real human and the one obtained through cellular phone dataset, a heuristic global optimization method is applied. based on traditional assumption that motion of buses tends to be near-uniform on road links, defining standard deviation of bus speed of its course as optimum objective function, the precise locations are searched as function variables, a single-objective constrained optimization problem is set. The compass search algorithm is used to solve the problem and meta-algorithm of MBH is applied to avoid local optimization.

The traffic travel speed estimating method is applied to calculate travel speed of 20 bus route in Huilongguan community – one of the most densely populated and representative in Beijing. As mentioned above, mobile phone dataset used in this study lack of actual location of devices, hence, to verify the practicability of this method, we develop a algorithm to generate a simulated cellular phone dataset with the ground truth information collected by a mobile phone application. Through comparisons between vehicle estimate location and actual location, original data sampling every 300 seconds on average, the error between base location and actual location vary from 500m to 2000m, the travel speed estimates method yielded an average error of 144.5 m, the target vehicles have an average speed of 33.40km/h, speed accounted for 30.79km/h and 30.60km/h at morning

rush-hour and evening rush hour respectively, We find the following characteristics: firstly, traffic pressure rise at 7:00-9:00 and 17:00-19:00, when the average traffic speed is significantly lower than other parts of the day. Then, the average traffic speed on the Badaling expressway is significantly higher than on the major roads of city. The result shows that travel speed estimates algorithm is effective. The present research can be applied to identify bus driver from massive spatiotemporal data, offer an accurate location of buses and estimate traffic travel speed on road links.

2. Literature review

Currently, traffic speed estimation is mostly based on information provided by loop detectors and cameras which can offer velocity, density, and flow at specific locations. In order to collect traffic conditions, many special sensors on the road have been deployed in recent years. Some kinds of new technologies capable of registering vehicle trajectories, such as triangulation of GPS-enabled mobile phones or cell phone signals, provide an opportunity to collect valuable real-time information for traffic situation. This type of data sources can provide data such as vehicle position, speed and acceleration (Herrera and Bayen, 2010). While vehicles such as taxis, buses and trucks equipped with Global Positioning System (GPS) are used to collect traffic data, obtaining reliable photos of traffic conditions on the highway remains a problem, as most private cars may be reluctant to share them. Private location information (Florida Department of Transportation, 2007; Guillaume Leduc 2008; Schneider et al. 2005). Thus, based on the current development trend of mobile technology, cellular phone data sources have broad prospects.

Previous research has explored the use of mobile sensors, such as GPS-supported mobile phones, to estimate traffic speed on highways (SaWal and Walland, 1995; Westerman et al). 1996, Ygnace et al. 2000, Bar Gera, 2007, Klaus et al., Herrera & Bayenra, 2010 et al., 2010), and low permeability model performance under varying degrees of congestion. The traffic speed of the arteries is not so deep, and the presence of traffic lights increases the complexity of the analysis.

The use of mobile phones as detectors to track the trajectory of vehicles and thus obtain real-time traffic has become an effective way (Kalabres et al., 2011; Lin et al., 2011). Due to its low cost and wide distribution, traffic condition estimation based on mobile phone signal data has become an alternative and feasible method (Bar-Gera, 2007; Wunnava et al., 2007). This technology is much cheaper than traditional data collection approach because it does not require a dedicated infrastructure and able to meets the real-time needs of UTMS1 (Wang et al., 2007). It has the potential to performance well in time and space coverage for transportation networks and useful data, and has a certain penetration rate among the population (Herrera et al.). In addition, mobile technology can easily provide measurement of travel time, or delay along the signal artery, and the sensor is not suitable.

The study on the use of motion sensors to estimate the traffic speed of a trunk road suggests estimating the speed of the connection by the average instantaneous speed or the average speed of the corresponding link (Cheu et al.). Zhang et al. (2007). After 5% of the estimated value, they achieved good results with a 94% penetration rate. Tao et al. (2012) proposed a similar model and reconstructed the trajectory of GPS data using map matching, Kalman filtering and data filtering, with an average error of six. A speed of 7% is estimated to be 10% penetration. Other studies used

methods based on Markov chain, logistic regression, and STARMA models (hunters, etc.). , 2009; squid, et al. 2010; Feng et al., 2011 (Ramezani and Geroliminis, 2012), obtained better results than the RST method. Especially Feng et al. (2012), estimated travel time is less than 2%. However, as the authors describe, the model does not work in the case of oversaturation.

Due to its low cost and wide distribution, traffic condition estimation based on mobile phone signal data has become an alternative and feasible method (Bar-Gera, 2007; Wunnava et al., 2007). The positioning technology of mobile phone signal data is usually better than that provided by GPS. Lower (Wunnava et al., 2007; Quddus et al. 2007), but there are a large number of mobile phones in the wide area, which will provide location data that makes positioning technology very attractive (Calabres et al., 2011). However, accuracy is still a huge challenge in estimating the speed of mobile phone signal data transmission.

In this study, we aim to estimate travel speed on road links based on mobile phone dataset. Dataset used in this study consisting of anonymous mobile phone signaling data, which calculating the mobile phone signaling data into approximate locations over time for devices. The dataset contains location estimations for more than ten million devices for 24 hours in Beijing

3. Method and Materials **(750 – 1500 WORDS)**

3.1. Study area and datasets

To test the above proposed approach, Huilongguan town, one of the largest townships in the northern part of Beijing, China, was selected to estimate traffic speeds of nearby roads. The town has a total population of about 450,000 people and an area of 34.5 square kilometers. Being one of most populated residential areas in Beijing, choked traffic has been an archenemy to urban transportation system in this area. The town streets and roads maps were sourced from OpenStreetMap and total 20 bus routes and related more than 400 bus stops covering this area were also retrieved from Beijing Public Transport Corporation (BPTC)

Cellular phone network datasets, covering the town extension area on early August, 2016, were collected, totally 3.6 billion 4G signaling records among 10 million mobile phone users with an average phone-station interaction interval of 280 seconds. Each record includes timestamp (TS), International Mobile Equipment Identity (IMEI), Tracking Area Code (TAC) and Cell Identity (CI), representing an interaction event between a mobile phone (IMEI) and a base station (TAC plus CI) at a dedicated time (TS). In most cases these trajectories are quite coarse with sparse sample points and followed a serious deviation from real pathways.

The field work of tracking bus movement along bus routes were carried out. Bus trajectories were recorded with a sampling interval of 1~2 seconds and a GPS positioning error of 5~10 meters through by a modified version of mobile GIS App (Geopaparazzi). Meanwhile, the interaction events between on-bus mobile phones and base stations along bus routes were also retrieved. These ground truth datasets will be used for model calibration and validation.

3.2. Identification of mobile phones carriers and target vehicles

Urban bus system is the major part of public transportation systems, which can cover the whole city. It could be a representative of the state of urban traffic. Thus, in order to monitor traffic

conditions as accurately and timely as possible, buses that run normally on a fixed traveling routes are used as probe vehicles to infer SMS of link road segmentally.

3.2.1. Matching a mobile phone carriers' trajectory to a specific bus route

Absorbing ideas from string matching, we modified the algorithm of longest common subsequence (LCS) to find bus drivers or conductors from the vast number of anonymous cell phone carriers based on their periodicity of cycling around a fixed route.

Let two cellular data trajectory sequences for the dimension of m and n respectively be defined as follows:

$$TR_A = ((a_{x,1}, a_{y,1}), \dots, (a_{x,n}, a_{y,n}))$$
$$TR_B = ((b_{x,1}, b_{y,1}), \dots, (b_{x,m}, b_{y,m}))$$

$a_{x,1}, a_{y,1}$ stand for the geographic coordinate of the ith point of trajectory sequences TR_A, let $HEAD(TR_A)$ be the subsequence consist of first $n-1$ elements of TR_A, which can be expressed as:

$$HEAD(TR_A) = ((a_{x,1}, a_{y,1}), \dots, (a_{x,n-1}, a_{y,n-1}))$$

The definition of LCS is given as:

$$LCSS_{periodic}(TR_A, TR_B) = \begin{cases} 0 & \text{if } (TR_A \text{ or } TR_B \text{ is Empty}) \\ 1 + LCSS(Head(TR_A), Head(TR_B)) & \text{if } (match(a_i, b_{j \bmod qlen}) = True) \\ \max[\, LCSS(Head(TR_A), TR_B), LCSS(TR_A, Head(TR_B))] & otherwise \end{cases}$$

Where $i = 1,2, \dots, n$; $j = 1,2, \dots,$ m

Considering that calculating LCS is a recursive problem, to compute the LCS efficiently, we applied a dynamic programming solution. start by constructing a two-dimensional array C in which partial results can be built up. List one of the sequences across the top and the other down the left, use $C[i,j]$ to record the length of LCS between x_i and y_i. Carry out the recursive calculation from bottom up. At this point, we can calculate it based on the relationship between x_i and y_i. Each $C[i,j]$ simultaneously records the last point of the match, which can be traced back to $C[i,j]$ after completion of the calculation.

the solution can be expressed as a recursive function as followed:

$$C[i,j] = \begin{cases} 0 & if(i = 0 \; or \; j = 0) \\ C[i-1, j-1] + 1 & if(i, j > 0 \; and \; x_i = y_i) \\ \max\{C[i-1, j], C[i, j-1]\} & otherwise \end{cases}$$

3.2.2. Identifying drivers and conductors by SVM

Base on the LCS results, a set of indices for a mobile phone carrier are defined to describe the degree of the match between his coarse trajectory and the related bus route, including periodicity, similarities and completeness. The SVM classification model is established by using the above indices as feature vectors and manual interpretation training dataset with category marks (whether is a bus driver/conductor).

The optimization hyperplane is to maximize the distance among the boundaries of different types. It can be defined as:

$$wx + b = 0$$

Where, w is a vector which is perpendicular to the hyperplane and b is a constant term. The samples are defined as vectors (x_i, x_j), $i = 1, \dots m.$ x is a multi-dimensional vector, which

represents the explanatory variables.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

Subject to $y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0, i = 1, \ldots, m$

Its dual form is

$$\min_{\alpha} \frac{1}{2} \alpha^T y_i y_j K(x_i, x_j) \alpha - e^T \alpha$$

$$y^T \alpha = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

Where, y reflects sample classification. $\xi_i \geq 0$ is a slack variable on behalf of the degree of error classification. α_i is Lagrange coefficient. e is a vector of all ones. $C > 0$ is defined as penalty parameter of error classification. $K(x_i, x_j)$ is the kernel function, which can transform the vectors in the low dimensional feature space into a high one.

A Radial Basis Function (RBF) SVM classification model in scikit-learn--a free software machine learning library for the Python programming language is used. The whole process of the model is realized using the Python2.7 programming.

3.3. Position estimation of target vehicle

3.3.1. Determining start-point and end-point of target vehicle trips

To estimate accurate location of target vehicle, it is necessary to tell whether target vehicle is sedentary or active, or, more specifically, to determine dwellings at both start-point and end-point of a bus route from their cycling trajectories. In order to judge the dynamic state of vehicles over time, kernel density estimator was applied to detect vehicle dynamic state changing event. Combining spatial approximate, we can furtherly detect the time period when the vehicle was parking in stations.

We firstly calculated the distance between a reference location (here we take coordinate of the departure station as reference location) and base station that connected with target devices from time to time, which can be regard as a finite data sample. We calculated kernel density of event record by cellular devices according to formula given below.

$$\hat{f}_h(x) = \frac{1}{n} \sum_{n}^{1} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{(x - x_i)}{h})$$

Where, K is the kernel with bandwidth h, a non-negative function that integrates to one and is also called as the scaled kernel and defined as $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$. xi denotes the distance of the *i*th point (base station tower) to the reference point (bus termini).

3.3.2. Estimating vehicle position by heuristic global optimization

Based on traditional assumption that motion of buses tends to be near-uniform on road links, we defined standard deviation of bus speed of its course as optimum objective function to evaluate the smoothness of locomotion when the vehicles running on a link-road, which can be descript by the formula below. The precise locations are searched as function variables in search space of coverage segment, with a number of constraints on keeping spatial-temporal

sequences on roads and leveraging dwelling times on bus stops and cross-roads, a single-objective constrained optimization problem is set.

$$\text{obj} = \text{minimize: } \sum_{i}^{n-1}\left[\frac{m_{i+1}-m_i}{t_{i+1}-t_i} - \frac{1}{n-1}\sum_{i}^{n-1}\left(\frac{m_{i+1}-m_i}{t_{i+1}-t_i}\right)\right]^2$$

$$\text{subject to: } m_i < m_{i+1}$$
$$lb_i < m < rb_i$$

Where, m_i is the accumulative mileage that target vehicle drive along bus route, t_i is the timestamp of ith interaction records, n is the numbers of total interaction records. Objective of optimization is to minimize obj under the given constraints: 1) the accumulative mileage m_i of following point should be greater than preceding point m_{i+1}. 2) search space of optimization was restricted to position nearby LCS result, which has left boundary lb_i and right boundary rb_i.

To solve the global optimization model and approach the accurate location of vehicles, the Compass Search Solver which is especially for constrained stochastic single-object algorithm was applied. After specified numbers of iterations or the optimum objective function reach given threshold, the iterations were completed and the model gave the adjust location of vehicles as an output. The compass search algorithm is used to solve the problem and meta-algorithm of MBH is applied to avoid local optimization.

3.4. Calculating space mean speed of link road

Space mean speed is measured over the whole roadway segment. Consecutive records of a roadway segment track the speed of individual vehicles, and then the average speed is calculated. It is considered more accurate than the time mean speed. In this study, we use estimated spatial-time location of target vehicle to calculate space mean speed.

We first partitioning bus route into a set of segments uniformly (50m of each segment length). Since that estimated spatial-time position of target vehicle is already calculated in section 3.3, the speed of each target vehicle can be inferred easily. For each segment, the space mean speed was calculated by estimated target vehicle speed as followed:

$$\boldsymbol{v_s} = \left((\mathbf{1}/\boldsymbol{n})\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\mathbf{1}/\boldsymbol{v_i})\right)^{-\mathbf{1}}$$

where n represents the number of target vehicles passing the roadway segment. The space mean speed is thus the harmonic mean of the speeds. The time mean speed is never less than space mean speed:

$$\boldsymbol{v_t} = \boldsymbol{v_s} + \frac{\boldsymbol{\sigma_s}^2}{\boldsymbol{v_s}}$$

Where $\boldsymbol{\sigma_s}^2$ is the variance of the space mean speed.

After taking whole set of target vehicle into space mean speed calculating, the traffic speed of link road was estimated segmentally.

4. Results and findings (1000 – 1500 WORDS)

In this section, we validate the result of target vehicle location estimation with both oversampling GPS data and base station record data collected simultaneously by a mobile phone application. First, a resampling process was applied on collected GPS data to generate simulative data that has same sampling distribution characteristic of cellular data. Then, generated simulative data was used to verify the accuracy of departure and arrival time estimation and spatial-time location estimation respectively.

4.1. Cross validation to evaluate the accuracy of identification of target vehicle

The identification of target vehicle is fundamental to speed estimation. Thus, it needs to be evaluated for accuracy and reliability. Through the establishment of the radial basis kernel function of SVM classifier, set up the training set, then trained discriminant model and identify the mobile phone signal data belong to target vehicles. We select 20 bus line inside research area and established a training set for SVM classifier, the training set collected 253 samples of cellular phone devices. result of cross inspection of each bus line is showed in Table 1 below, the overall recall rate of whole 20 bus line reaches 83.95% and precision rate reaches 93.75%.

Table 1. Validation on identification of driver and conductor (of a public vehicle)

Bus route	Precision	Recall	F1-Score
307	0.913	0.954	0.933
344	0.957	0.937	0.947
407	0.916	0.846	0.880
428	0.956	0.916	0.936
…	…	…	…
overall	0.938	0.840	0.885

The result showed in Table 1. is focus on testing identification model towards cellular phone carriers. To offer the precision evaluation of target vehicle identification, we applied an Fréchet distance-based clustering algorithm to clustered cellular phone carriers on same vehicle into groups as target vehicles, then we verified the clustered vehicle identification result with similar method. Result of cross inspection of each bus line is showed in Table 2. below, the overall recall rate of whole 20 bus line reaches 83.24% and precision rate reaches 81.21%.

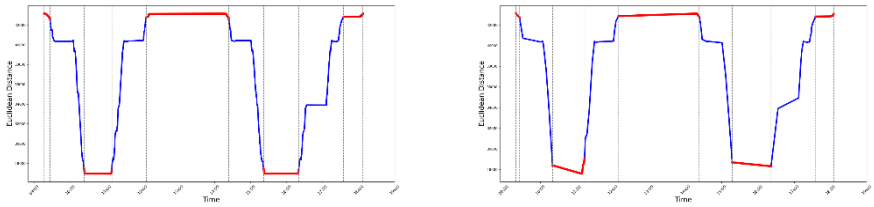
Table 2. Validation on identification of target vehicle

Bus route	Precision	Recall	F1-Score
307	0.873	0.843	0.858
344	0.896	0.901	0.898
407	0.854	0.831	0.842
428	0.857	0.863	0.860
…	…	…	…
overall	0.832	0.872	0.851

4.2. Evaluation of target vehicle location estimation

(1) Departure and arrival time estimation

Curve in fig. 1(a) and fig. 1(b) show the Euclidean distance between target vehicle and origin station changes over time of oversampled GPS trajectory and resampled simulative cellular data trajectory. The start-point and end-point of each trip were detected by a kernel density estimator described in section 3.3.1, and time period when target vehicle is active or sedentary are marked in red and blue respectively. Although it is obvious that resampled trajectory lost certain detail characteristic and distort slightly, the estimator provides a relatively accurate estimate of start-point and end-point.



(a) Estimation of GPS data **(b) Estimation of resampled data**

Fig. 1. Departure and arrival time estimation

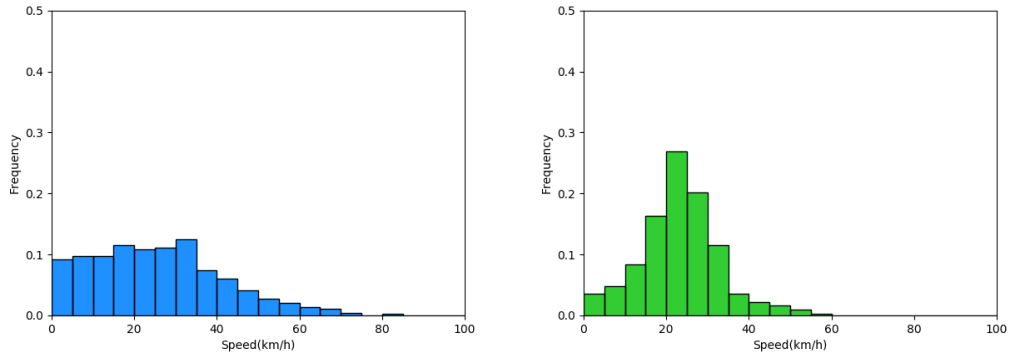
To examine start-point and end-point estimation quantitatively, comparison between grounded and estimated departure-time/ arrival-time was showed in Table 3. The average estimation error of departure-time is 29.75 seconds, which of arrival-time is 28.5 seconds. Proposed departure and arrival time estimation model performance well for its accuracy.

Table 3. Grounded and estimated Departure-time/ Arrival-time

Trips	GPS Departure-time	Estimated Departure-time	Departure-time Estimation error	GPS Arrival-time	Estimated Arrival-time	Arrival-time Estimation error
Trip 1	9:59:10	9:59:12	2	10:28:05	10:28:48	43
Trip 2	11:05:38	11:04:41	57	11:38:01	11:37:48	13
Trip 3	15:01:10	15:01:22	12	15:26:16	15:26:34	18
Trip 4	16:45:08	16:45:56	48	17:14:18	17:13:38	40

(2) Spatial-time location estimation

Fig.2(a) and fig.2(b) show the distribution proportion of target vehicle speed on the link road before and after the heuristic global optimization. The contrast between the two figures show that target vehicle speeds tend to be smooth after optimization, and unrealistic traffic speed values that higher or lower than the reasonable range caused by inherent inaccuracy and Sparse Sampling of cellular phone data are reduced. e.g., optimization model performance to minimizing deviations of traffic speed values' trend in adjacent time intervals.



(a) **(b)**

Fig. 2. distribution proportion of target vehicle speed
(a) before the heuristic global optimization (b) after the heuristic global optimization

We then analysis the spatial deviation of estimated position of target vehicles. An MBH-compass search algorithm was applied to solve the vehicle position-time position optimization problem and offer estimation of vehicles' spatial-time location. The partial output of optimize model was listed in Table 4. By interpolating spatial-time location estimation result, we made a visualized color strip of vehicle location change over time in fig.3, two color strips represent accurate position and estimated position of target vehicle shows high consistency.

Table 4. Spatial-time location estimation result

Time	Mileage	X coordinate	Y coordinate	Time	Mileage	X coordinate	Y coordinate
10:02:26	36118.53	12950904	4864737	10:17:00	38918.11	12952997	4865498
10:03:23	36407.21	12951192	4864749	10:19:18	39370.23	12952998	4865932
10:07:27	37221.02	12952005	4864780	10:23:08	39963.69	12953591	4865954
10:10:30	37779.06	12952563	4864799	10:23:25	40100.62	12953534	4866067
10:12:13	38164.97	12952949	4864815	…	…	…	…

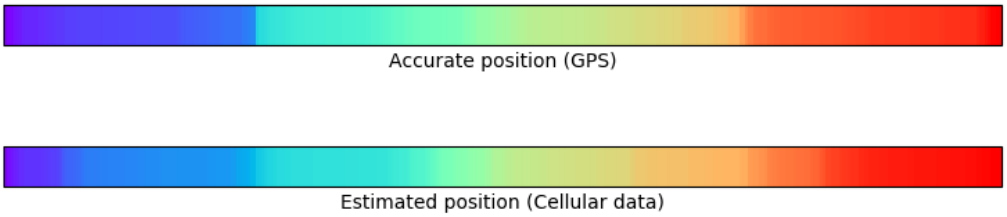


Fig. 3. accurate position and estimated position of target vehicle change over time

Error between estimated position infer by cellular tower location and ground position collected by GPS before and after the global optimization are showed in Fig.4 The bus trajectories represented with cellular tower positions are deviated averagely 318.4m from their real trajectories, however, the reconstructed spatial-temporal bus trajectories reduced error significantly, 144.5m, after global optimization. Thus, we can state that our approach performs well minimizing deviations of traffic speed values' trend in adjacent time intervals and offer an accurate position of target vehicles.

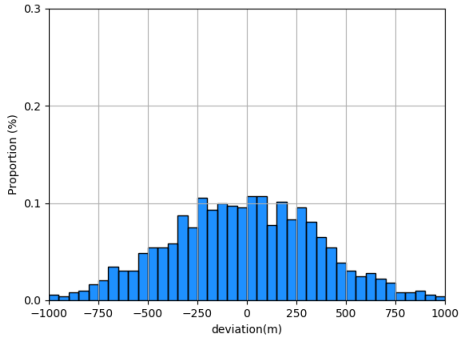


Fig. 4. (a)

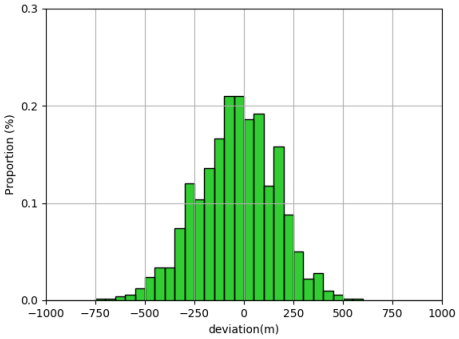


Fig. 4. (b)

4.3. Result of vehicle position and traffic speed estimation

Finally, we calculated the statistical results of road traffic speed hourly of among study area, and a traffic speed map was produced to illustrate traffic status of major roads at Huilongguan town as shown in Fig.5 By statistics, the target vehicles have an average speed of 33.40km/h, speed accounted for 30.79km/h and 30.60km/h at morning rush-hour and evening rush hour respectively, We find the following characteristics: firstly, traffic pressure rise at 7:00-9:00 and 17:00-19:00, when the average traffic speed is significantly lower than other parts of the day. Then, the average traffic speed on the Badaling expressway is significantly higher than on the major roads of city.

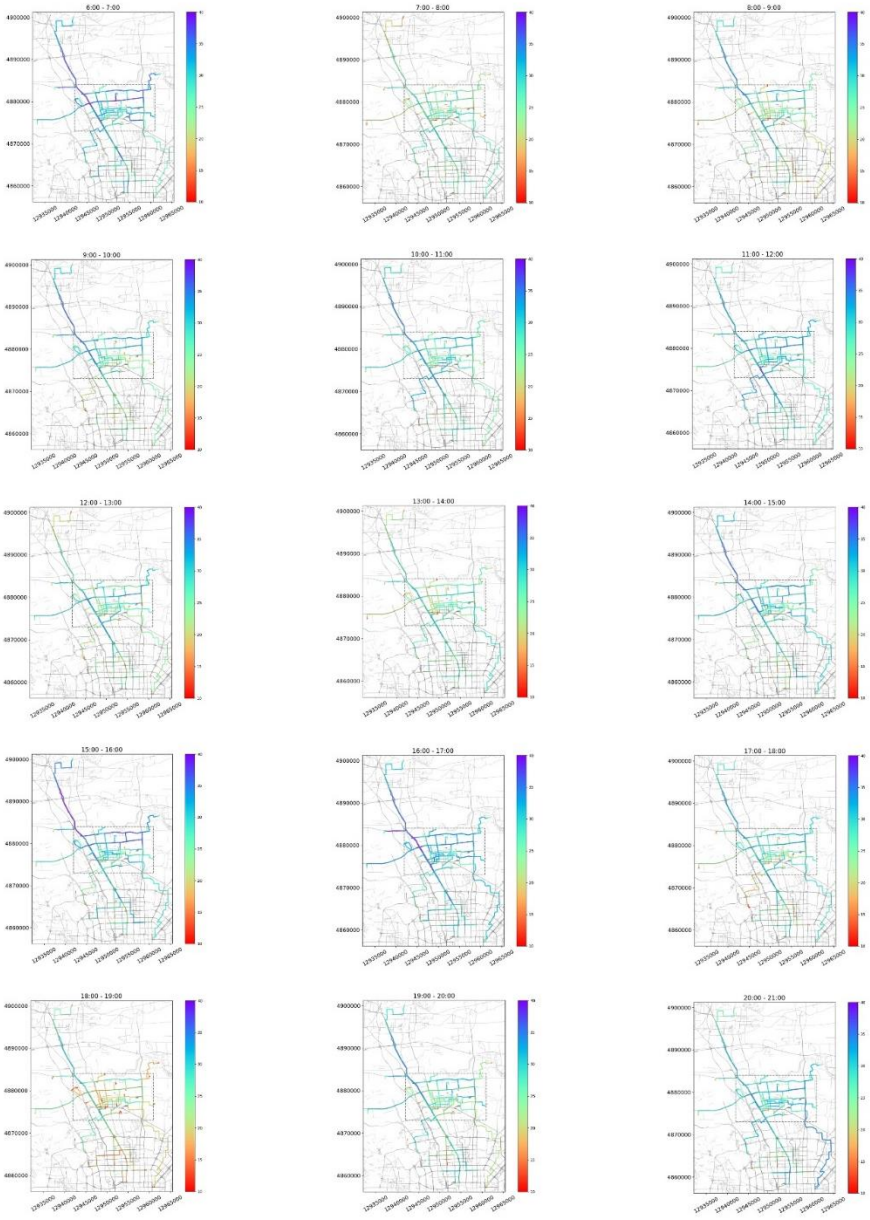


Fig. 5. traffic speed map at Huilongguan town

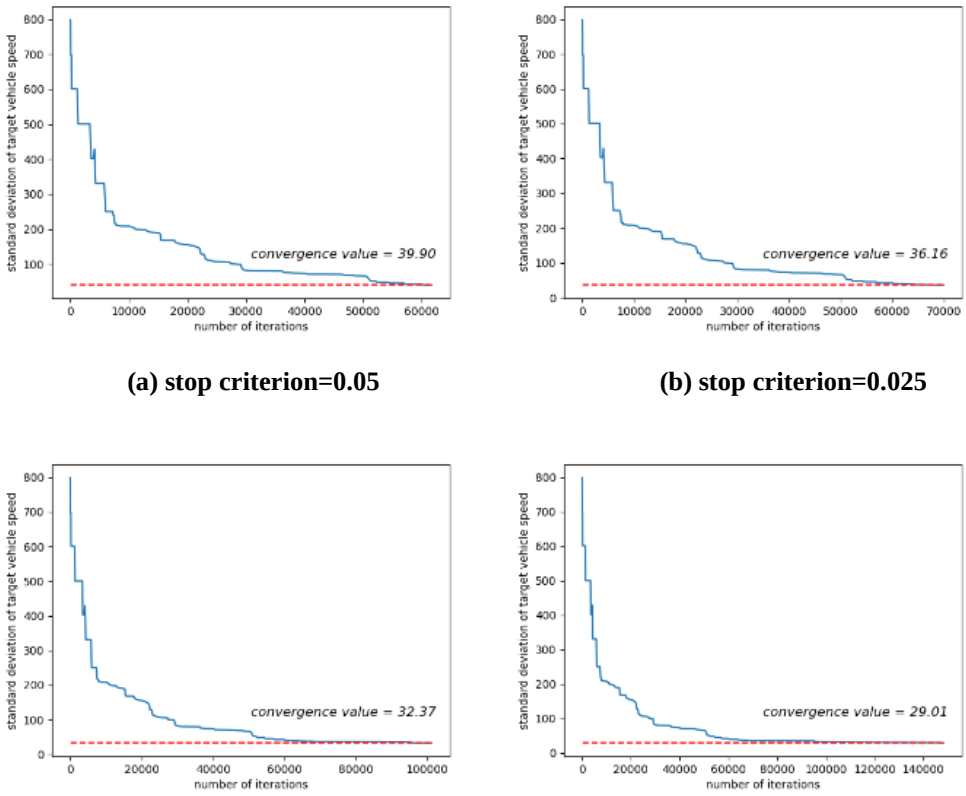
5. Discussion & conclusions

The method proposed in this article provide a new approach to estimate traffic speed based on inaccurate and sparse sampled cellular data, and the effectiveness of the method was validated in section 4. However, it should be note that this study examined only by simulative data generated from GPS data that has characteristics of sampling. Thus, in this section, following topics about key parameter and model performance will be discussed:

5.1. Stopping criterion for iteration of heuristic global optimization

In section 3.3.2, we applied MBH-compass search algorithm is used to solve the vehicle position-time position optimization problem. MBH-compass search algorithm is an evolutionary algorithm which approach global optimal solution by iteration. There are three key parameters of

MBH-compass search algorithm: objective function, constraints and stopping criterion, the former two parameters were already discussed in section 3.3.2. Thus, we test proposed method by calculating number of iterations, model running time, objective function value and estimation error with different stopping criterion.



(a) stop criterion=0.05 **(b) stop criterion=0.025**

(c) stop criterion=0.0.0125 **(d)stop criterion=0.00625**

Fig. 6. Objective function values change with iteration

Objective function values change with iteration of MBH-compass search algorithm shows in fig.6(a)-fig.6(d) with a stop criterion of 0.005, 0.025, 0.0125 and 0.00625 respectively. With each stop range, objective function value descends in steps with increase of iteration times, and finally converged to different objective function values. Detailed information includes maximum number of iteration, convergence value, model running time and position estimation error shows in Table 5. From the Table 5, one can tell that MBH-compass search algorithm can effectively reduce position estimation errors along with the increasing iterative times, however, the time cost for model running increase as well. When stop range varies from 0.0125 to 0.00625, the estimation error shows no significant differences, which indicated that estimation accuracy cannot obviously be improved while running time increased about 50% (more than 400 seconds). Thus, we gave 0.0125 as a suggested value of stopping criterion for MBH-compass search algorithm to balance model time cost and accuracy requirement.

Table 5. Optimization result with different stop criterion

Stop criterion	Number of iterations	Convergence value	Running time(s)	Estimation error(m)
0.1	202	696.85	13.9	476.5

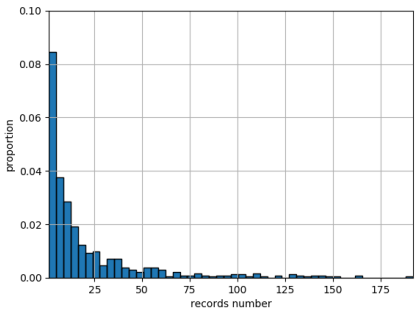
0.05	61687	39.90	317.6	229.2
0.025	69979	36.16	454.2	187.4
0.0125	101193	32.37	814.0	141.8
0.00625	148321	29.01	1287.2	138.3

5.2. Impact of different sampling rate on spatial and temporal location estimation

Target vehicle recognition and accurate position estimation in this study is based on cellular station transition over time, sampling rate of cellular data is definitely one of key factors of speed estimation. Considering that sampling rate of cellular data collected from different data source varies significantly, it is necessary to discuss applicability of this method with data of sampling rate.

To test model performance at different data sampling rate, a set of resampled GPS data with varying sampling intervals from 1min to 30min was generated as simulative data. we measure the accuracy of position estimation of each resampled GPS data and presented following result.

For original oversampled GPS data, there are total 640 cellular base station used during recording. As the GPS data statistic records numbers of each cellular base station show in fig.7(a) and dwell time of each cellular base station show in fig.7(b), more than 10 percent of cellular base station only has one connection with mobile phone and 50 percent of cellular base station has less than 9 connections with mobile phone. The average dwell time for each cellular base station is 21.4 seconds, however, for some base station the dwell time could reach more than 2 minutes which could cause by a congestion on the link road or a cellular base station with widespread coverage.



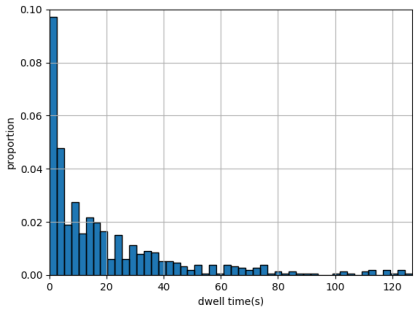


Fig. 7. (a) GPS data statistic records numbers **Fig. 7. (b) dwell time of each cellular base station**

Fig.8 shows numbers of cellular station transition (i.e. the numbers of cellular station connected) varies over sampling rate which range from 1.7 seconds (original GPS sampling rate) to 30 minutes. With decrease of sampling rate, the base station remained reduced as well, which lead to greater difficulty for target vehicle recognition and accurate position estimation. The number of cellular stations connected drop significantly when trajectory data was oversampled, that is because a high proportion of cellular station has relatively short-term connection with target mobile phone device which are highly probable left out after resampling. When the sampling interval is 280 seconds which is same as cellular phone dataset used in this study, only 83 base station remains – about 86% of base were left out after resampling.

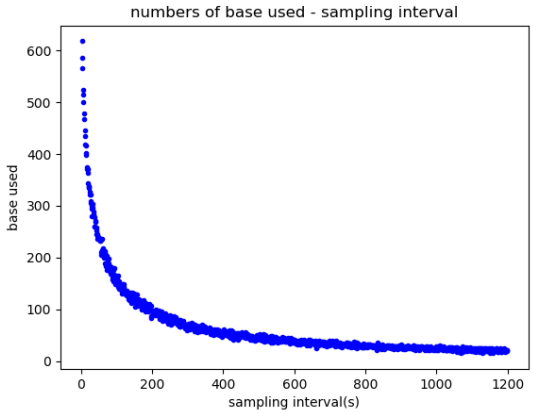


Fig. 8. Numbers of cellular station transition varies over sampling rate

To explore how sampling rate of cellular data influence vehicle position estimation, we calculate average estimation error as metric of position estimation at varying sampling intervals from 1min to 30min. The average estimation error result shows in fig.9, the position estimation error increase with lower sampling rate. When sampling interval is below 10min, the model shows a reliable performance with an estimation error less than 300m. However, when sampling interval is above 15min, the average estimation error increase sharply to more than 600m.

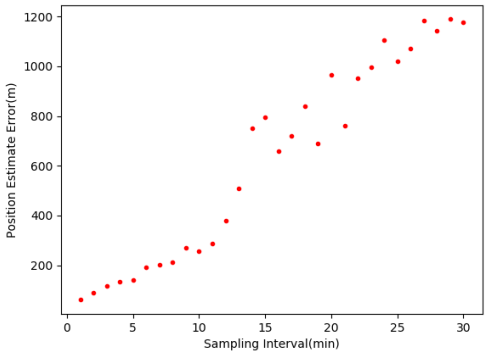


Fig. 9. Average estimation error result with different sampling rate

5.3. Conclusion

In this paper, we first modified the algorithm of longest common subsequence (LCS) to identify target vehicles from the vast number of anonymous cell phone carriers based on their periodicity of cycling around a fixed route. Then, the support vector machine (SVM) is adopted to pick up those phone carriers with a high similarity value and label them as target vehicle. A kernel density estimator was exploited to separate drivers' movement time and rest time by analyzing dwellings at both start-point and end-point of a bus route from their cycling trajectories. Finally, we established a heuristic global optimization model to approach the accurate location of vehicles when they are running normally on a link-road and assess the result by using real challenging positioning (GPS and Cellular-based) data traces.

We arrive at the conclusion that our method allows for traffic speed estimation using mobile cellular data. The experimental results validated by GPS data show that (1) Our approach performs

well in recognizing target vehicles (transit bus) from large volume mobile data. (2) It also contributes to estimate the accurate position of target vehicles that and would be valuable on the traffic speed estimation. (3) In study area of this study, traffic pressure rise at 7:00-9:00 and 17:00-19:00, when the average traffic speed is significantly lower than other parts of the day and the average traffic speed on the expressway is significantly higher than on the major roads of city.

6. Acknowledge

First and foremost, I would like to show my deepest gratitude to my supervisor, Prof. Song XianFeng, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance in every stage of the writing of this thesis. Without his enlightening instruction, impressive kindness and patience, I could not have completed my thesis. His keen and vigorous academic observation enlightens me not only in this thesis but also in my future study. I shall extend my thanks to colleague in my laboratory who give me a comfortable learning atmosphere. Without their kindness and help, this thesis could not have reached its present form. Last but not least, I want to thank all my friends, for their encouragement and support.

7. References

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study fro.m Israel. Transportation Research Part C, Emerging Technologies, 15(6), 380–391. doi:10.1016/j.trc.2007.06.003

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2008). Review of traffic data estimations extracted from cellular networks. IET Intelligent Transport Systems, 2(3), 179–192. doi:10.1049/iet-its:20080003

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-time urban monitoring using cell phones: A case study in Rome. Intelligent Transportation Systems. IEEE Transactions on, 12(1), 141–151.

Guillaume Leduc. (2008). Road traffic data: Collection methods and applications. Working Papers on Energy, Transport and Climate Change, 1, 55.

Gundlegard, D., & Karlsson, J. M. (2009, October). Route classification in travel time estimation based on cellular network signaling. In Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on (pp. 1-6). IEEE. doi:10.1109/ITSC.2009.5309692

Hansapalangkul, T., Keeratiwintakorn, P., & Pattara-Atikom, W. (2007, June). Detection and estimation of road congestion using cellular phones. In Telecommunications, 2007. ITST'07. 7th International Conference on ITS (pp. 1-4). IEEE.

Hongsakham, W., Pattara-Atikom, W., & Peachavanish, R. (2008, May). Estimating road traffic congestion from cellular handoff information using cell-based neural networks and K-means clustering. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on (Vol. 1, pp. 13-16). IEEE. doi:10.1109/ECTICON.2008.4600361

Innovative Data Collect Research Project. (2007). Update on the State of the Innovative Traffic Data Collection Industry. Florida Transportation Department.

Lin, B. Y., Chen, C. H., & Lo, C. C. (2011). A traffic information estimation model using periodic location update events from cellular network. In Intelligent Computing and Information Science (pp. 72–77). Springer Berlin Heidelberg. doi:10.1007/978-3-642-18134-4_12

Lü, W., Zhu, T., Wu, D., Dai, H., & Huang, J. (2008). A heuristic path-estimating algorithm for large-scale real-time traffic information calculating. Science in China Series E: Technological Sciences, 51(1), 165–174. doi:10.1007/s11431-008-5013-6

Mohan, P., Padmanabhan, V. N., & Ramjee, R. (2008, November). Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In Proceedings of the 6th ACM conference on Embedded network sensor systems (pp. 323-336). ACM. doi:10.1145/1460412.1460444

Narupiti, S., Wajanasathienkul, W., Rotwannasin, P., &Lertworawanich, P. (2013). Evaluation of Offline and Online Speed-based Travel Time Estimation on Expressway. Asian Transport Studies, 2(4).

Orlik, P. V., & Rappaport Stephen, S. (1998). A model for tele traffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions. Selected Areas in Communications. IEEE Journal on, 16(5), 788–803.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications, 36(2), 3336–3341. doi:10.1016/j.eswa.2008.01.039

Pattara-Atikom, W., Peachavanish, R., & Luckana, R. (2007, September). Estimating road traffic congestion using cell dwell time with simple threshold and fuzzy logic techniques. In Intelligent

Transportation Systems Conference, 2007. ITSC 2007. IEEE (pp. 956-961). IEEE. doi:10.1109/ITSC.2007.4357756

Pattaramalai, S., Aalo, V. A., & Efthymoglou, G. P. (2009). Evaluation of call performance in cellular networks with generalized cell dwell time and call-holding time distributions in the presence of channel fading. Vehicular Technology. IEEE Transactions on, 58(6), 3002–3013.

Puntumapon, K., & Pattara-Atikom, W. (2008, May). Classification of cellular phone mobility using Naive Bayes model. In Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE (pp. 3021-3025). IEEE. doi:10.1109/VETECS.2008.324

8. Authors' biography

Kemin Zhu was born in Hunan, China, in 1993, He is currently doing his PhD studies College of Resources and Environment, Chinese Academy of Sciences, Beijing. His main research field is spatio-temporal data mining.

Junli Liu was born in Shandong, China, in 1992, He is currently doing his PhD studies College of Resources and Environment, Chinese Academy of Sciences, Beijing. His main research field is spatio-temporal data mining.

Xianfeng Song was born in Hebei, China, in 1969, He received the Ph.D. degree in Institute of Geographic Sciences and Natural Resources Research , Chinese Academy of Sciences in 1998. He is currently a Professor with the Department of Resources and Environment, University of Chinese Academy of Sciences. His research topics are spatio-temporal data mining, mobile GIS/WEBGIS system, swat model, soil erosion model.