EJTIR

Issue 15(4), 2015
pp. 639-661
ISSN: 1567-7141
tlo.tbm.tudelft.nl/ejtir

Estimating traffic delays and network speeds from low-frequency GPS taxis traces for urban transport modelling

Bin Deng[1]
Martin Centre for Architectural and Urban Studies, University of Cambridge, United Kingdom.

Steve Denman[2]
Centre for Smart Infrastructure and Construction, University of Cambridge, United Kingdom.

Vassilis Zachariadis[3]
Centre for Smart Infrastructure and Construction, University of Cambridge, United Kingdom.

Ying Jin[4]
Department of Architecture, University of Cambridge, United Kingdom.

The collection of actual traffic delays and road traffic speed data is essential in modelling urban transport resource efficiency, congestion and carbon and pollutant emissions, which is in turn part of core empirical basis for evidence-based policy making for improving urban sustainability. This data collection has also been one of the most expensive and time-consuming tasks, which restricts how well and how often the models can be built and validated, often to the extent that urban transport models have to rely on severely outdated data with sparse coverage. New smart data such as GPS vehicle traces has raised the prospect of remedying the data shortage, but for operational and data protection reasons often only low-sampling frequency traces are available. This paper proposes a novel method for estimating actual, congested link-speeds from low-sampling frequency taxi GPS traces that are publicly available. The method is based on a path inference process and is applied over a detailed road network in a large city region. It shows that low frequency GPS trajectories can significantly improve the spatial and temporal resolutions of traffic speed data for transport modelling and policy analysis. This opens up the prospect of improving road operation performance, managing travel demand and optimizing urban circulation.

Keywords: *low-frequency GPS taxi traces, route choice, map-matching, congested link speed, transport modelling.*

1. Introduction

Taxis play an important role in urban transport, complementing traditional public transport modes. As urbanization expands, the demand for taxis is increasing, particularly in fast developing cities such as Beijing. As a result, taxi trips are contributing more to urban traffic volumes. For example, there were 66,600 licensed taxicabs in Beijing in 2012, which generated over 1.2 million ridden trips per day (Zhu et al., 2013). Nowadays large numbers of these taxicabs are equipped with GPS (Global Positioning System) applications, transforming these vehicles into

[1] A: Department of Architecture, 1-5 Scroope Terrace, Cambridge, CB2 1PX, U.K. E: bd315@cam.ac.uk
[2] A: Department of Architecture, 1-5 Scroope Terrace, Cambridge, CB2 1PX, U.K. E: sd560@cam.ac.uk
[3] A: Department of Architecture, 1-5 Scroope Terrace, Cambridge, CB2 1PX, U.K. E: vz209@cam.ac.uk
[4] A: Department of Architecture, 1-5 Scroope Terrace, Cambridge, CB2 1PX, U.K. E: Ying.Jin@aha.cam.ac.uk

ubiquitous probes of transport networks. The resulting information provides significant opportunities in understanding urban transport dynamics. The nature of taxi driving in busy urban areas; high passenger demand; significant levels of congestion and the desire for drivers to maximize profit through optimization of route, allows greater potential in the estimation of network speed from these datasets.

In this paper we seek to determine whether publicly released low sampling frequency GPS traces generated by urban taxis can be used effectively to identify traffic delays, particularly to explore the extent to which they are useful in estimating the congested driving speeds of urban road networks as an input to city-scale land use and transport models.

The congested link speed plays a crucial role in urban transport modelling. Historically this information has been generated iteratively from observed traffic volume on each road link. Besides the time-consuming process of collecting and processing traffic-flow data, the conventional iterative calculation itself is a complex task. The existence of an alternative method for estimating congested link speeds would potentially cut the cost in time and labour; improve the efficiency om data collection; and enhance the precision (particularly with regard to temporal accuracy) of the congested link speed estimation.

In order to evaluate the effectiveness of our proposed methods, we base our study on a publicly available set of taxi traces from Beijing. This data provides a minimal amount of information, limited to location, timestamp and taxicab ID and presents some typical shortcomings of such datasets: there is a short collection period (seven days, only three of which are work days); there is no behavioural context (no information on whether the taxicab is occupied, available, parked or looking for fares); low positional accuracy and low (and inconsistent) sampling frequency (the nature of the urban environment means signal quality is reduced and in some cases lost completely for short periods of time).

The question is whether data-sources of this kind can offer insights into the spatial and temporal patterns of the traffic conditions on heavily congested urban networks. If so, the process of the traffic assignment module in transport models could be significantly improved in a time-and-labour-efficient way, taking advantage of the widespread use of the GPS technology and the immediate availability of the sensed data. In addition, the increasing availability of the processing power and advanced computational route assignment method (Hagen-Zanker and Jin, 2012; Zachariadis et al., 2013) offer possibility to derive route choices over large origin and destination matrices so long as there are ways to estimate driving speeds and traffic conditions across the road network.

Addressing the problem, this paper represents a revised map-matching method to estimate the congested link speed for the purpose of improving the route choice module for urban transport modelling. More specifically, there are four major challenges to consider:

- The first challenge is network representation. Availability of detailed network is often a restriction, as such datasets are often costly to access or build or held by state controlled organizations. Thus we develop a semi-supervised network-generation process that relies on publically available mapping data from Open Street Map (2013). The developed network represents all traversable links with a structure of multi-lane and multi-directional, as well as complex junctions.
- The second challenge is spatial transformation. We propose a method for map-matching GPS positioning coordinates in continuous space (latitude-longitude pairs) onto the modelled road network, most importantly, for inferring paths (sequences of network road-links) between successive (typically sparse) positioning observations based on time-minimizing shortest paths.
- The next challenge is the statistical significance. We have to establish whether the collection of samples (from pairs of successive GPS taxi traces) that are used to estimate

the driving speed and traffic delay for each of the network links offer statistically significant results; i.e. whether the inferred speeds from the GPS data that correspond to a network link are distributed relatively close to their mean value. In order to justify the proposed method it must be shown that the inferred link speeds are both consistent and realistic in space and time. We provide a systematic testing of our results using a sample validation procedure.

- The last challenge is the validation of the method. A sample validation procedure is executed, and we compared the travel time between observational origin-to-destination trips and the estimated travel times based on the calculated congested link speed.

2. Related work

GPS was primarily developed for US military purposes in the 1960s (Kaplan and Hegarty, 2005). Today GPS technology has already established itself as the most popular global positioning method. The applications of GPS cover a wide range of scientific fields, such as topography, geodesy, hydrography, photogrammetry, transport (Mintsis et al., 2004). Such applications have been embedded into various portable electronic devices, such as smart phones, tablet computers, and vehicle-mounted devices (such as fleet tracking systems and satellite navigation systems). In many modern cities, taxis equipped with GPS devices have transformed them into ubiquitous probe vehicles.

As urban traffic patterns become increasingly complex, more precise and detailed information are now required by transport planners and researchers when inform policy decision. While the high adoption rate of taxi GPS applications and the spatial-temporal data collected as a result, provide us with new opportunities in understanding urban transport dynamics. Meanwhile, the availability of analysis and visualization platforms (such as Geographic Information Systems (GIS)) has enable researchers to tackle with bid datasets relating the complex driving behaviour (which typically contain millions of spatial-temporal records) at a very detailed level (Mintsis et al., 2004; Zheng et al., 2008).

Recent research can be categorized into two emerging themes:

- First is focused on the microscopic standpoint of drivers' individual behaviour, such as the optimization of location or route choice (Manley, 2012), location identification (Ashbrook et al., 2003; Kang et Al., 2005; Schmandt and Marmasse, 2004; Zhou et al., 2007), route determination (Wang et al., 2011), analysis of mobility patterns (González et al., 2008; Fang et al., 2009; Liao et al., 2007), traffic condition prediction (Bar-Gera, 2007; Herrera et al., 2010; Shi and Liu, 2010; Work et al., 2010; Zou et al., 2005; Nanthawichit et al., 2003; Liu and Ma, 2009), and mobility intelligence (Zheng and Xie, 2011; Zheng et al., 2009; Liu et al., 2010; Bohte and Maat, 2009);
- Second is comparatively new, which is in a view of macroscopic traffic characteristic investigation, such as the hot-zone extraction (Zhu et al., 2013), traffic monitoring (Shi and Liu, 2010), urban traffic dynamics studies (Geroliminis and Daganzo, 2009; Hellinga and Liping, 2002; Ye et al., 2012), and very recently route choice modelling (Ben-Elia et al., 2010; Jenelius and Koutsopoulos, 2013; Fadaei Oshyani et al., 2014; Zhan et al., 2013).

The field of congested speed estimation, using sparse taxi GPS data is rather new, and remains largely unexplored. Meanwhile in the case of macroscopic network-based analysis, the core process is the translation from continuous-space positioning information into topologically valid network-based representations. This is facilitated using map-matching methods, which typically integrate positioning data with spatial road networks to identify the correct continuous-space to network-space transformation; e.g. to infer the link in which a data-point corresponds to (Greenfeld, 2002; Ochieng et al., 2003; Quddus et al., 2007; Yuan et al., 2011; Yuan et al., 2013). As

expected, the accuracy of a map-matching algorithm will be determined by the quality of the spatial road network and the sampling rate of the GPS data. However, even though the performance of an algorithm relies on the quality and characteristics of input data, the underlying inference method will significantly affect its effectiveness (Chen et al., 2005; Parkinson et al., 1996).

This is particularly relevant when input data is of low sampling rate, as in the case of the dataset in this paper, where GPS loggers occasionally lose signal in dense urban areas, and their frequencies are set below 0.01Hz to reduce power consumption. Low frequency data requires map-matching algorithms that consider the topological validity of the inferred matches.

There have been several map-matching methods developed to deal with low-frequency data, such as (Chen et al., 2014; Hunter et al., 2013; Miwa et al., 2012; Wang et al., 2011; Ye et al., 2012). Particularly Lou et al (2009) propose such a map-matching algorithm, ST-matching that combines (1) the spatial geometric and topological structures of the road network and (2) the temporal/speed constraints of the trajectories. This is used to construct a candidate graph from which global trajectory route is selected based on a scoring system.

Rahmani and Koutsopoulos (2013) propose a simultaneous map-matching and path inference methodology for sparse GPS traces where the only information available is latitude, longitude, and timestamp. The method identifies a set of candidate links in the vicinity of each GPS observation and find a matched point on each of those links. Subsequently, all pairs of matched points of consecutive observations are connected with shortest paths between them. Since the algorithm is designed with real-time applications in mind the shortest path calculation at time t(k) is based on the estimated link speeds at time t(k-1). Their study had the advantage of known trues paths in which to evaluate the effectiveness of their method. Their findings show that their method is robust and performs favourably compared to other methods that incorporate additional observed data (such as heading and instantaneous speed) Rahmani and Koutsopoulos (2013).

The method proposed in this paper builds on the work of Rahmani and Koutsopoulos (2013) and Lou et al using very similar foundations. We also propose the adoption of a candidate link approach to the initial map-matching problem followed by the estimation of shortest paths on a topologically structured network. However there are a number of differences in methodology, scale, observational datasets, and research aims, discussed in more detail below.

The methodology we propose has been designed to facilitate multiple iterations of shortest path calculation. An initial iteration calculates time minimised shortest paths based on free flow road speeds allocated to each link based on road type (Figure 1). This can either be applied in a temporal context (similarly to Rahmani and Koutsopoulos, 2013), or be based on iterative updates to road speed based on previous iteration and subsequent refinement of the respective shortest path calculations.

This provides two opportunities; firstly as the final estimated link speed is not wholly reliant on an initial shortest path calculation based on free-flow speed, the initial route choice becomes less important assuming a suitably large dataset. Over multiple iterations, this approach will optimize link speed; secondly, and particularly suitable to smaller and/or time series datasets, congested link speed can be refined over time as further datasets are collected or are made available.

The majority of research in this area is focused on relatively limited spatial scales, or tested using small sample networks (Zhan et al., 2013; Rahmani and Koutsopoulos, 2013; Hellinga and Liping, 2002; Ye et al., 2012). In contrast, the intended use of our method is within the context of city scale land use transportation modelling; as such we are more focused on the urban or city scale.

Zhan et al (2013) propose a comprehensive method for the calculation of link travel times based on minimising the least squared difference between expected travel times and observed travel

times. They utilise a large taxi dataset from New York with known origin and destinations for defined fares. They use a sample network for Manhattan, and in comparison to the research outlined in this paper, the network complexity is relatively low: 193 nodes and 381 links. A very different approach is required in their work due to the nature of both data and network; the urban environment of Manhattan, in particular the gridded road network where multiple paths may involve very similar costs, provides very different challenges to the complex urban network of Beijing.

The dataset available for our study is very large and contains very limited attribution, with no information on taxi status, such as origin and destination points; or alternative sensor information such as acceleration. We do not have the advantage of knowing the true paths associated with each trajectory, as with Lou et al. (2009) and Rahman and Koutsopoulos (2013). This provides both limitations and opportunities.

In terms of limitations, we will never know the true path taken so we cannot validate our results based on comprehensive ground truth data. There are also numerous errors in the source data, such as inconsistencies in observed travel time versus observed distances inferring impossible speeds. We believe this is an inherent characteristic of this type of data and as such we propose methods to utilise typically available datasets in this context.

In terms of opportunities, the size of the dataset allows a selective approach and provides opportunities to remove observations that are ambiguous. Due to the large volumes of data, we are also able to estimate congested link speed at high spatial-temporal resolution (thirty minute time intervals during the morning peak for the entire extent within Beijing's fifth ring road). Our iterative approach reduces the requirement for absolute accuracy of route choice. We also validate our approach using a sample validation procedure and a comparison of travel time with observations.

In summary, the approach gains insight into the processing of complex and low-frequency spatial data in a large scale, with a relatively high temporal resolution; and provides a new point view in the generation of congested link speeds.

3. Methodology

In the following section, we propose methods for addressing the three key challenges identify in the introduction: network representation, spatial transformation (map-matching and path estimation) and statistical significance.

3.1 Development of the modelled network

Our modelled road network represents all traversable links within the fifth ring road of Beijing and is suitably detailed to accommodate our map-matching method. We use a hybrid generation process whereby major links are manually categorized by link type: expressways; second to sixth ring road; level one-to-four city major roads; and assigned one way restrictions and elevation types. We use geometry from Open Street Map to generate infill links representing minor roads with respective link characteristics (low speeds; bi-directional movement; one lane in each direction). Our network consists of 38,000 links in total (Figure 1).

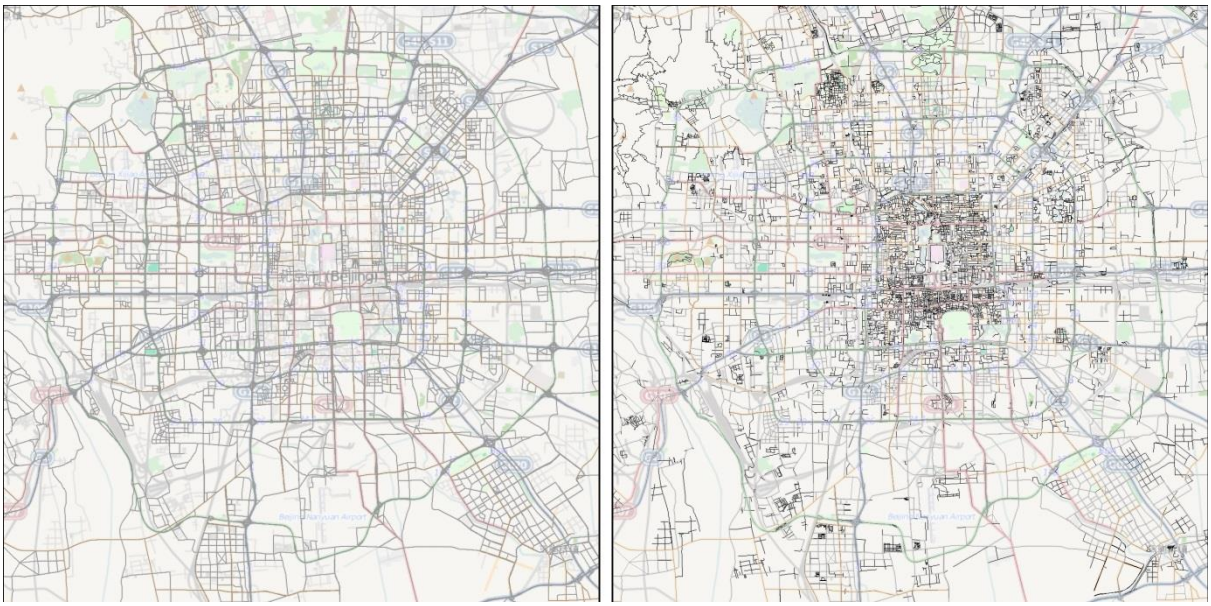

Figure 1. Network representation

3.2 Map-matching and path inference

The second challenge that we previously identified, is the development of a methodology to map-match continuous space GPS trajectories to the modelled network and the inference of paths from typically sparse observations. We propose map matching and path inference methods utilizing time minimised shortest path (Dijkstra Algorithm) with an aim to estimate traffic delays across a significant proportion of the modelled network.

Most map-matching algorithms map current or neighbour locations to the road network. The accuracy is generally low as the map-matching routine generally only considers current position, ignoring previous and following locations. These traditional map-matching algorithms rely on high sampling rate frequency and become less reliable as the uncertainty in data increases (Lou et al., 2009). Traditional methods would lose the detail between observations, relying simply on the position and associated speed at a known location measured by the GPS device. We adopt a shortest path based method (similar to that proposed by Rahmani and Koutsopoulos, 2013) that calculates the likely real word position on the network whilst estimating the path travelled between observations allowing more accurate path estimation and supplying significant increases in the speed observations available throughout the entire network.

Inferring paths from sparse GPS observations using shortest path is well documented (Lou et al., 2009; Rahmani and Koutsopoulos, 2013; Patterson et al., 2003; Fadaei Oshyani, 2011). Distance minimised shortest path calculations have been shown to be unrealistic when compared with actual routes. A study of London using a similar GPS datasets from taxis has shown that a distance minimised shortest path predicts on average only forty per cent of the actual route travelled (Manley, 2012).

Intuitively, taxi drivers are experienced drivers who can usually find out the fastest route to send passengers to a destination based on their extensive local knowledge (Yuan and Zheng, 2010). We therefore believe that estimating inferred path using time minimised shortest path is a more realistic assumption.

The GPS trajectory dataset

The GPS trajectory dataset used in this paper consists of 1,500,000 observations from 10,357 taxis in Beijing covering the working period Monday 4th to Wednesday 6th February 2008 (Table 1).

The data was obtained from Microsoft Research Asia. Each observation includes details of taxi identifier, location and timestamp.

Table 1. The result of the filtering process

	4th Feb	5th Feb	6th Feb
All observations	562,612	509,690	422,795
Within fifth ring road	312,869	266,807	162,293
Within network tolerance	171,939	134,212	69,467
Within speed tolerance	311.052	265,226	161,273
Within length tolerance	171,246	142,112	75,829
Used in path inference	115,018	90,977	46,614

The median sampling time of trajectories is 219 seconds (Table 2). Assuming an average speed of 20 kph then the travel distance between GPS observations would 1,200 meters. This poses a challenge when estimating link speed and traffic delays over the entire modelled network.

Table 2. GPS sampling time

Date	Counts of Records		Sample Time /s	
(5am-11am)	Total	Filtered	Mean	Median
4th Feb 2008	562,612	115,018	206.10	221
5th Feb 2008	509,690	90,977	219.61	251
6th Feb 2008	422,795	46,614	231.97	300
Total	1,495,097	252,609	219.23	seconds

The trajectory dataset has been filtered to remove those observations that are inaccurate or unsuitable for map-matching: Trajectory lengths of less than 100 meters are removed; Point to point trajectory speeds of greater than 120 kph are removed; Trajectories are filtered based on proximity to modelled network (this is described in more detail below).

Trajectories outside of set tolerances from the modelled network are removed from the map-matching process. The tolerances used are shown in Table 2 and represent a buffer equal to:

$$b_{R^m} = (d_l\left(f(R^m)\right) x\, n_l(R^m)) + (d_s\left(f(R^m)\right) x\, n_s(R^m)) + d_u\left(f(R^m)\right) \quad (1)$$

where b_{R^m} is the buffer distance of link R^m, $f(R^m)$ is the classification function (returns R^m's link type), $d_l\left(f(R^m)\right)$ is the typical lane width for R^m's link type, $n_l(R^m)$ is the number of lanes in R^m, $d_s\left(f(R^m)\right)$ is the typical separator band width for R^m's link type, $n_s(R^m)$ is the number of separators in R^m and $d_u\left(f(R^m)\right)$ is the typical utilities width for R^m's link type.

The network proximity tolerance aims to consider the total width of the road when allocating trajectory start and end points to a candidate network link. This allows greater tolerances to road types that are generally wider to account for the positional accuracy of the GPS location.

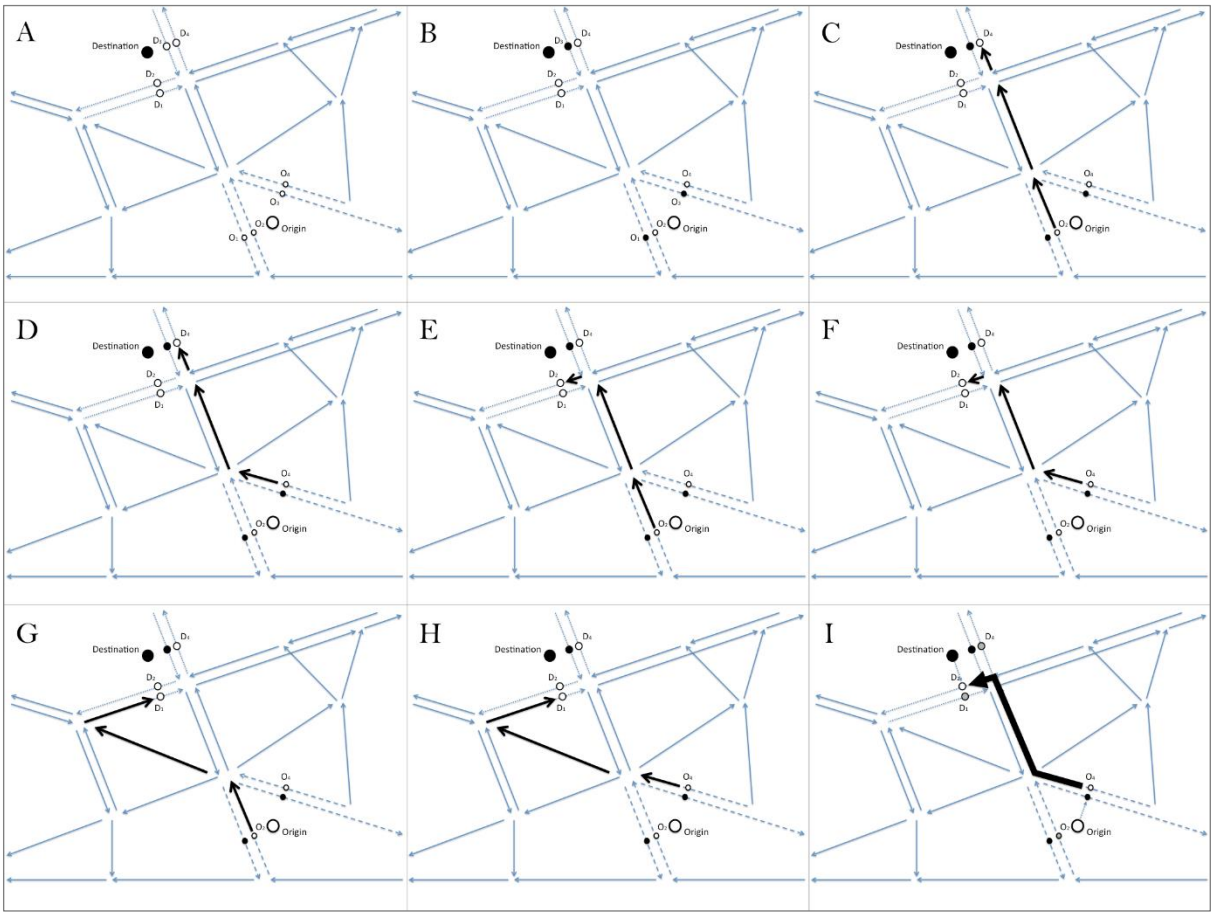

Figure 2. The combined map-matching and path inference process. For all pairs of successive GPS points (Origin-Destination $P^1 \rightarrow P^2$) the proposed algorithm (1) finds all candidate start links CsL^n and end links CeL^n within the appropriate distance buffers b_{R^m} and calculates their respective candidate start CsP^n and end CeP^n points (A); (2) identifies the on-network points that do not hold appropriate connectivity properties (B); (3) for each combination of on-network OD points calculates the time-minimising path (C-H)[5]; and (4) selects the OD pair with the minimum time-minimising path (I)

This paper defines the taxi trajectory T as a pair of sequential taxi locations. Each trajectory therefore consists of a start point P^1 and end point P^2. Each point consists of a location $(x(P^n), y(P^n))$ and a timestamp $T(P^n)$. A trajectory can therefore be described as $T: P^1 \rightarrow P^2$ and the observed time difference between consecutive points can therefore be described as $T(P^2) - T(P^1)$.

The initial problem is how to allocate each of the trajectory points P^1 and P^2 to a precise location on a network link (any location along $R^1 \rightarrow R^2 \rightarrow \cdots \rightarrow R^n$). The proposed solution is the generation of candidate start CsL^n and end links CeL^n for each trajectory. Candidate links are therefore all links within a search tolerance ε of trajectory points P^1 and P^2. Candidate start CsP^n and end CeP^n points are then generated by calculating the nearest point on each candidate link to P^1 and P^2 (Figure 2).

We define ε using the network proximity tolerance defined in Table 2, which is dependent upon link type. We therefore search wider for candidate links representing higher status roads to account for their greater widths and the likelihood of additional locational inaccuracies on these road types.

[5] The time traversing each of the links is, in first instance, based on free flow link speeds. In case of more than one iteration, iteration k uses the link speeds estimated at iteration $k-1$.

The problem becomes as follows: for a set of candidate start and end locations pairs how do we define the most probable pair and, given that we know the precise start and end location of each trajectory, how do we estimate the route taken between consecutive points and how do we use this to estimate speed along the entire network?

Map-matching and path estimation

We propose the use of a time minimised shortest path to estimate candidate paths taken between each pair of candidate start and end locations for each trajectory. We then select the most likely path from all candidates for any given trajectory as the path with smallest time cost. We describe this route as the initial path. The initial path can be described by a series of road links such that travel time between $P^1 \rightarrow P^2$ is minimised based on free flow road speeds. We are then able to infer the precise start and end locations for each trajectory based on the initial path, which is used in later analysis.

We propose further refinement of the estimated speeds by introducing an iterative process. In each round of iteration, the refined time-minimising paths are calculated using the estimated link speeds of the previous iteration. In this case, the refined paths can be described by a series of road links such that the travel time between point $P^1 \rightarrow P^2$ is minimised based on estimated link speeds derived from the paths of the previous iteration.

Estimation of link traffic delays and driving speeds

In order to estimate a representative speed for each modelled network link, we split the initial path (or refined path in later iterations) for each trajectory based on link type and associated free flow travel time of the component network links. We calculate free flow travel time[6] for each portion of the path based on free flow link speed and link length ($t_{R^m} = l_{R^m}/s_{R^m}$). We then calculate path travel time for each path based on observed path length and trajectory time difference ($T(P^2) - T(P^1)$). We then allocate a proportion of path travel time to each path portion based on free flow travel time to path travel time ratio and then allocate a speed value for each link traversed on that path. Therefore, in the case of the pair of successive trajectory points P^1 and P^2 which are map-matched to path $R^1 \rightarrow R^2 \rightarrow \cdots \rightarrow R^n$, where R^m are network links, the estimated time of traversing link R^k after iteration k the pair of successive trajectory points P^1 and P^2 will be:

$$t_k^{P^1 \rightarrow P^2}(R^m) = \left(T(P^2) - T(P^1)\right) \cdot \frac{distance_k(R^m)/speed_{(k-1)}(R^m)}{\sum_n distance_k(R^n)/speed_{(k-1)}(R^n)} \quad (2)$$

where $T(P^1), T(P^2)$ are the timestamps for the successive GPS points P^1 and P^2, $speed_{(k-1)}(R^m)$ is the speed estimation for link R^m after iteration $k-1$, and $distance_k(R^m)$ is the length of the cost-minimising path between points P^1 and P^2 inside network link R^m in the current iteration k.

Table 3. The results of path inference

	Day 3	Day 4	Day 5
Trajectories processed	115,018	90,977	46,614
Links within 5th ring road	47,262	47,262	47,262
Number of routes generated	110,414	20,113	11,483
Number of speed values	1,203,646	209,941	122,151
Links with speed values	27,260	20,157	17,326
Average route length (km)	3.3	-	2.9

[6] In subsequent iterations free flow link speeds are replaced by the estimated link speeds of the previous iteration.

Using equation 2, we are then able to calculate an estimated mean speed for each modelled network link by day and time of day, along with statistics such as speed range, standard deviation, number of observations and the difference between free-flow speed and observed speed. In the k^{th} iteration of the process, for link R^m, these metrics will be based on a statistical sample that contains all pairs of successive GPS points P^1 and P^2 that contain part of R^m in the cost-minimising path between P^1 and P^2. Therefore, the estimated link speed of R^m after iteration k will be equal to:

$$speed_k(R^m) = \frac{1}{\|M_k(R^m)\|} \cdot \sum_{P^1 \to P^2 \in M_k(R^m)} \left[distance_k(R^m) \Big/ t_k^{P^1 \to P^2}(R^m) \right] \quad (3)$$

Where $M_k(R^m)$ is the set that contains all the $[P^1, P^2]$ GPS point pairs, for which the cost-minimizing path in iteration k traverses (either partially or fully) network link R^m. The results of the path-inference process are shown in Table 3. These will be analysed in section 3.3.

3.3 Statistical analysis of resulting speed estimates

Following equation 3, the objective of this piece of analysis is to explore whether the aggregate traffic delays and speeds $speed_k(R^m)$ of the network links are based on statistically significant distributions of individual estimations $t_k^{P^1 \to P^2}(R^m)$. This is an essential part of the proposed methodology; in order to justify the use of the link-based speed estimations we should demonstrate that they are both consistent and realistic in space and time; i.e. that speed estimations $speed_k^{P^1 \to P^2}(R^m)$ of any GPS point pairs $[P^1, P^2]$ in $M_k(R^m)$ with similar timestamps $T(P^1), T(P^2)$ will not vary dramatically around an arbitrary mean value, but will be tightly distributed around it.

To establish that this is the case, we calculate the relative standard deviation (RSD) for all links for different time periods and plot the results as a function of the size of the sample $\|M_k(R^m)\|$. Moreover, we want to establish that the size of the RSD for link R^m is not dependent on the spatial location of the R^m or at least any spatial variation does not introduce systematic biases that would undermine the usefulness of the speed estimations. Having said that, we expect to observe systematic RSD variation between different link types (e.g. motorways versus "in-fill" links) because the circulation characteristics of each link type is unique.

Figure 3 illustrates road-link RSD as a function of size of the sample (top-left plot). As expected, as the sample size increases the variation of the relative deviation of different links decreases, because of the rule of big numbers. The top-right plot shows the cumulative distribution of sample sizes of the links. At least half of the links base their estimated speed $speed_k(R^m)$ on 16 samples or more. The two plots at the bottom show the respective RSD distributions for two short periods of time (30 minutes each). Our expectation, temporally concentrated samples should result in lower RSDs, is confirmed; average RSD for all links is 0.39 for the 6.00-11.00am period, and 0.32 and 0.33 respectively for the 7.30-8.00am and 9.30-10.00am periods. Moreover, the postulation that different types of links should have different levels of RSD is also confirmed. The average RSDs of infill-links and low tier road-links are 0.40 and 0.39 respectively. On the other hand, the average RSDs of road-links with speed limits equal to 90kph and 120kphare 0.29 and 0.22 respectively. The higher RSD variation of the low speed network reflects their multiple functions, higher probability to encounter congestion and the behaviour of the taxi drivers (collection of passengers etc.). Similarly, the lower RSD values of the high-speed links reflect arterial conditions; and in the case of the 120 kph links, the freedom to apply preferred speed by choosing from multiple lanes.

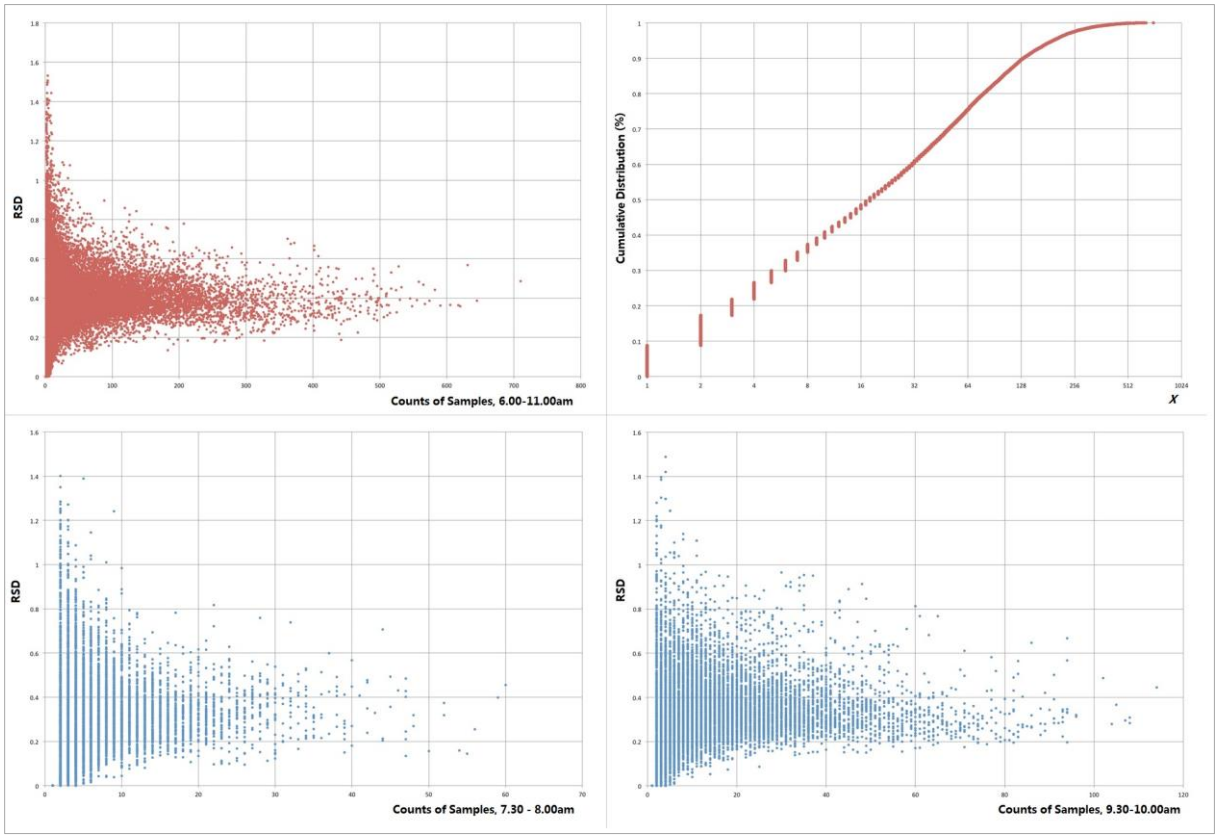

Figure 3. Relative Standard Deviation. Top-left plot shows RSD as a function of sample size for the full length of the morning peak period (6.00 – 11.00am). Top-right plot shows the cumulative distribution of links with sample size less than x. Bottom plots show RSD as a function of sample size for the 7.30-8.00am and the 9.30-10.00am time slots. As expected, as the sample size increases) the relative deviation for the shorter time periods tends to lowers values. Average RSD for all links is 0.39 for the 6.00-11.00am period, and 0.32 and 0.33 respectively for the 7.30-8.00am and 9.30-10.00am periods.

Figure 4. Relative Standard Deviation (RDS) of speed estimation (left) and number of speed estimations (sample size) in each link (number of taxi GPS traces used to estimate link speed) (right).

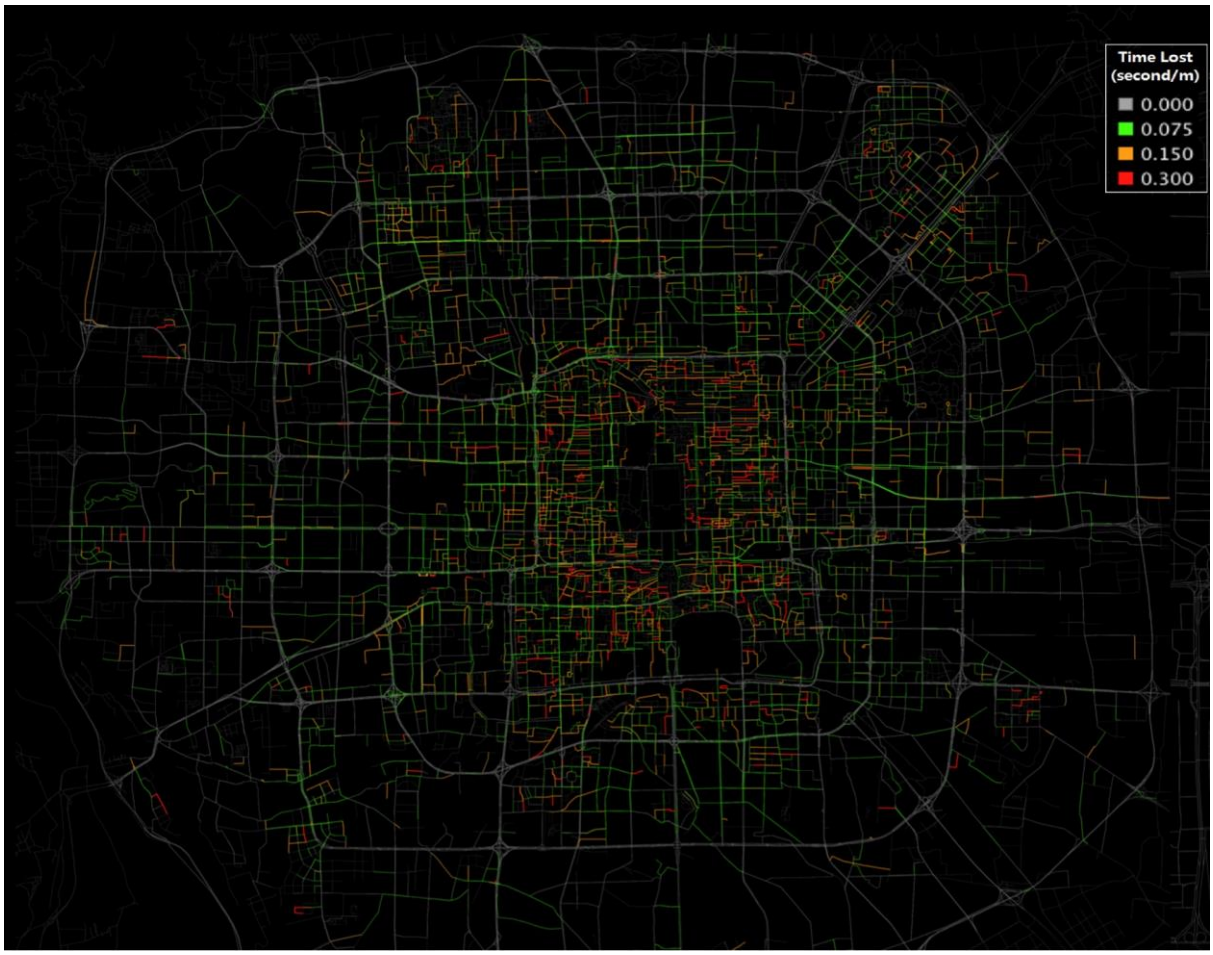

Figure 5. Time lost in traffic based on the taxi-GPS speed estimations (in seconds per metre)

Figure 4 illustrates the spatial distribution of the relative standard deviation for the full morning peak period including the shoulder hours (6.00-11.00am) and the respective sample sizes; the lack of spatial pattern in the distribution of the RDS sizes is good news, because it implies that there is no systematic, spatially introduced bias in the speed estimation. On the other hand, the spatial distribution of the sample sizes results in a distinct pattern that highlights the hierarchical structure of the road network. Figure 5 shows the resulting time $T_{lost}(R^m)$ lost in traffic based on the taxi-GPS speed estimations for the full morning-peak period (6.00-11.00am). It represents the time over free flow required to traverse a metre of road (measured in seconds per traversed metre or road):

$$T_{lost}(R^m) = \frac{1}{l_{R^m}} \cdot \left(\frac{l_{R^m}}{speed_k(R^m)} - \frac{l_{R^m}}{speed_{f.f.}(R^m)} \right) = \frac{1}{speed_k(R^m)} - \frac{1}{speed_{f.f.}(R^m)} \quad (4)$$

where $\text{speed}_{\text{f.f.}}(\text{R}^{\text{m}})$ is the free-flow speed of the link R^{m}. This measure reflects the level of congestion in the network; i.e. how much more time than a vehicle would need if there was no traffic will be needed per metre under the morning peak traffic conditions. Since this is calculated per metre there is no visual distortion related to the length of each link. Finally, Figure 6 shows the average estimated link speed by type of network link and Figure 7 shows the average estimated link speed on major expressways in Beijing, both for the three days of testing.

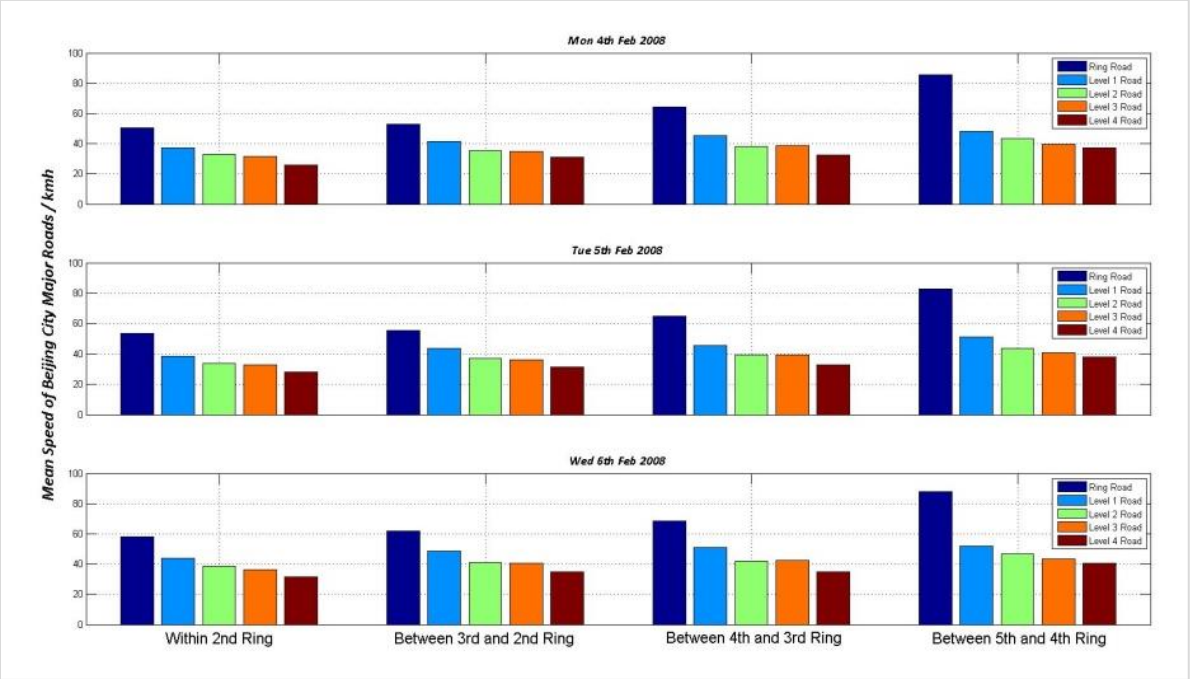

Figure 6. Average estimated link speed by type of link for the three days of testing

The average link speed increases as we move outwards from the centre of Beijing.

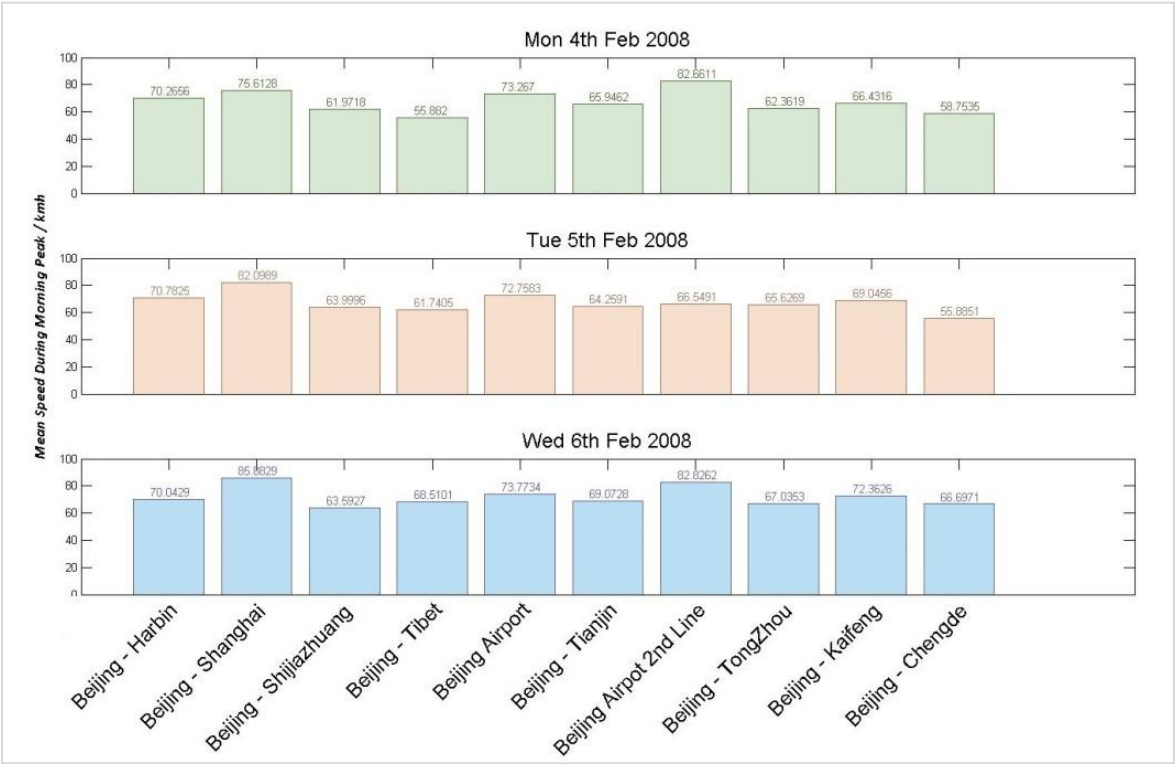

Figure 7. Average estimated link speed on 10 expressways in Beijing for the three days of testing

It should be noted that the 3 days of testing are the last weekdays before the Chinese New Year and therefore, variations between them should be expected because the travel demand changes as we move closer to the public holiday.

3.4 Validation of Method

In this section, we now introduce a sample based validation procedure to assess the accuracy and effectiveness of the method, by comparing travel times between modelled and observed data.

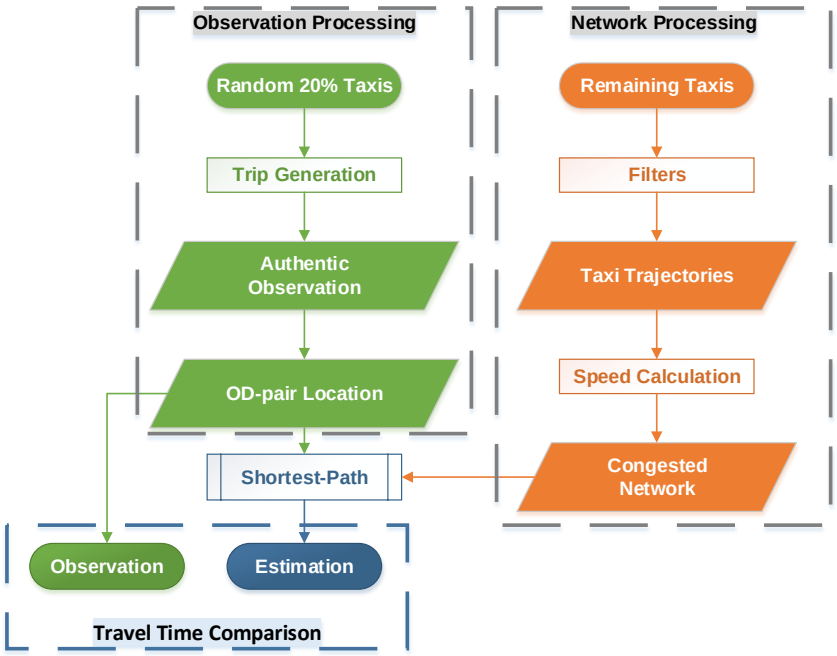

Figure 8. Flow chart of validation process

The process is illustrated in Figure 8. From the 10,357 taxis forming the dataset for the 4th February (Monday), 20% are randomly selected and their trajectories combined to form an observational dataset. The trajectories from the remaining eighty per cent of taxis are combined to form an experimental dataset which is used to generate congested link speeds. Then the travel time from estimated shortest-path and observed trip duration will be compared.

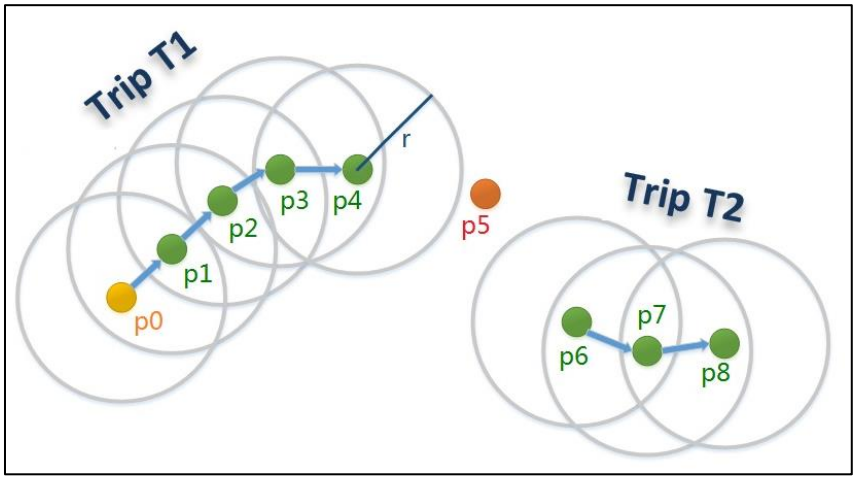

Figure 9. Generation of observed trips

As highlighted, taxi status information is unavailable, while the dataset is provided as a series of continuous locations and timestamps for each taxi. This inhibits the comparison of travel times using the modelled versus observed approach. In order to create a meaningful comparison, we combine consecutive trajectories from the observational dataset to form taxi trips. We assume that taxi drivers would prefer to choose the shortest path based on their experiences and awareness of the congested network speed.

Trips are formed from consecutive trajectories when the following criteria are all met:

- All trajectories belong to one taxi;
- The time interval between any two consecutive trajectories is shorter than a defined maximum time interval. This requirement is applied to ensure the continuity of the trip;

- There are at least three consecutive trajectories in each trip;
- The start point and end locations of each trip are different to ensure only moving taxis are simulated;

The ratio of the actually observed travel distance against the start-to-end-point straight line length is below the 90% value of the ratios for all trips. Due to the inaccuracies in the original dataset, this rule is further applied to remove trips with overwhelmingly longer distance or faster speeds. This occurs in reality when taxi drivers wonder around within a specific region; while in the datasets, this is reflected by significant errors of the ratio of observed accumulated trip length (calculate as the straight line distance between all trajectories forming a trip) and the straight line distance between trip start point and trip end point. We then remove the ten per cent of observation versus modelled data points with the highest ratio (20.85 and 3.96 respectively for sixty-second-interval and ninety-second-interval trips).

Table 4. Generated Observed Traces

Max Time Interval	Raw Trip (counts)	Filtered Trip (counts)	Authentic Trip (seconds)
60 seconds	1417	891	846
90 seconds	1320	821	780
Total	2737	1712	1626

As shown in Figure 9, p0-p8 points are nine consecutive trajectories for one taxi. The processing starts from the earliest point p0; if time difference between any two consecutive points smaller than r, then a trip is generated (such as Trip T1). The trip stops at point (p4), where the subsequent trajectory interval is greater than r. This process is repeated for all trajectories in the observational dataset, with points not forming part of a trip ignored.

Our observational dataset consists of trajectories from 2,071 taxis, twenty per cent of the entire 10,357 taxis in the original dataset. We generate observed trips during 6am-11am on the 4th February with r values of 60 and 90 seconds. Table 4 demonstrates the number of generated observed trips for each of the maximum-interval segmentations. There are 1712 original trips generated in total, and to guarantee the authenticity of the observations, these trips are then further filtered respectively for each segmentation by the 95% travel time value.

So far the pre-processing of observations has been implemented, and consequently there are 1626 authentic trips will be used in the following comparison analyses.

Generation of the congested link speed from the experimental dataset

Congested link speed is estimated for each 30 minute period during the extended morning peak (from 6am to 11am) calculated from the experimental dataset and our path estimation and congested link speed calculation methods, outlined in chapter 3. Where link speed has not been calculated for an individual link due to insufficient information in the experimental dataset, we assign the average congested link speed for the network by link type (as show in Table 5). This forms a series of comprehensive time based networks within the fifth ring road of Beijing.

We then calculate travel time for each trip based on a time minimised shortest path using the congested link speed corresponding to the trip start time. For example, a trip beginning at 10:06am is modelled using the 10:00am congested link speed.

Table 5. Average congested link speed (kph) by link type in each time segmentation

	Average Congest Link Speed (kph)										
Link Type	6:00	6:30	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30	Morning Peak
4th level/ unclassified	22.2	19.9	22.2	20.6	20.9	19.3	19.2	18.1	18.3	18.1	18.6
2nd ring /3rd level road	28.5	29.0	28.0	28.3	27.6	25.9	25.2	23.7	23.4	23.6	25.1
3rd ring /2nd level road	33.2	33.2	32.7	32.6	31.2	29.5	28.4	27.0	27.0	26.8	29.2
4th ring /1st level road	40.2	39.5	39.1	36.8	37.5	35.6	32.4	30.1	31.2	30.8	33.6
5th ring / expressway	55.2	57.0	56.0	54.6	52.7	49.6	47.0	45.1	44.9	45.1	47.9
6th ring	74.1	78.8	71.2	75.9	71.8	64.2	71.5	67.8	61.3	66.4	68.9
6th ring/ expressway	61.4	69.3	62.7	63.2	63.1	59.9	58.4	53.8	52.1	55.6	58.1
expressway	80.2	76.4	76.4	72.4	72.6	67.1	68.2	67.0	66.1	65.4	69.7

Analysis of validation results

Figure 10 and Figure 11 illustrate the comparison between modelled travel time and observed travel time, for trips with sixty-second and ninety-second maximum interval respectively. The comparison suggests that when trips are generated using the sixty-second maximum interval, our method is capable of producing modelled-trip travel times that have a broadly satisfactory relationship to the observations. The heteroskedastic pattern shown in Figure 10 (i.e. the margin of error for the modelled versus observed becoming larger as the durations of the trips are larger) is to be expected, as we compare an average travel time estimated off the model network against individual taxi trajectories. However, the modelled travel time in aggregate is 5% higher than the observed, as indicated by the linear fitting slope of 1.05. When we compare the estimated trip times with those observed taxi trajectories with pausing intervals ranging from 0 up to 90 seconds, then there is a much higher discrepancy: the estimated trip times are some 19% below (i.e. with the slope being 0.81). We cannot be certain why this is the case, but since the traffic signal intervals are rarely more than sixty seconds long, this sharp change in the comparison indicates that introducing those taxi traces with 60-90-second interval into the sample may not be advisable. Such delayed intervals might be caused by either the observed times having been prolonged by non-traffic delays (e.g. detours and waiting for customers), or some other errors in the observed data. This will need better quality data being available in the future.

There is also the potential issue of the reduction of the number of observations used to estimate congested link speed. Although the total number of observation in the original dataset is very large, the network is also very dense, meaning that is some instances few observations are used to estimate congested link speed. Where these links are then used in the modelled trip shortest path we would expect to see fluctuations in the modelled travel time.

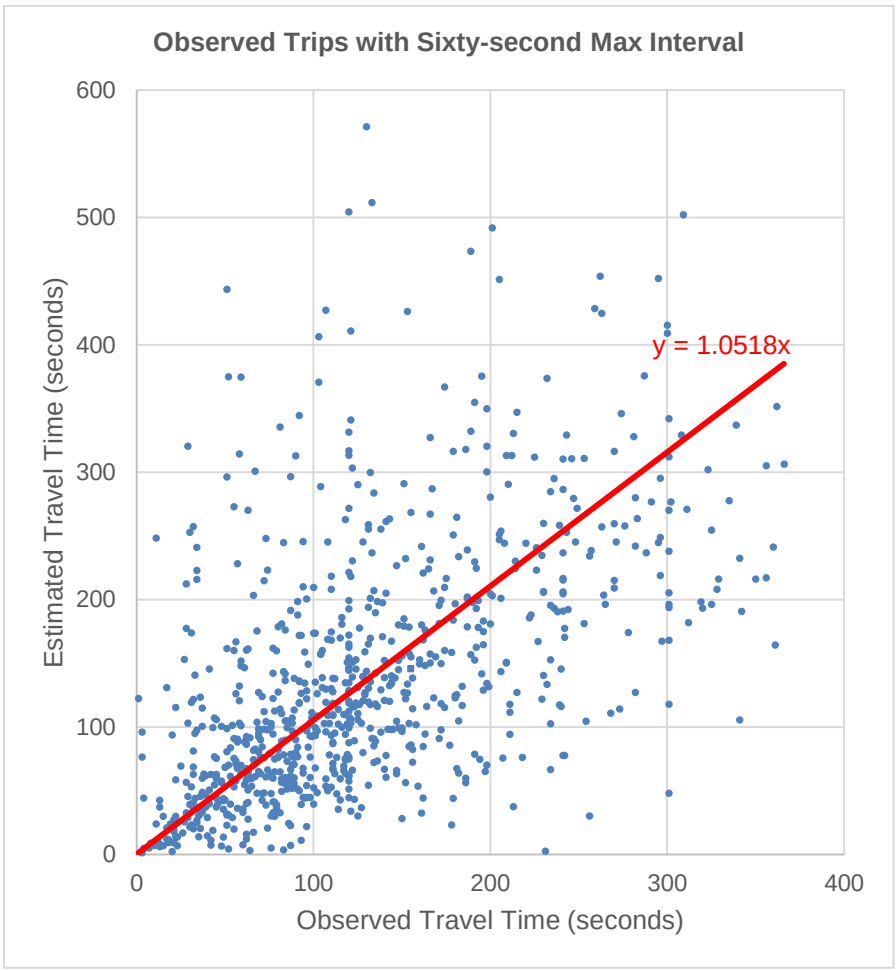

Figure 10. Estimated-observed travel time comparison for trips with 60-seconds max interval

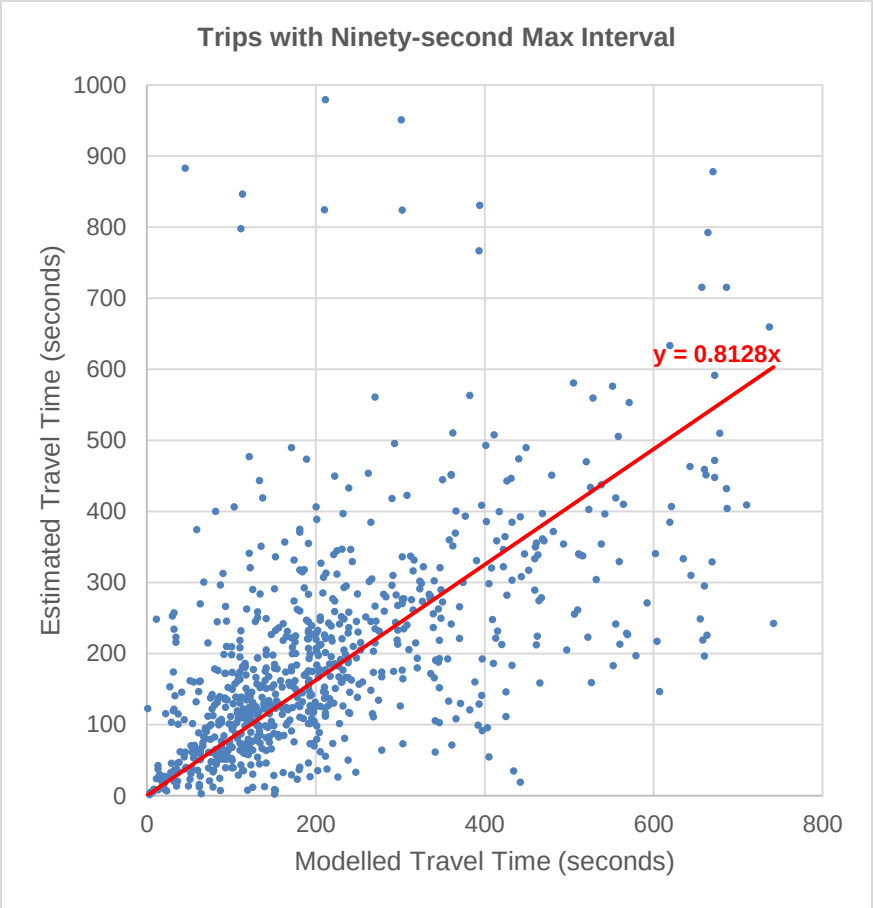

Figure 11. Estimated-observed travel time comparison for trips with 90-seconds max interval

Although the linear fitting curves are acceptable, there are errors when comparing individual pairs. More precisely speaking, the median values of proportional absolute error of estimated travel times are 36.8% and 37.4% respectively for sixty-second and ninety-second interval trips. This means that for all 1626 observational trips, there are 50% (813) of them, have estimated travel time with error smaller than 37%, when comparing with the observed travel time. This can be explained by limiting factors already discussed; the inconsistencies in the accuracy of the original dataset itself; limiting the congested speed estimation to just 80% of original data; and difficulties in inferring trips from what is a very limited dataset in terms of identifying taxi status.

We have shown through this validation exercise that the methods outlined in this paper provide satisfactory congested link speeds across the network. These congested link speeds produce estimated travel times for a randomly generated sample of trips that are within acceptable levels when compared with observed travel times. This is especially true when the trips are generated using time intervals of no longer than 60 seconds. Meanwhile the median values show that there are 50% of the 1626 estimations having errors smaller than 37%. But those errors are inherent with such datasets with low or inconsistent-precision of the original observations and exaggerated by the difficulty in comparing shortest-path routes and the real routes taken by a taxi driver where status information is unknown.

Having said that, the validation result verifies the method as acceptably accurate, especially when the time interval of observation is no longer than 60 seconds.

4. Conclusions and further work

We have tested the feasibility of estimating traffic delays and link speeds for a strategic transport model of Beijing, from low sampling frequency taxi GPS traces, generated over three consecutive days. The available GPS dataset contained only basic information: a series of time-stamped coordinates for each vehicle. We build on previous work, to derive the link speed data at the city scale, from high volume but low-sampling-frequency GPS traces.

The algorithm combines hierarchical, geometric and topological structures of the road network and temporal and speed constraints of the trajectories. It relies on an information-rich network representation, which is generated through a semi-automated process. The matching tolerance varies between different network link types, which has a positive impact on map-matching precision and increases the probability of considering a valid subset of trajectories. Moreover, the path inference process is based on time minimisation rather than observational probability. This allows the application of an iterative refining process which results in a deterministic set of inferred paths that guarantee minimisation of time between any pair of consecutive GPS traces. All the proposed processes can be implemented in standard GIS tools, however the authors recommend a different approach in order to optimise computational efficiency; much of the map-matching and path estimation tools used in this paper have been developed using Python.

We argue that the analysis of low-frequency taxi GPS traces can facilitate a significant improvement of the temporal resolution of conventional transport models and potentially generate reliable link speed estimations, especially when the used dataset includes contextual information (vehicle status etc.). However, our analysis has shown that even in the case of basic time-place datasets (which covers the majority of available GPS datasets), speed estimations could be robust; half of the results have an RSD below 34.9%, 34.2% and 32.3% for the three days of data collection. Meanwhile, the results of method validation using the 20% sample taxis, also indicates that 813 observations among all 1626 sample trips, have a proportional absolute error below 37%. Those errors are inherent and not avoidable due the limitation factors mentioned above. Thus the results can be concluded as this is arguably justifies the use of the estimations to increase the temporal resolution of link speeds estimated using more static (and possibly computationally intensive) approaches, such as conventional transport models.

4.1 Further work

This is a relatively unexplored research area and as such it provides exciting future opportunities for simplifying the process of transport model development. The main short-term objectives are (1) further development of the proposed methods (map-matching, path inference and network generation) and (2) addressing a series of identified shortcomings and un-tackled issues, including:

- Further automation of the network generation process and populating link attributes directly from publicly available resources
- Development of methods to tackle zero speed observations in order to identify status (e.g. parking, traffic lights, congestion) in basic time-place datasets
- The implementation of the methods into an integrated software package to improve automation and computational performance

At the same time, we are keen to explore several other directions: Further development of methods using other data sources (such as the full Beijing taxi dataset consisting of 30 days of data) to investigate how precision is affected by data availability and identify potential thresholds in terms of data requirements; apply the method to other study areas and compare finding from our Beijing study; investigate the spatiotemporal traffic patterns for different link types and search for distinct characteristics.

Acknowledgements

We would like to thank Microsoft Research Asia for releasing publically the dataset which has been critical to the development of this paper, and wider work on the development of the strategic transport model. We are also grateful to Open Street Map and contributors for the development of comprehensive road topology for Beijing (and around the world), without which the network creation would have been prohibitively costly. We would also like to thank the editor and reviewers for their comments and suggestions, which have helped shape the paper and methods.

References

Ashbrook, D. and Starner, T. (2003). Using GPS to learn significant locations and predict user movement across multiple users. *Personal and Ubiquitous Computing*, 7(5), 275-286.

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C*, 15(6), 380-291.

Ben-Elia, E. and Shiftan, Y. (2010). Which road do I take? A learning-based model of route-choice behavior with real-time information. *Transportation Research Part A: Policy and Practice*, 44(4), 249-264.

Bohte, W. and Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(13), 285-297.

Chen B.Y., Yuan H., Li Q., Lam, W.H.K., Shaw, S.L. and Yan K. (2014). Map matching algorithm for large-scale low-frequency floating car data. *International Journal of Geographical Information Science*, 28(1), 22-38.

Chen, W., Li, Z., Yu, M. and Chen, Y. (2005). Effects of sensor errors on the performance of map matching. *Journal of Navigation*, 58, 273-282.

Drane, C.R., MacNaughtan, M.D. and Scott, C.A. (2003). *Location and Tracking System*. U.S. Patent 6,522,890.

Fadaei Oshyani, M. (2011). Estimating route choice models using low frequency GPS data, *M.Sc. thesis*, Supervisor A. Karlström, TSC-MT 11-024, KTH.

Fadaei Oshyani, M., Sundberg, M. and Karlström, A. (2014). Consistently estimating link speed using sparse GPS data with measured errors. *Procedia-Social and Behavioral Sciences*, 111, 829–838.

Fang, H., Hsu, W.J. and Rudolph, L. (2009). Mining user position log for construction of personalized activity map. *ADMA*, 444–452.

Geroliminis, N. and Daganzo, C.F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9), 759-770.

González M.C., Hidalgo C.A. and Barabási, A.L. (2008). Understanding individual human mobility patterns. *Nature*, 453, 779–782.

Greenfeld, J. (2002). Matching GPS observations to locations on a digital map. *Transportation Research Board 81st Meeting* Washington, D.C., USA

Hagen-Zanker, A. and Jin, Y. (2012). A new method of adaptive zoning for spatial interaction models. *Geographical Analysis*, 44(4), 281-301.

Hellinga, B. and Liping, F. (2002). Reducing bias in probe-based arterial link travel time estimates. *Transportation Research Part C*, 10, 257-273.

Herrera, J. C., Work, D. B., Herring, R., Ban, X. J., Jacobson, Q. and Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C*, 18(4), 568-583.

Jenelius, E. and Koutsopoulos, H.N. (2013). Travel time estimation for urban road networks using low frequency probe vehicle data. *Transport Research Part B*, 53, 64-81.

Kang, J.H., Welbourne, W., Stewart, B., and Borriello, G. (2005). Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review*, 9(3), 58-58.

Kaplan, E.D. and Hegarty, C. (2005). *Understanding GPS: Principles and Applications*. Artech house, London

Liao, L., Fox, D. and Kautz, H. (2007). Extracting places and activities from GPS traces using hierarchical conditional random fields. *International Journal of Robotics Research*, 26(1), 119-134.

Liu, H. X. and Ma, W. (2009). A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. *Transportation Research Part C: Emerging Technologies*, 17(1), 11-26.

Liu, L., Andris, C. and Ratti, C. (2010). Uncovering cabdrivers' behaviour patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6), 541-548.

Lou, Y., Zhang, C., Zheng, Y. and Xie, X. (2009). Map-matching for low-sampling-rate GPS trajectories. In *Proceedings of ACM SIGSPATIAL Conference on Geographical Information Systems*.

Manley, E. Navigating the city: Minimising distance but not minimal distance. Available online at: http://urbanmovements.co.uk/2012/06/20/navigating-the-city-minimising-distance-but-n/#comments [Accessed 12 March 2014]

Mintsis, G., Basbas, S., Papaioannou, P., Taxiltaris, C. and Tziavos, I.N. (2004). Applications of GPS technology in the land transportation system. *European Journal of Operational Research*, 152(2), 399-409.

Miwa, T., Kiuchib, D., Yamamotoa, T. and Morikawac, T. (2012). Development of map matching algorithm for low frequency probe data. *Transportation Research Part C: Emerging Technologies*, 22, 132–145.

Nanthawichit, C., Nakatsuji, T., Suzuki, H. Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. *Transportation Research Record: Journal of the Transportation Research Board*, 1855.1 (2003): 49-59.

Ochieng, W.Y., Quddus, M. and Noland, R.B. (2003). Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, 2(55), 1-14.

Parkinson, B.W. (1996). *Progress in Astronautics and Aeronautics: Global Positioning System: Theory and Applications*. American Institute of Aeronautics and Astronautics.

Patterson, D.J., Liao, L., Fox, D. and Kautz, H. (2003). Inferring high-level behavior from low-level sensors. *UbiComp* 2003, 73-89.

Quddus, M.A., Ochieng, W.Y. and Noland, R.B. (2007). Current map-matching algorithms for transport applications: state-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5), 312-328.

Rahmani, M. and Koutsopoulos, H.N. (2013). Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, 30, 41-54.

Schmandt, C. and Marmasse, N. (2004). User-centered location awareness. *Computer*, 37(10), 110-111.

Scott, C.A. and Drane, C.R. (1994). Increased accuracy of motor vehicle position estimation by utilising map data: vehicle dynamics, and other information sources. In *Proceedings of Vehicle Navigation and Information Systems Conference*.

Shi, W. and Liu, Y. (2010). Real-time urban traffic monitoring with Global Positioning System-equipped vehicles. *IET Intelligent Transport Systems*, 4(2), 113-120.

Shi, W. and Liu, Y. (2010). Real-time urban traffic monitoring with Global Positioning System-equipped vehicles. *IET Intelligent Transport Systems*, 4(2), 113-120.

Timothy Hunter, T., Abbeel, Alexandre, P. and Bayen, M. (2013). The path inference filter: model-based low-latency map matching of probe vehicle data. *Springer Tracts in Advanced Robotics*, 86, 591-607.

Wang, W., Jin, J., Ran, B. and Guo, X. (2011). Large-scale freeway network traffic monitoring: A map-matching algorithm based on low-logging frequency GPS probe data. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 15(2), 63-74.

White, C.E., Bernstein, D. and Kornhauser, A.L. (2000). Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1), 91-108.

Work, D.B., Blandin, S., Tossavainen, O.P., Piccoli, B. and Bayen, A.M. (2010). A traffic model for velocity data assimilation. *Applied Mathematics Research Express*, 1, 1-35.

Ye, Q., Szeto, W. and Wong, S. (2012). Short-term traffic speed forecasting based on data recorded at irregular intervals. *IEEE Transactions on Intelligent Transportation Systems*, 13(4), 1727-1737.

Yuan, J. and Zheng, Y. (2010). An interactive voting-based map matching algorithm. In *proceedings of the International Conference on Mobile Data Management 2010 (MDM 2010)*, Kansas City, Missouri, USA.

Yuan, J., Zheng, Y., Xie, X. and Sun, G. (2011). Driving with knowledge from the physical world. *The 17th ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD'11)*, New York, NY, USA, ACM.

Yuan, J., Zheng, Y., Xie, X. and Sun, G. (2013). T-Drive: Enhancing driving directions with taxi drivers' intelligence. *Transactions on Knowledge and Data Engineering*, 25(1), 220-232.

Zachariadis, V., Jin, Y. and Hagen-Zanker, A.(2013). Reducing the computational requirements of models of spatial interaction using hierarchical clustering of spatial choices, *AUM 2013*, Cambridge, UK.

Zhan, X., Hasan, S., Ukkusuri, S. and Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C*, 33, 37-49.

Zheng, Y. and Xie, X. (2011). Learning travel recommendations from user-generated GPS traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1), 1-29.

Zheng, Y., Liu, L., Wan,g L. and Xie, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the web. In *Proceedings of the 17th International conference on World Wide Web*, 247-256, ACM New York.

Zheng, Y., Zhang, L., Xie, X. and Ma, W.Y. (2009). Mining interesting locations and travel sequences from GPS trajectories. *Proceedings of the 18th International conference on World Wide Web*, 791-800, ACM New York, NY.

Zhou, C., Frankowski, D., Ludford P., Shekhar, S., and Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transportation on Information System (TOIS)*, 25(3), 1-31.

Zhu, B., Huang, Q., Guibas, L. and Zhang, L. (2013). Urban Population Migration Pattern Mining Based on Taxi Trajectories. Paper presented at the 3rd International Workshop on Mobile Sensing, April 2013, USA.

Zou, L., Xu, J.M. and Zhu, L.X. (2005). Arterial speed studies with taxi equipped with global positioning receivers as probe vehicle. *Proceedings of the 2005 International Conference on Wireless Communications, Networking and Mobile Computing*, 1343-1347.

Resources

Cloudmade - Open Street Map (OSM) vector data. Available online at: http://downloads.cloudmade.com/ [Accessed 13 April 2013].

Open Street Map, 2013. Open Database License. Available online at: http://www.OpenStreetMap.org [Accessed 13 April 2013]. For more information see: http://www.openstreetmap.org/copyright.

T-Drive trajectory data sample, Microsoft Asia. Available online at: http://research.microsoft.com/apps/pubs/?id=152883 [Accessed 21 May 2013]