Transportation Research Part C 26 (2013) 78–98

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A probabilistic map matching method for smartphone GPS data ☆

Michel Bierlaire, Jingmin Chen *, Jeffrey Newman

Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 19 November 2010
Received in revised form 2 August 2012
Accepted 3 August 2012

Keywords:
GPS data
Probabilistic map matching
Path observation generation
Network-free data
Route choice modeling

ABSTRACT

Smartphones have the capability of recording various kinds of data from built-in sensors such as GPS in a non-intrusive, systematic way. In transportation studies, such as route choice modeling, the discrete sequences of GPS data need to be associated with the transportation network to generate meaningful paths. The poor quality of GPS data collected from smartphones precludes the use of state of the art map matching methods. In this paper, we propose a probabilistic map matching approach. It generates a set of potential true paths, and associates a likelihood with each of them. Both spatial (GPS coordinates) and temporal information (speed and time) is used to calculate the likelihood of the data for a specific path. Applications and analyses on real trips illustrate the robustness and effectiveness of the proposed approach. Also, as an application example, a Path-Size Logit model is estimated based on a sample of real observations. The estimation results show the viability of applying the proposed method in a real route choice modeling context.

1. Introduction

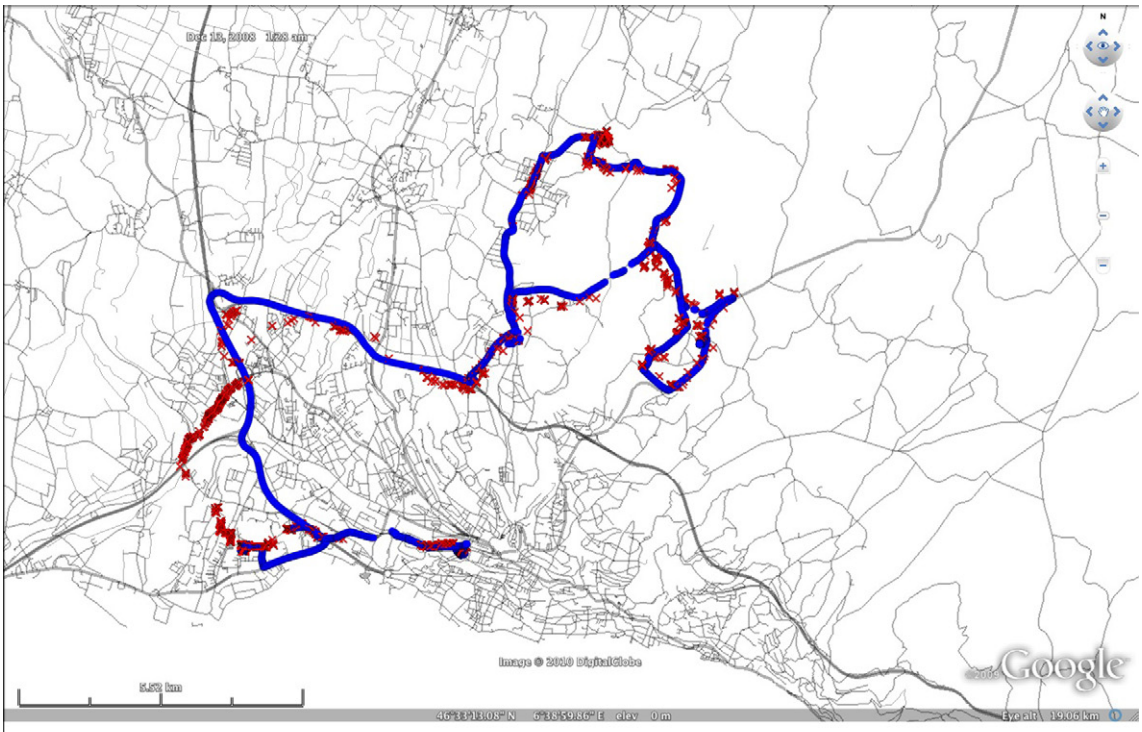
Developing technology has long been harnessed to supplement or replace parts of travel behavior surveys. Tools such as GPS devices have been used to track movements of individuals in a systematic way, instead of relying merely on travel diaries and prompted recall questioning. Tracking survey participants using a specialized GPS device provides some challenges. In particular, people may forget to charge the device, or leave it at home. Nowadays, many people carry a wireless phone. They already manage the tasks of charging and remembering to carry it, at least as well as for any special survey device. Therefore, we propose, as in Stopher (2008), to bundle the survey data collection into a phone.

An important feature of most GPS capable cell phones is Assisted-GPS, which reduces warm-up time for getting the first GPS reading to seconds. This advantage provides more opportunities to observe full tracks of the user's trips without losing the beginning parts of trips. However, the GPS device consumes a great deal of energy. Due to practical constraints, such as limited phone storage space and expensive data transmission cost, data cannot be recorded at a high rate. In our experiments, we use a time interval of 10 s. Also, the data is not as accurate as those collected from dedicated GPS devices. For instance, in the Nokia N95 model used for our experiments, the GPS antenna is embedded under the keyboard, which is generally covered by the screen when the phone is not being actively used. Furthermore, most people carry the cell phone in their pocket or handbag. This weakens the GPS signal.

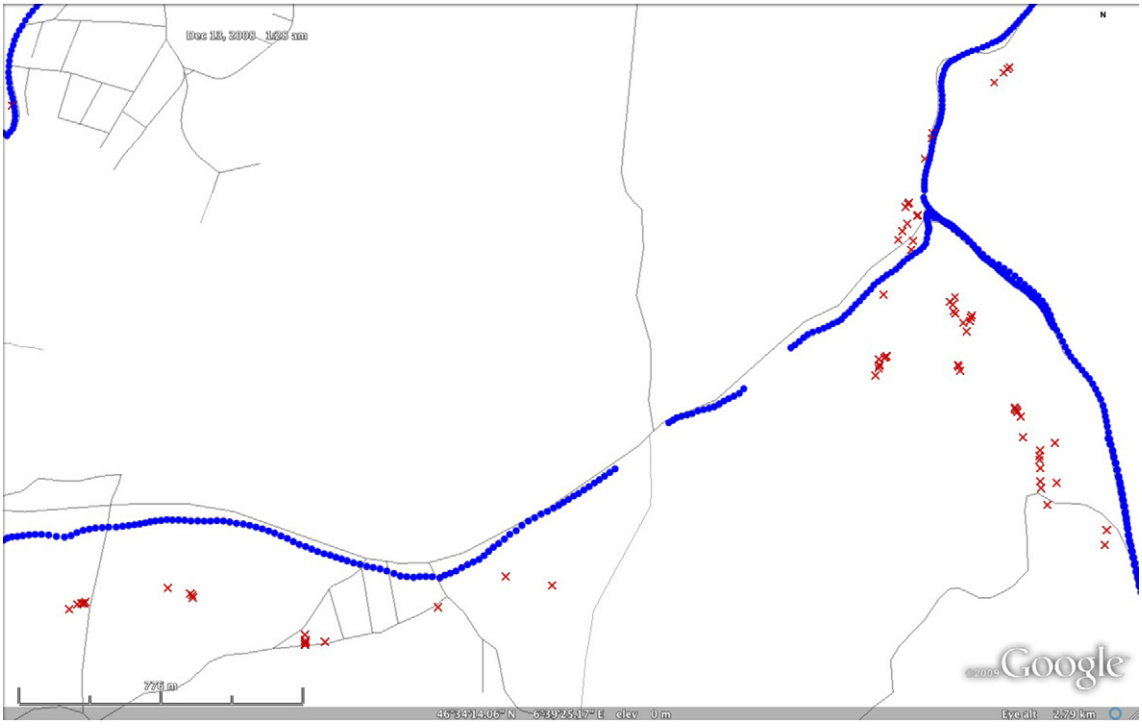
We conducted an experiment where a N95 cell phone and a dedicated GPS device (a MobilityMeter, of the type used by Flamm et al. (2007)), were both carried by the same person during a day. Both devices are configured to record GPS fixes with 1 s interval. The two tracks are reported in Fig. 1, where the blue circles (appearing darker on a black-and-white copy) represent the tracks provided by the MobilityMeter, and the red x's (appearing lighter on a b&w copy) represent the tracks

☆ This research is supported by the Swiss National Science Foundation Grant 200021/131998 – Route choice models and smart phone data.
* Corresponding author. Tel.: +41 79 59 32 532.
E-mail address: jingmin.chen@epfl.ch (J. Chen).

0968-090X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2012.08.001



(a) in a region



(b) zoom in

Fig. 1. GPS traces from N95 and a GPS device.

provided by the N95 smartphone. Fig. 1a shows 6083 points from N95, and 12165 points from MobilityMeter. The availability rate of N95 is 88.7%, while that of MobilityMeter is 99.0%.[1] Throughout the paper, the transportation network data used is provided by OpenStreetMap (www.openstreetmap.org), which is an open source map data service. Although statistical investigation of GPS data accuracy (e.g. Blewitt et al., 1992) is out of the scope in this paper, we can still observe from the visualization that, intuitively, the MobilityMeter GPS data are more consistent with each other in terms of continuity, while N95 GPS data are more scattered. Also, the MobilityMeter GPS data seem closer to the roads.

The raw GPS data are usually matched to the transportation network in order to be useful for many applications. In particular, navigation systems motivate the study of such map matching (MM) techniques. A comprehensive review of 35 MM algorithms for navigation applications since 1989 is presented by Quddus et al. (2007). And a validation strategy for MM algorithms is proposed by Quddus et al. (2005). Since they are designed for navigation applications, current MM algorithms aim at providing on-line deterministic identification of the real road/arc from a single GPS point. However, they do not guarantee that detected roads are connected to form a meaningful path, even if some MM algorithms (e.g. Greenfeld, 2002; Ochieng et al., 2003) do consider connectivity and contiguity of the arcs. In some transport studies where on-line identification is not required and intensive computation is allowed, researchers are also interested in the actual path for the whole trip. For examples, "path" observations are the input for route choice models (Bierlaire and Frejinger, 2008); some novel navigation techniques learn "routing" strategies from GPS data recorded from experienced road network users (e.g. (Yuan et al., 2010)); and "route" travel time can be estimated from GPS data recorded from floating cars in the transportation network (e.g. (Ebendt et al., 2010)).

The adaptation of multiple hypotheses technique (MHT) (Pyo et al., 2001) in MM enables modelers to generate a connected path from a GPS trace representing geographical locations during a trip. Several algorithms (e.g. Marchal et al., 2005; Schuessler and Axhausen, 2009a) maintain at each GPS point a set of path candidates. For each candidate, a score is calculated based on the dissimilarity between GPS points and arcs in terms of distance, speed and/or heading difference, though heading was found to be unreliable for this application (Schuessler and Axhausen, 2009a). The work by Schuessler and Axhausen (2009a) focuses on the computational efficiency of the MM method, and shows excellent results along that line, with dense and accurate GPS data. However, from the experiments that we have conducted (see Section 3.4), it appears that the method is not suitable for smartphone data, where the focus should be in managing the inaccuracy and low density of the data.

An integrated particle filter modeling framework for detecting transportation modes and traveling roads is proposed by Liao et al. (2007). In their approach, a *state* combines various mobility patterns, including the transportation mode and the current road. A Rao–Blackwellized particle filter is used as the framework, while the probability of the traveler switching from one mode to another depends on his proximity to available transportation facilities. A Kalman filter is used to model the dynamic process of traveling on the network and retrieving the GPS fix. In order to fit in the Kalman filter framework, a great deal of simplification is required.

The best MM techniques for sparse data generate a unique best fitting path, but in some applications a unique path is not required. One such application is route choice modeling with network-free data, as presented by Bierlaire and Frejinger (2008). They have introduced an estimation procedure for route choice models that accepts a probabilistic representation of the observed paths, accounting for errors in measurement. An observation does not need to be a unique path, but can be represented by a set of potential paths, along with a probability that the location measurements are indeed recorded from each path. Probabilistic MM algorithms in the literature rely on Dead Reckoning equipping cars or other sensors that smartphones do not embed (e.g. Ochieng et al., 2003). The scores calculated by MM algorithms with MHT techniques, while often heuristically effective, in general lack the theoretical foundation necessary to serve as the probabilities that the corresponding paths are the true path. The simplicity of the score calculation cannot ensure its correctness if there are outlier observations. Moreover, in such a post-processing algorithm (as opposed to real time algorithm for navigation tools), "inaccurate" data is eliminated in the process of data filtering (Schuessler and Axhausen, 2009b), with the risk that some useful information is also excluded.

This paper proposes and implements an advanced and practical probabilistic MM algorithm. It takes advantage of both geographical and temporal information in GPS data to measure the likelihood that the data has been generated along a given path. The likelihood measurement accounts for the inaccuracy of both the smartphone GPS data and the representation of the underlying transportation network. The proposed path generation procedure is capable of dealing with the sparsity of the smartphone GPS data. This method can reduce the impact of noise in GPS readings, and provides probabilistic path observations for further applications.

The next section introduces the GPS data recorded from the smartphones, and the context where the data was recorded. Section 3 derives the probabilistic measurement model for measuring the likelihood that a GPS trace is recorded while traveling on a path. This model relies on a network performance model. Although stand-alone traffic simulators can be used, a simple traffic model using only information available from the GPS records is presented. The probabilistic measurement model is illustrated on some example paths with a real smartphone GPS trace. Potential paths need to be generated before their likelihoods can be calculated. As MM algorithms are not suitable for the smartphone GPS data, a new path generation

[1] Warming time is not accounted in calculating the availability rate. If a device does not record data in more than 10 min, it is considered as 'off' and this time period is not accounted in calculating the availability rate.

algorithm, accounting for the sparsity of the smartphone GPS data, is proposed in Section 4. The proposed approach is applied on 25 real smartphone GPS traces, and some examples are illustrated. In Appendix C we perform sensitivity analyses on model parameters and GPS sampling interval. Some conclusions are included in Section 5. In Appendix B, we apply the proposed method to a smartphone user's GPS data, and model his driving route choice behavior from generated path observations.

2. Context and data

Let $G = (N,A)$ denote a transportation network, where N is the set of all nodes and A is the set of all arcs. The horizontal position of each node $n \in N$ is represented by $x_n = \{\text{lat}, \text{lon}\}$, which is a pair of coordinates consisting of latitude and longitude. The shape of the physical route of arc a is described by an application

$$\mathcal{L}_a : [0,1] \rightarrow \mathbb{R}^2. \tag{1}$$

For a point on the arc, its position x is generated from a unique number ℓ between 0 and 1 such that $x = \mathcal{L}_a(\ell)$. In particular, $\mathcal{L}_a(0)$ is the coordinates of the up-node, and $\mathcal{L}_a(1)$ is the coordinates of the down-node of arc a. For example, if the arc is a straight line from node u to node d, then

$$\mathcal{L}_a(\ell) = (1-\ell)x_u + \ell x_d. \tag{2}$$

Indeed, straight lines are used in transport network data to represent arcs in practice. The performance of the network is characterized by a model

$$x = S(x^-, t^-, t, p) \tag{3}$$

predicting the position x at time t of an individual at position x^- at time t^-, and following path p. It is a random variable with probability distribution function

$$f_x(x|x^-, t^-, t, p). \tag{4}$$

Typically, this model is obtained from a calibrated traffic simulator. However, for practical purposes, analytical models can also be used (see Section 3.3).

Location data is recorded by devices which are carried by travelers when they are traveling on the transportation network. The device makes location measurements combining various sensors such as GPS readings, GSM cell tower information, WLAN base stations, etc. We denote one measurement by

$$\hat{g} = (\hat{t}, \hat{x}, \hat{\sigma}^x, \hat{v}, \hat{\sigma}^v, \hat{h}),$$

which is a tuple containing: $\hat{t}$, a time stamp; $\hat{x} = (\hat{x}_{\text{lat}}, \hat{x}_{\text{lon}})$, a pair of coordinates; $\hat{\sigma}^x$, the standard deviation of the horizontal error in the location measurement; $\hat{v}$, a speed measurement (km/h) and, $\hat{\sigma}^v$, the standard deviation of the error in that measurement; $\widehat{h}$, a heading measurement, that is the angle to the north direction, from 0 to 359, clockwise. We assume that the data has been preprocessed so that we have access to a sequence of measurements $(\hat{g}_1, \ldots, \hat{g}_T)$ corresponding to a given trip.

The experiments described in this paper use smartphone data collected from Nokia N95 smartphones. Dataset A, presented in Section 3.4, contains only one GPS trace with 10 points. It has been collected by one of the authors, with known true path; Dataset B, presented in Appendix C, contains 25 GPS traces and has been collected by 3 anonymous individuals, without known true paths. These GPS traces are recorded while the users are traveling in urban and outskirt areas. The GPS sampling interval is set to be 10 s. Both datasets are produced by a large smartphone data collection campaign in Switzerland, which has been conducted since September 2009 by Nokia Research Center in Lausanne, Ecole Polytechnique Fédérale de Lausanne (EPFL), and IDIAP Research Institute, Switzerland. About 180 participants continuously collect smartphone data for research purposes in 2 years. The details about the data collection campaign is described in Kiukkonen et al. (2010).

3. Probabilistic measurement model

In this section, we focus on the derivation of the probabilistic measurement model for a set of GPS data. More precisely, we compute the probability of all the observed GPS points $(\hat{g}_1, \ldots, \hat{g}_T)$ given a hypothetical path p:

$$\Pr(\hat{g}_1, \ldots, \hat{g}_T | p).$$

This probability is an essential input for the network-free data modeling approach (Bierlaire and Frejinger, 2008). In this section, we introduce a new modeling framework to derive this model and its components.

3.1. Measurement equations

We now derive the probability that a given path p generates the data $(\hat{g}_1, \ldots, \hat{g}_T)$. For the sake of simplification, we focus on the measurement equation for the locations $(\hat{x}_1, \ldots, \hat{x}_T)$, that is

$$\Pr(\hat{x}_1, \ldots, \hat{x}_T | p), \tag{5}$$

which is decomposed recursively:

$$\Pr(\hat{x}_1,\ldots,\hat{x}_T|p) = \Pr(\hat{x}_T|\hat{x}_1,\ldots,\hat{x}_{T-1},p)\Pr(\hat{x}_1,\ldots,\hat{x}_{T-1}|p). \tag{6}$$

The recursion starts with the model $\Pr(\hat{x}_1|p)$:

$$\Pr(\hat{x}_1|p) = \int_{x_1\in p} \Pr(\hat{x}_1|x_1,p)\Pr(x_1|p)dx_1, \tag{7}$$

where the integral spans all locations x_1 on path p. For the first point, if we do not have any prior on the location, $\Pr(x_1|p)$ is a constant equal to the inverse of the length L_p of p. The model $\Pr(\hat{x}_1|x_1,p) = \Pr(\hat{x}_1|x_1)$ describes the measurement error of the smartphone device.

It is generally assumed that the errors in longitudinal and latitudinal directions (e_{lon} and e_{lat} respectively) are independently normally distributed (van Diggelen, 1998). Therefore, the distance between the true location and the measured coordinates $e = \sqrt{e_{lon}^2 + e_{lat}^2}$ follows a Rayleigh distribution. The probability that coordinates $\hat{x}_1$ is recorded from a location x_1 is defined as the probability that the distance between x_1 and $\hat{x}_1$ is less than the true error. Then, we have

$$\Pr(\hat{x}_1|x_1) = \Pr(e > \|\hat{x}_1 - x_1\|_2) = \exp\left(-\frac{\|\hat{x}_1 - x_1\|_2^2}{2\hat{\sigma}^2}\right). \tag{8}$$

As the variance σ^2 is unknown, we use $\hat{\sigma}^2 = \sigma^2_{\text{network}} + \left(\hat{\sigma}_1^x\right)^2$ as an estimate, where $\sigma^2_{\text{network}}$ captures the difference between the coded network and the actual roads and paths, and $\left(\hat{\sigma}_1^x\right)^2$ captures the measurement error of the GPS device. Quddus et al. (2005) explain that errors in network data effect the quality of the MM results, therefore the network error parameter σ_{network} is introduced here.

Combining (7) and (8), we obtain

$$\Pr(\hat{x}_1|p) = \frac{1}{L_p}\int_{x_1} \exp\left(-\frac{\|\hat{x}_1 - x_1\|_2^2}{2\hat{\sigma}^2}\right)dx_1. \tag{9}$$

The next step of the recursion derives

$$\Pr(\hat{x}_1,\hat{x}_2|p) = \Pr(\hat{x}_2|\hat{x}_1,p)\Pr(\hat{x}_1|p), \tag{10}$$

where $\Pr(\hat{x}_1|p)$ is defined by (9). We write

$$\Pr(\hat{x}_2|\hat{x}_1,p) = \int_{x_2\in p} \Pr(\hat{x}_2|x_2,\hat{x}_1,p)\Pr(x_2|\hat{x}_1,p)dx_2. \tag{11}$$

The first term in (11), $\Pr(\hat{x}_2|x_2,\hat{x}_1,p) = \Pr(\hat{x}_2|x_2)$, is again modeling the measurement error of the device, and can also be defined by (8), combined with the same simplifications as described above. The second term predicts the position at time $\hat{t}_2$ of the traveler. It is written as

$$\Pr(x_2|\hat{x}_1,p) = \int_{x_1\in p} \Pr(x_2|x_1,\hat{x}_1,p)\Pr(x_1|\hat{x}_1,p)dx_1. \tag{12}$$

The first term in (12) models the movement of the traveler, which is captured by (3), that is

$$\Pr(x_2|x_1,\hat{x}_1,p) = f_x(x_2|x_1,\hat{t}_1,\hat{t}_2,p),$$

where f_x is the density function (4) of the traffic model. The second term can be derived from Bayes rule:

$$\Pr(x_1|\widehat{x_1},p) = \frac{\Pr(\hat{x}_1|x_1,p)\Pr(x_1|p)}{\int_{x_1}\Pr(\hat{x}_1|x_1,p)\Pr(x_1|p)dx_1}.$$

As $\Pr(x_1|p) = 1/L_p$ is constant for a given p, we have

$$\Pr(x_1|\hat{x}_1,p) = \frac{\Pr(\hat{x}_1|x_1,p)}{\int_{x'_1\in p}\Pr(\hat{x}_1|x'_1,p)dx'_1} \tag{13}$$

which is a normalized version of (8). This completes the definition of (10).

The recursion in (6) requires that, at iteration k, the probability

$$\Pr(\hat{x}_k|\hat{x}_1,\ldots,\hat{x}_{k-1},p)$$

is calculated. It can be generalized from (11) and (12) that

$$\begin{aligned}\Pr(\hat{x}_k|\hat{x}_1,\ldots,\hat{x}_{k-1},p) = {} & \int_{x_k}\Pr(\hat{x}_k|x_k,\hat{x}_1,\ldots,\hat{x}_{k-1},p)\\ & \int_{x_{k-1}}\Pr(x_k|x_{k-1},\hat{x}_1,\ldots,\hat{x}_{k-1},p)\Pr(x_{k-1}|\hat{x}_1,\ldots,\hat{x}_{k-1},p)d_{x_{k-1}}dx_k,\end{aligned} \tag{14}$$

where $\Pr(\hat{x}_k|x_k,\hat{x}_1,\ldots,\hat{x}_{k-1},p)=\Pr(\hat{x}_k|x_k)$ is given by (8), and $\Pr(x_k|x_{k-1},\hat{x}_1,\ldots,\hat{x}_{k-1},p)=\Pr(x_k|x_{k-1},p)$ is the traffic model $f_x(x_k|x_{k-1},\hat{t}_{k-1},\hat{t}_k,p)$. The last part of (14), $\Pr(x_{k-1}|\hat{x}_1,\ldots,\hat{x}_{k-1},p)$, is the posterior pdf of the true location x_{k-1} given observed GPS trace $\hat{x}_1,\ldots,\hat{x}_{k-1}$ and path p. This distribution is not tractable, and we must simplify it, and replace it by

$$\Pr(x_{k-1}|\hat{x}_1,\ldots,\hat{x}_{k-1},p)\approx\Pr(x_{k-1}|\hat{x}_{k-1},p). \tag{15}$$

Therefore, we can use the same derivation that leads to (13) to obtain

$$\Pr(x_{k-1}|\hat{x}_{k-1},p)=\frac{\Pr(\hat{x}_{k-1}|x_{k-1},p)}{\int_{x'_{k-1}\in p}\Pr(\hat{x}_{k-1}|x'_{k-1},p)dx'_{k-1}}. \tag{16}$$

The derivation above involves many integrals over the full path. Although these integrals have low dimension, they can be cumbersome to compute, especially when the path p is long. In Section 3.2, we describe how to decompose the integrals, and to use the concept of Domain of Data Relevance (DDR) introduced by Bierlaire and Frejinger (2008) to simplify the computation.

3.2. Computing integrals

The measurement equations involve various integrals along a path p of the form

$$I=\int_{x\in p}f(x)dx, \tag{17}$$

that are complicated to compute in real applications. We describe here how to exploit the topology of the network to compute these integrals.

First, we decompose the path into arcs to obtain

$$I=\sum_{a\in p}\int_{x\in a}f(x)dx. \tag{18}$$

For each arc, we use the shape model (1) to obtain a unidimensional integral

$$\int_{x\in a}f(x)dx=\int_{\ell=0}^{1}f(\mathcal{L}_a(\ell))|\partial\mathcal{L}|d\ell, \tag{19}$$

where

$$|\partial\mathcal{L}|=\sqrt{\left(\frac{d(\mathcal{L}_a(\ell))_{\text{lat}}}{d\ell}\right)^2+\left(\frac{d(\mathcal{L}_a(\ell))_{\text{lon}}}{d\ell}\right)^2}. \tag{20}$$

For example, if the linear model (2) is used, we have

$$|\partial\mathcal{L}|=\|x_u-x_d\|_2. \tag{21}$$

Second, we truncate the domain of the integrals to save computation time where negligible quantities are involved. For a given GPS observation $\hat{x}$, Bierlaire and Frejinger (2008) define the DDR as the physical area where the piece of data is relevant. In our context, a point x is considered to be in the DDR of $\hat{x}$ if following conditions are satisfied:

- the probability $\Pr(\hat{x}|x)$ is above a given threshold θ;
- if $\hat{v}>10$ km/h, the difference between the GPS heading and the arc direction is less than 60°. The arc direction is approximated as the direction from its up node to down node.

In our implementation, we have used a value $\theta=0.65$. It corresponds roughly to points in a diameter of 100 m when the σ parameter of the GPS device is 100 m, and the σ for the network coding is assumed to be 30 m. Indeed,

$$\exp\left(-\frac{\|\hat{x}-x\|_2^2}{2\hat{\sigma}^2}\right)\geqslant\theta$$

is equivalent to

$$\|\hat{x}-x\|_2\leqslant\sqrt{-2(\hat{\sigma})^2\ln\theta},$$

and the upper bound 96.9 is obtained with $\theta=0.65$ and $\hat{\sigma}=104.4=\sqrt{100^2+30^2}$. This is illustrated in Fig. 2, where the parts of arcs AB and AC represented by a solid red line are inside the DDR of the data point $\hat{x}$.

Clearly, the value of the parameters should be adjusted to account for the features of the relevant application, and the quality of the associated data. Also, the complexity of the computation of the integrals increases with the size of the DDR. A large DDR means more computation. On the other hand, too small a DDR may artificially produce a zero probability

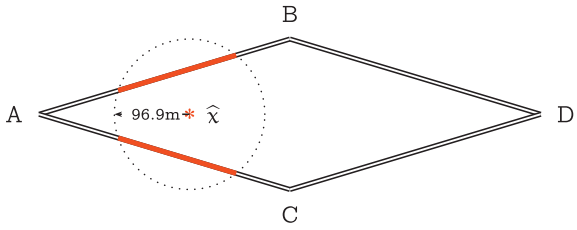


Fig. 2. Domain of Data Relevance.

for the measurement equation, which is undesirable. As discussed by Bierlaire and Frejinger (2008), the specification of the DDR should correspond to a good trade-off between accuracy and computational burden. Some sensitivity analyses regarding these parameters are performed in Appendix C.

3.3. Traffic model

In our framework, the traffic model is designed to predict the position of the GPS device over time. More precisely, it predicts the position x of the device at time t if the position at time t^- is x^-, and the device is traveling along path p.

This is the typical role of dynamic traffic simulators (such as AIMSUN (Barceló and Casas, 2005), MITSIM (Yang and Koutsopoulos, 1996), DynaMIT (Ben-Akiva et al., 2001), Dynasmart (Mahmassani, 2001), among many). However, it is not always practical to use a calibrated traffic simulator in a MM context. Therefore, we suggest to use a simple analytical model such as the one described below.

First, we define the operator that computes the distance between two points x and y lying on path p, and denote it by

$$d_p(x, y). \tag{22}$$

This operator is easily implemented using the same decomposition of paths into arcs described in Section 3.2. We write the traffic model in terms of speed instead of position, considering the random variable

$$v = \frac{d_p(x^-, x)}{t - t^-} \tag{23}$$

with pdf

$$f_v\left(\frac{d_p(x^-, x)}{t - t^-}\right). \tag{24}$$

In our experiments, the traveling speed of the device is recorded every 10 s, therefore its distribution can be derived from the observed speed data. For the distribution of speed, we assume a mixture of a negative exponential distribution and a log normal distribution. The first is designed to capture the instances where the vehicles are stopped at intersections, or traveling at low speed before or after that stop. The second is designed to capture vehicles moving at regular speed. The distribution is

$$f_v(v) = w\lambda \exp^{-\lambda v} + (1 - w)\frac{1}{v\sqrt{2\pi\tau^2}} \exp^{-\frac{(\ln v - \mu)^2}{2\tau^2}}, \tag{25}$$

where w (the weighting), λ (the scale parameter of the negative exponential distribution), μ (the location parameter of the log normal distribution), and τ (the scale parameter of the log normal distribution) are parameters to be estimated. Dataset B, including 1041 GPS records, is used for the estimation. Following are some statistics of dataset B: total number of GPS points (1041); number of GPS points per trace (minimum: 16, mean: 35.9, maximum: 53); duration of the trip (minimum: 180 s, mean: 387 s, maximum: 795 s). Fig. 3 shows the normalized histogram of the recorded speed data and the estimated speed distribution. Table 1 reports the parameters estimated by maximum likelihood.

3.4. Illustration

We illustrate likelihood results for 4 example paths associated with a real GPS trace (dataset A) recorded from N95 smartphone. The 4 paths are shown in Fig. 4 as red solid lines. The GPS points are also shown in each figure as blue points. The direction of the trajectory is illustrated by the arrow besides each path and by the GPS points's annotation with their orders being recorded. This particular trip is chosen to be analyzed because it is recorded while traveling by car in a dense transportation network. And the path is known with certainty as the traveler was one of the authors.

Fig. 4a shows the true path. If we look at the GPS points, the ambiguity of the coordinates readings and the density of the transportation network makes the actual path difficult to be recognized from the N95 data alone. It can be observed that some of the GPS points (e.g. 7 and 8) deviate more than 30 meters from the actual path. Consequently, another path shown in Fig. 4b also seems intuitively reasonable enough to be the actual path if we only compare the geographical dissimilarities.

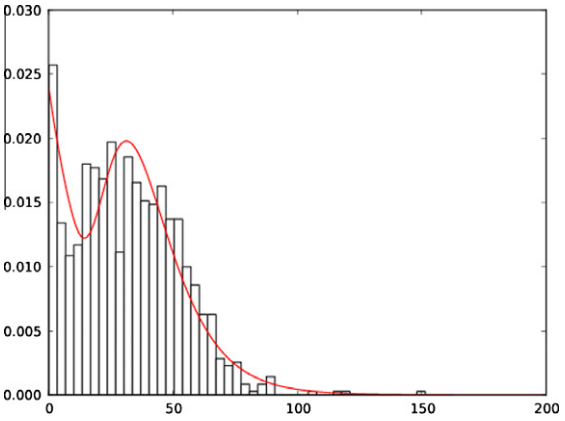


Fig. 3. The speed distribution.

Table 1
Parameters estimates for the speed distribution.

Parameter	Estimate	Standard error
w	0.423	0.0636
λ	0.057	0.0097
μ	3.672	0.0314
τ	0.396	0.0282

Parameters estimated by *R*.

Another path candidate shown in Fig. 4c is intuitively less possible to be the actual path. The last one (Fig. 4d seems very problematic, but is actually generated by the deterministic MM algorithm developed by Schuessler and Axhausen (2009a) (without using speed penalty term in the score function).

We calculate the natural logarithm of the measurement likelihood (5), termed the measurement loglikelihood, for all paths,[2]

$$\ln \Pr(\hat{x}_1, \ldots, \hat{x}_T | p). \tag{26}$$

We notice that the real path gains the highest loglikelihood, −14.1. The loglikelihood is lower for paths that are intuitively less possible to be the true one (−15.7 and −16.3 for path 2 and path 3 respectively). The value for path 4 is $-\infty$, because the path does not pass through DDR of some GPS points (e.g. the last one).

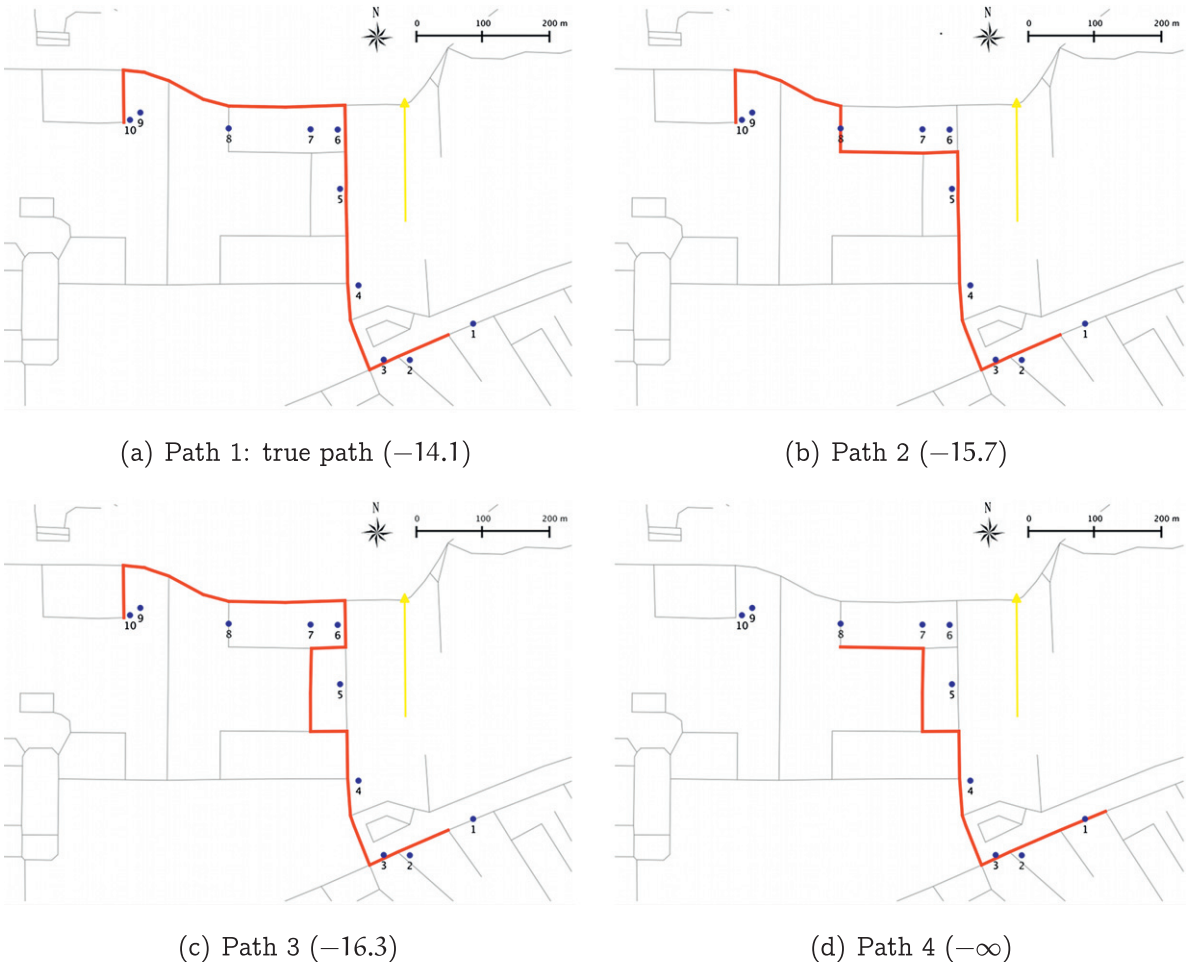
Path 4 generated by the deterministic MM algorithm seems strange due to the incapability of the algorithm to deal with sparse data. In fact, the beginning of the path is correctly identified. The path terminates earlier than the real destination because the number of arcs in the path is constrained by the algorithm not to be higher than the number of GPS points. Indeed, the matched path has exactly 10 arcs, and it gains the lowest MM score among all paths with not more than 10 arcs. The drawback of this algorithm is described in Section 4 in details. In fact, if the GPS data has higher density, the deterministic MM algorithm may be able to identify the true path. For example, Fig. 5 shows the MobilityMeter data recorded at the same time. The deterministic MM result produced by Schuessler and Axhausen (2009a, 's) algorithm is the true path with this data.

4. Path generation algorithm

For a sequence of GPS data, the method presented in Section 3 assigns a likelihood to a given path p. We focus now on generating candidate true paths.

State of the art MM algorithms are designed for dense data, where it can be safely assumed that nearly every arc on a path generates at least one GPS point. For instance, Marchal et al. (2005) and Schuessler and Axhausen (2009a) generate path candidates by considering each GPS point one by one in the chronological order. At each iteration k, they generate a set P_k of path candidates assumed to match the GPS points up to k. These paths are generated by topologically extending the paths in P_{k-1} by not more than one arc. Hence, in order to allow for correctly identifying the true path, the GPS device has to record at least

[2] If we further expand (6), the measurement likelihood (5) becomes $\Pr(\hat{x}_1, \ldots, \hat{x}_T | p) = \Pr(\hat{x}_1 | p) \prod_{k=2}^{T} \Pr(\hat{x}_k | \hat{x}_1, \ldots, \hat{x}_{k-1}, p)$, which is the multiplication of many probability values that are smaller than 1. Consequently, the measurement likelihood (5) is close to zero. Throughout this paper we present the logarithm of it, as it is common for likelihood.



(a) Path 1: true path (−14.1)

(b) Path 2 (−15.7)

(c) Path 3 (−16.3)

(d) Path 4 (−∞)

Fig. 4. Results from N95 GPS data.

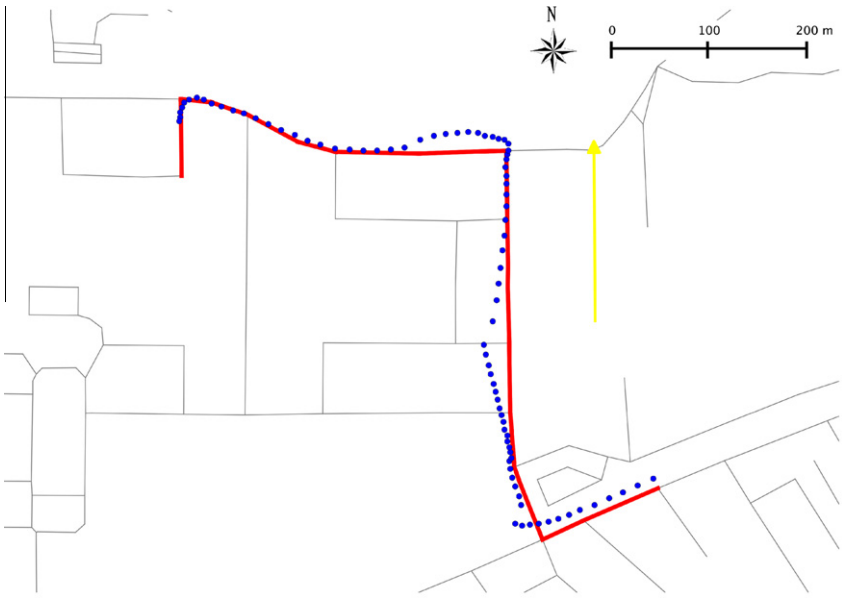


Fig. 5. MobilityMeter data and deterministic MM result.

one GPS point on each arc. It can clearly be observed from Figs. 1, 4 and 5 that the dedicated GPS device data is consistent with the "high density" hypothesis, while the smartphone data is not. Also, the example in Fig. 4d shows that the MM algorithm is not appropriate for smartphone GPS data.

In order to address this problem, we propose a path generation algorithm designed for sparse data. It uses a similar iterative process as the one described above. But the path extension procedure at each iteration is not limited to only one arc. The algorithm ignores GPS points that have a speed lower than 8 km/h, labeled as "stationary". When the device is more or less stationary, while it may generate data that is relevant for comparing path likelihood, it is not generating information that is useful in path extension. Two exceptions are the first and the last GPS points; even if their speed values are low, they reveal information about the origin and the destination. Therefore, they are not labeled as "stationary". This algorithm is described in Algorithm 1. Detailed explanation of some procedures (numbered lines) are given as follows:

Algorithm 1. Path generation algorithm

Input: A GPS trace $(\widehat{g}_1, \ldots, \widehat{g}_T)$ with non-"stationary" GPS points.
Input: The underlying transportation network $G = (N, A)$.
Result: A set of candidate paths P_T.

```
// Deal with the first GPS point.
P_1 ← empty set of paths;
for each arc a ∈ A do
    DDR_1 ← the DDR of the first GPS point;
    if a intersects DDR_1 then
        include a as a path in P_1;

// Iterative process.
for k ← 2 to T do
    P_k ← empty set of paths;
    foreach p ∈ P_{k−1} do
        // Path extension procedure.
        n ← the end node of p;
1       spt ← a bounded shortest path tree rooted at n;
        foreach arc a ∈ spt do
            if a intersects DDR_k then
                sp ← shortest path connecting p and a;
                a_0 ← first arc of sp;
                a_1 ← last arc of p;
                if a_0 ∈ DDR_k or a_0 is not reverse of a_1 then
                    // exclude unreasonable U-turns
                    p_new ← join p, sp and a;
                    include p_new in P_k;
2                   update likelihood Pr(x̂_1, ..., x̂_k | p_new);
    if ||P_k|| > 20 then
3       eliminate some paths from P_k;
```

1. The bound for the shortest path tree is derived from an assumption about the maximum possible speed and the time interval between t_{k-1} and t_k. The leaf nodes of the bounded shortest path tree are the first nodes detected by the Dijkstra algorithm that violate the bound. In our experiments, the bound is defined by $1.5(t_k - t_{k-1})\hat{v}_{\max}$, where $\hat{v}_{\max}$ is the maximum speed value among the observed speeds $\hat{v}_{k-1}$, $\hat{v}_k$, and the speed calculated by $\|\hat{x}_k - \hat{x}_{k-1}\|_2/(t_k - t_{k-1})$. The factor 1.5 is a safety margin to minimize the risk of missing a relevant observation.
2. In the update of likelihood, all the GPS points up to $\hat{g}_k$, including "stationary" points, are included. As explained in Section 3.1, the likelihood (6) is updated recursively and at each iteration, only Eq. (14) needs to be calculated.
3. The path elimination procedure limits the size of P_k at each iteration if it is too large. After many tests, we use 20 as the threshold, which balances the tradeoff between algorithm speed and result effectiveness. It is designed to speed up the algorithm by eliminating less relevant branches produced from the path extension procedure. The path elimination procedure is performed by selecting and keeping following paths:

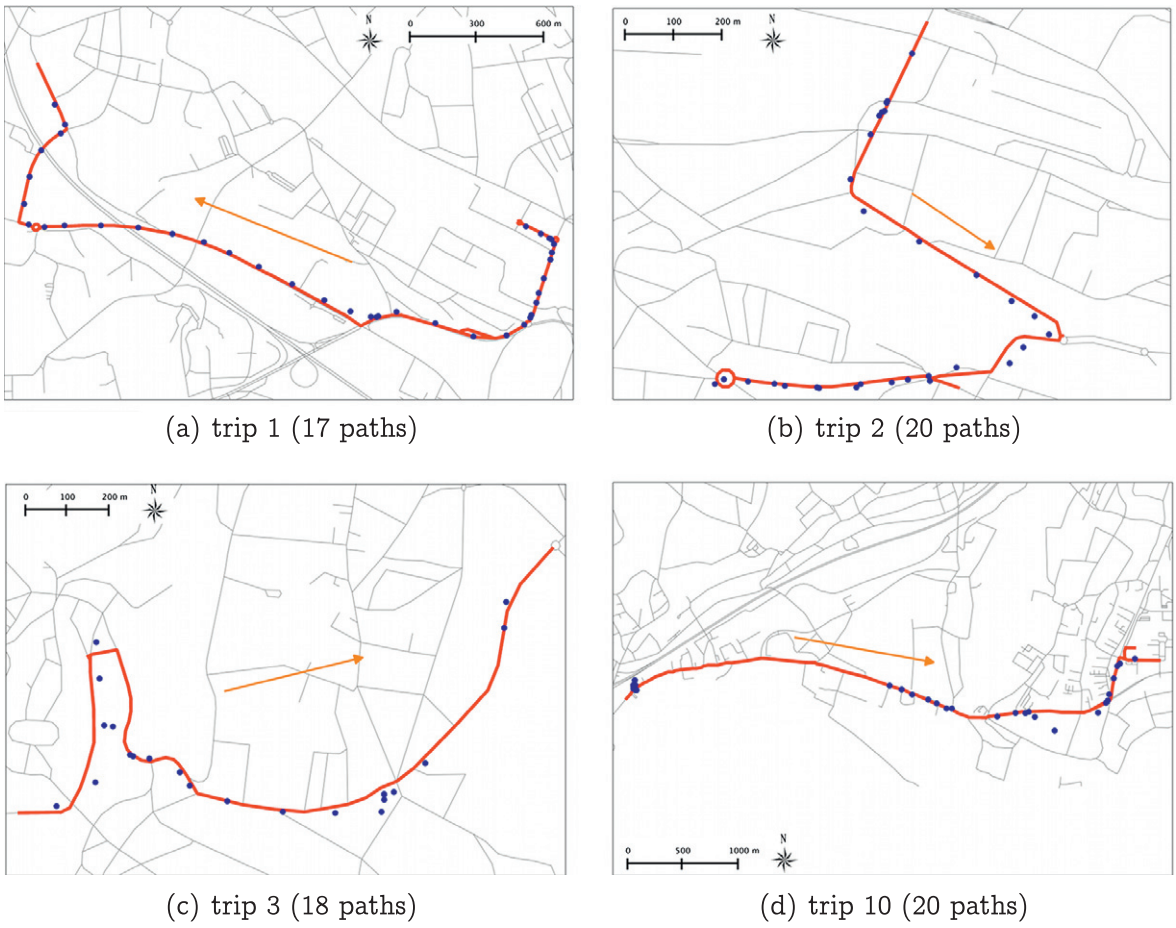


Fig. 6. Examples for some GPS traces.

(a) The 2 shortest paths in P_k are selected.
(b) Paths are randomly selected from P_k according to the likelihood (5). In practice, the likelihood is normalized to obtain scores summing up to one before the random selection. Path candidates are drawn using simulation until the cumulative normalized likelihood exceeds a predefined number (e.g. 0.8).
(c) For each arc $a \in spt_k$ and a intersects DDR_k, P_{ak} is defined as P_k's subset that only contains paths going via a. We then apply a similar simulation procedure as in Step 2 on P_{ak}, but only to draw one path. This is meant to guarantee that each arc associated with the latest GPS point has at least a path in P_k after the elimination procedure.

The algorithm is implemented as a software package in C++. We illustrate some results generated from 25 real GPS traces (dataset B). Fig. 6 shows 4 examples, and 6 more examples are included in Appendix A. Again, each GPS trace is associated with many path candidates, and they overlap to a large extent, as shown on the maps. The results in general look reasonable, as each path is close to its corresponding GPS trace.

For the same trip, the differences of the generated paths show the uncertainty of the probabilistic map matching result. On one hand, the uncertainty is due to the imprecision of the GPS data. On the other hand, most of the uncertainty belongs to the end of the trips. This can be explained by the mechanism of the likelihood model. The likelihood model utilizes the dependency between adjacent GPS points. Each GPS point in fact provides information to help in identifying its upstream trajectory. The end of a trip always gains less information since it has less (or none) downstream GPS points.

5. Conclusions

We propose a probabilistic MM method for matching a set of paths with GPS data. A probabilistic measurement model is derived, which calculates the probability that a GPS recording device would have generated a sequence of measurements while following a given path. It is based on a structural model and a measurement model, which captures the movements and the recordings of the GPS device respectively.

The uncertainty derived from the inaccuracy of both the GPS data and the transportation network is explicitly taken into account. The application to real data shows that the probability values of the actual path and some other paths are realistic and meaningful.

A path generation algorithm is also proposed that accounts for the sparsity of the data. The methodology has been applied on real smartphone data collected in Switzerland. Appendix B presents the estimation of a simple route choice model and it shows that the proposed methodology indeed allows the use of GPS data from smartphones in this context.

In the probabilistic measurement model, some of the parameters' values are based on engineering intuition. These parameters are σ_{network}, the standard deviation of network error; θ, which defines the diameter of the DDR; and the heading constraint (60°) for excluding arcs from the DDR. Sensitivity analyses presented in Appendix C prove the robustness of the proposed probabilistic MM approach with respect to these somehow arbitrary values. A sensitivity analysis is also performed on the sampling interval of the GPS data. The sampling interval, originally 10 s, is increased to be 20, 30 and 60 s, thus the data become sparser. The uncertainty in the path results increases with the sparsity of the GPS data. This is consistent with the intuition that more GPS data brings more information, thus less uncertainty.

Future extensions of this work include investigation into the use of other types of data provided by smartphones, such as the detection of cell towers, WiFi base stations, or other bluetooth devices, as well as physical activity detected by accelerometers, gyroscopes, and magnetometers. Also, the analysis of other travel decisions, such as mode choice, should be considered, similar to the work by Liao et al. (2007). Moreover, we are interested in developing a path generation method based on Monte Carlo Simulation (Flötteröd and Bierlaire, 2011). Finally, the efficiency of the algorithm will have to be adapted to deal with large dataset.

Acknowledgments

This research is supported by the Swiss National Science Foundation Grant 200021/131998 *Route choice models and smart phone data*. We have benefited from discussions with Emma Frejinger and the members of the Nokia Research Center in Lausanne, especially Niko Kiukkonen, whose help in data collection was invaluable.

Appendix A. Path generation examples

Figs. 7–12 show 6 more probabilistic MM results, in addition to the 4 shown in Section 4.

Appendix B. Modeling route choice behavior from real data

The motivation to develop the probabilistic path generation algorithm is the estimation of route choice models. In this section, we illustrate the estimation of the parameters of a route choice model for a given smartphone user from data recorded from his phone while driving. We use the network-free data modeling and the Path Size Logit (PSL) as the route

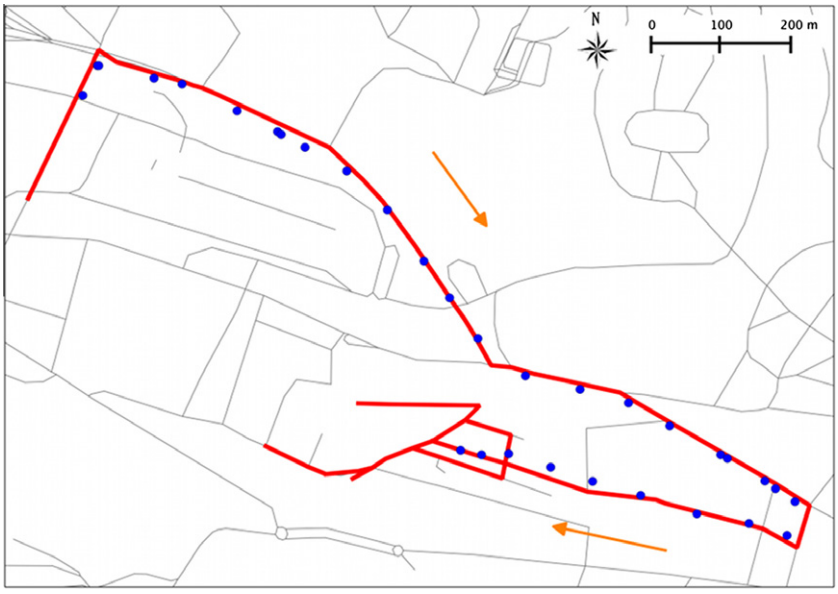


Fig. 7. Trip 3 (29 paths).

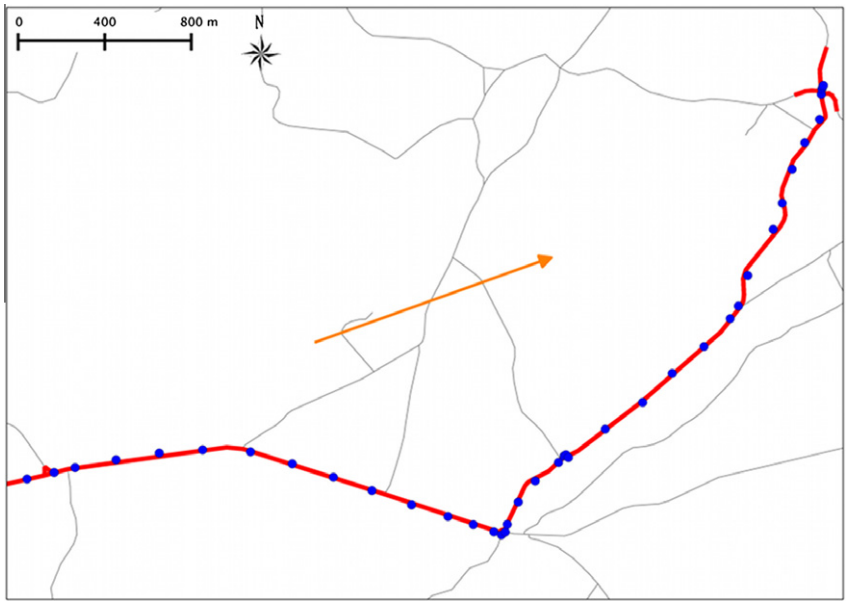


Fig. 8. Trip 5 (22 paths).

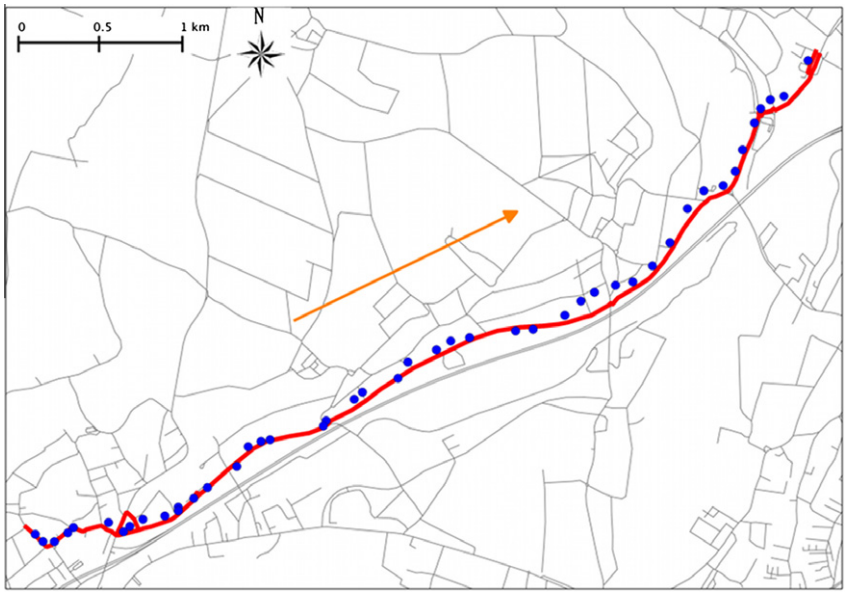


Fig. 9. Trip 6 (6 paths).

choice model (Bierlaire and Frejinger, 2008). For a GPS trace $(\hat{x}_1, \ldots, \hat{x}_T)$ that represents a trip, P and S are the generated potential true paths and OD pairs. The likelihood function for this GPS trace is given by

$$\Pr(\hat{x}_1, \ldots, \hat{x}_T|S) = \sum_{s \in S} \Pr(s|S) \sum_{p \in P^s} \Pr(\hat{x}_1, \ldots, \hat{x}_T|p) \Pr(p|C(s); \beta), \tag{27}$$

where

- S is the set of relevant OD pairs,
- $\Pr(s|S)$ is the probability that the actual OD pair is s. In this study, it is defined as $\Pr(s|S) = 1/|S|$ if $s \in S$, and 0 otherwise;
- $P^s \subseteq P$ is the set of generated path candidates corresponding to OD pair $s \in S$;
- $\Pr(\hat{x}_1, \ldots, \hat{x}_T|p)$ is the GPS measurement likelihood (5) calculated from the proposed method;

Fig. 10. Trip 7(12 paths).

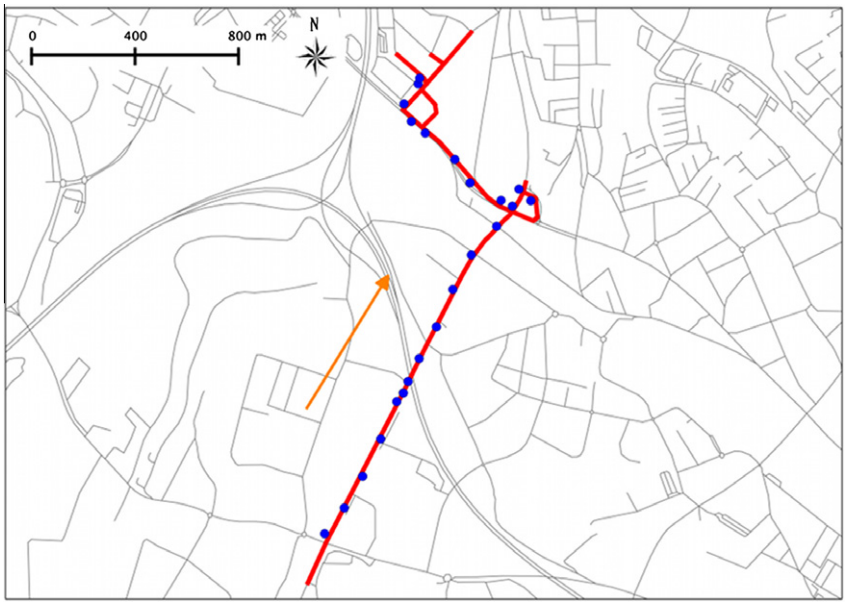


Fig. 11. Trip 8 (13 paths).

- $\Pr(p|C(s);\beta)$ is the route choice model, where $C(s)$ is the choice set for OD pair s, and β are the parameters to be estimated. In this study, a PSL specification is used.

In the following subsections, we focus on the specification and the estimation of the choice model $\Pr(p|C(s);\beta)$, where $p \in P^s$ is a probabilistically chosen path with OD pair s.

B.1. Choice set generation: importance sampling

Choice set generation is an important procedure in route choice modeling. In this study, we employ the stochastic choice set generation algorithm proposed by Frejinger et al. (2009). This method assumes that the relevant choice set $C(s)$ is the set of all possible paths in the network connecting OD pair s. In order to develop a tractable choice set for use in estimating the parameters of the choice model, path alternatives are sampled using a biased random walk algorithm, with arc weights at

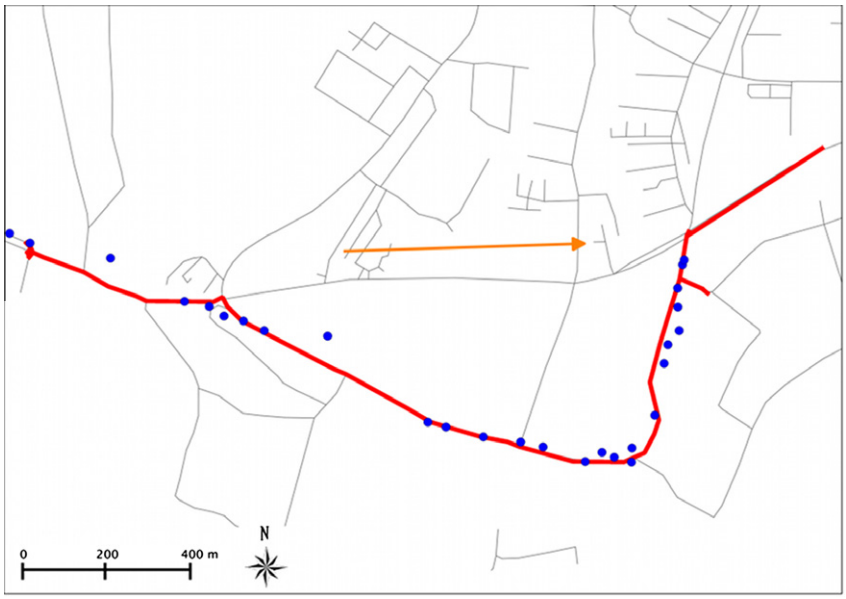


Fig. 12. Trip 9 (36 paths).

each node set by the ratio of the length of the shortest path to the destination using any arc and using the target arc. The sampling bias is subsequently corrected in the choice model.

The random walk procedure as presented in Frejinger et al. (2009) does not allow passing the destination and subsequently returning to it. However, GPS observations show this behavior is not uncommon, as it can represent normal parking search behavior. In order to incorporate the sampling bias correction in the choice model, the positive conditioning property requires that it is at least possible to sample the observed path choice, so we modify the random walk algorithm to allow for this.

Let h be the number of times the random walk algorithm has visited the destination node. At each such visit, the walk terminates with probability $P(h)$, where $P(h)$ is increasing with h, and the walk continues with probability $1 - P(h)$. In our experiments, we used

$$P(h) = 1 - 0.5^h. \tag{28}$$

If the walk proceeds, it does so using the original procedure, although we modify the arc weights for selecting an arc departing the destination, to reflect the fact that the walk has to leave the destination node. Otherwise, the shortest path from the current node (i.e. the destination) to the destination has obviously zero length, resulting in undefined arc selection probabilities. We correct this by simply imposing a condition that a shortest path must contain at least one arc, thus forcing a strictly positive result.

With this modification, the probability $q(p)$ for sampling a path p is now

$$q(p) = P(h_p)\prod_{i=1}^{h_p-1}(1 - P(i))\prod_{a\in\Gamma_p} q(a|\mathcal{E}_a),$$

that is, the probability that the algorithm has been continued $h_p - 1$ times and stopped once. Γ_p is the (ordered) sequence of arcs in path p, $\mathcal{E}_a$ is the list of arcs with the same up-node as a,[3] $q(a|\mathcal{E}_a)$ is the probability for selecting arc a among all arcs in $\mathcal{E}_a$, h_p is the number of times that the destination node is in path p.

To estimate the models presented in the next section, choice set samples were created by generating 50 random walks for comparison against each possible true path, using the method described above with length representing generalized cost and Kumaraswamy parameters $b_1 = 30$ and $b_2 = 1$, plus the observed paths as calculated by the previously described algorithms (see Ben-Akiva and Lerman (1985) for a discussion of model estimation in general).

B.2. Model estimation

We estimate the parameters of a simple model in order to illustrate the procedure. We use 19 real trips recorded from a single user's smartphone. Table 2 presents some statistics about the trips, as well as about the paths generated by the pro-

[3] Note that we use a slightly different notation than Frejinger et al. (2009) to simplify the presentation.

Table 2
Statistics of the recorded 19 trips.

	Min	Average	Max
Number of GPS points per trip	16	36	58
Approximate travel time per trip (s)	179	397	795
Length of the generated paths (km)	1.93	3.98	6.42
Number of traffic signals of the generated paths	0	2.84	5.0

Table 3
Estimation result.

Coefficient	Value	Std. Error	*t*-Test	*p* Value
β_{EPS}	0.242	0.138	4.98	0.00
β_{ℓ}	−33.7	16.4	−5.28	0.00
β_{sg}	−2.74	3.67	−2.39	0.02

Number of observations: 19.
Null log likelihood: −776.1.
Final log likelihood: −708.9.
Adjusted rho-square: 0.083.
Model estimated by BIOGEME (Bierlaire, 2003).

cedure described in Section 4. For each trip, the length and the number of traffic signals are weighted by the normalized measurement likelihood:

$$\Pr(p|\hat{g}_1,\ldots,\hat{g}_T) = \frac{\Pr(\hat{g}_1,\ldots,\hat{g}_T)|p}{\sum_{p'\in P}\Pr(\hat{g}_1,\ldots,\hat{g}_T|p')}. \tag{29}$$

The deterministic term of the utility function of path p in the PSL model is specified as

$$V_p = \beta_{EPS}\ln EPS_p + \beta_{\ell}L_p + \beta_{sg}NbSignals + Corr_p, \tag{30}$$

where *NbSignals* is the number of traffic signals along the path; EPS_p is the Extended Path Size (EPS), which accounts for the path overlapping and corrects for the sampling; $Corr_p$ is the choice set sampling correction term. We refer to Frejinger et al. (2009) for more details about *EPS* and sampling correction.

Table 3 reports the coefficient estimates. All coefficients have their expected signs (positive for the EPS coefficient as is consistent with established route choice theory (Frejinger, 2008), and negative for coefficients on path length and number of traffic signals) and they are all significantly different from zero.

Clearly, the size of the sample is too small to consider this as a final model. However, it illustrates the feasibility of the overall approach on a real data set.

Appendix C. Sensitivity analysis

In the probabilistic measurement model, some of the parameters' values are based on engineering intuition. These parameters are σ_{network}, the standard deviation of network error; θ, which defines the diameter of the DDR; and the heading constraint (60°) for excluding arcs from the DDR. Sensitivity analyses, presented in on these parameters to test the robustness of the proposed probabilistic MM approach to these somehow arbitrary values. A sensitivity analysis is also performed on the sampling interval of the GPS data.

C.1. Experimental design

For any of the above mentioned parameters, although its precise value is not easy to decide, a reasonable bound can be derived based on available information. Therefore, the sensitivity analysis is performed as applying the proposed approach on the same dataset with the parameter's value varying within these bounds, and analyzing how the variation affects the results.

A probabilistic MM result for a GPS trace is a set of paths with associated measurement likelihood values. It contains a lot of information which takes effort to read. Hence, we have to define aggregate indicators for the sensitivity analyses. Intuitively, all paths in a result should be almost the same, as have been illustrated with some examples in the last section. The differences among the paths can be understood as the uncertainty of the result, which is caused by the ambiguity in the GPS data. Hence, we first define an aggregate similarity indicator $O_1(P)$ to measure the overall overlapping of all paths P in a

result. Second, we also want to compare results produced by different parameter values. Therefore, another indicator $O_2(P_1,P_2)$ is defined to measure the similarity (overlapping) between two sets of paths, P_1 and P_2.

We start by defining how one path p overlaps with all paths in a set $P(p \in P)$:

$$O(p,P) = \begin{cases} 1 & \text{if } \|P\| = 1 \\ \sum_{a \in p} \frac{L_a}{L_p} \frac{\sum_{p' \in P} \delta_{ap'} - 1}{\|P\| - 1} & \text{otherwise} \end{cases}, \tag{31}$$

where $\|P\|$ denotes the size of the path set; $\delta_{ap'}$ is a dummy variable, valued 1 if path p' goes via arc a, and 0 otherwise. This definition is inspired by the concept of Path Size, which is widely used in route choice modeling (see Ben-Akiva and Bierlaire, 2003) to measure how an alternative overlaps with other paths in the choice set. The O value, between 0 and 1, can be roughly understood as the proportion of the path p that overlaps with all paths in P. The more the overlapping, the higher the value. If p is the only path in P ($\|P\| = 1$), i.e. perfect overlapping in P, $O = 1$; if p does not overlap with any other path at all, $O = 0$.

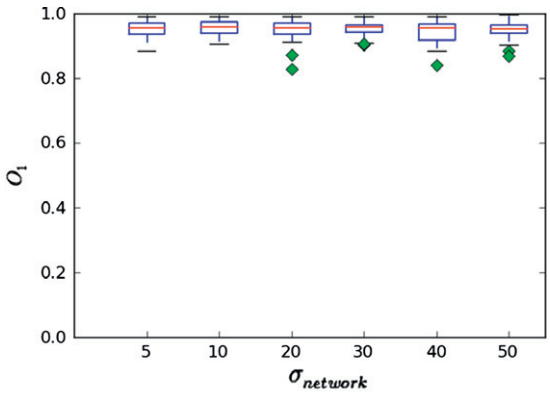
Based on this definition, we simply take the average to measure P's overall overlapping:

$$O_1(P) = \frac{1}{\|P\|} \sum_{p \in P} O(p,P).$$

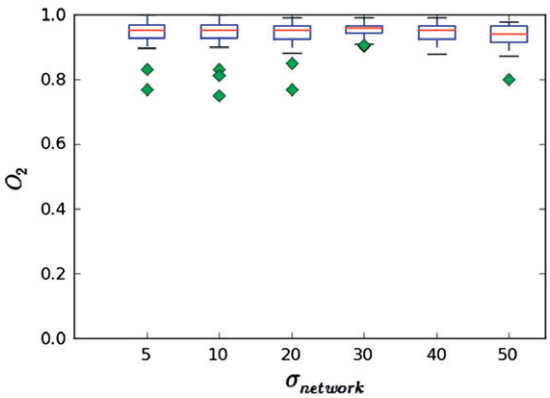
So O_1 also values between 0 and 1. And a higher value indicates a higher degree of overlapping in P. We expect this indicator close to 1 because all paths in a result should be similar. Indeed, The O_1 values for example results shown in Fig. 6 are: 0.966, 0.952, 0,962, 0.963, which are all close to 1. We also expect that the uncertainty of the result is insensitive to parameter variations. Hence O_1 value should be stable with respect to parameter variations.

Similarly, O_2 indicator for comparing two path sets is defined as:

$$O_2(P_1,P_2) = \frac{1}{\|P_1\|} \sum_{p \in P_1} O(p, P_2 \cup \{p\}).$$



(a) O_1



(b) O_2

Fig. 13. Sensitivity analysis for network error parameter.

In particular, if $P_1 = P_2$, then $O_2(P_1,P_2) = O_1(P_1) = O_2(P_2)$. Similar to O_1, O_2 indicator also values between 0 and 1. A higher value indicates a higher degree of overlapping between the two path sets. If any path in one set does not overlap with any path in the other set, $O_2 = 0$. If all paths in P_1 and P_2 are identical, $O_2 = 1$. In the sensitivity analysis for a parameter, P_2 is always set to be the path set produced by the default value. For example, in analyzing σ_{network}, P_2 is fixed to be $P_{\sigma_{\text{network}}=30}$, which is the path set produced by using $\sigma_{\text{network}} = 30$ (we follow the same notation convention throughout this section). Then, O_2 always indicates the overlapping of the paths generated from an alternative setting (e.g., $\sigma_{\text{network}=5}$) against $P_{\sigma_{\text{network}}=30}$. We expect the proposed method is robust in the sense that results produced by different parameter values are similar to each other. Therefore, this indicator should have high value (close to 1) for any parameter value being used in the analyses.

C.2. Network error

There are two sources of the network error. First, the OpenStreetMap network data are collected from GPS devices, so the error in the GPS records is introduced. The amplitude of this error is difficult to be estimated because many people contribute to the network data and they use different kinds of GPS devices. A survey on commercial GPS devices suggests that the error from commercial GPS receivers is less than 15 m in 95% of all cases (Ehsani et al., 2009). Second, in the network data, a road is represented as an abstract line without width. Therefore, we need to account for the width of the real road in the network error. This part of the error is also difficult to estimate because the details about the infrastructure are not easily accessible. In Switzerland, a third class road with one lane has minimum 2.8 m width and a first class road with 2 lanes has minimum 6 m width (Swisstopo, 2011). A motorway has not more than 4 lanes per direction (according to Swiss motorway website, http://www.autobahnen.ch), and according to Switzerland standard (3.20–3.75 m width per lane, OFROU (2011)), the maximum width per direction is 15 m.

Based on the above analysis, we believe that in most of the cases, σ_{network} is very unlikely to go beyond 50 m or below 5 m. So we perform a sensitivity analysis on σ_{network} with values 5 m, 10 m, 20 m, 30 m, 40 m and 50 m.

Fig. 13 reports the distribution of O_1 and O_2 across the 25 trips in dataset B, using a boxplot representation. We can notice from this figure that O_1 is in general close to 1, which indicates a high degree of overlapping and a low degree of uncertainty in the path results. And this indicator is insensitive to the σ_{network} value variation. Moreover, no mater which parameter

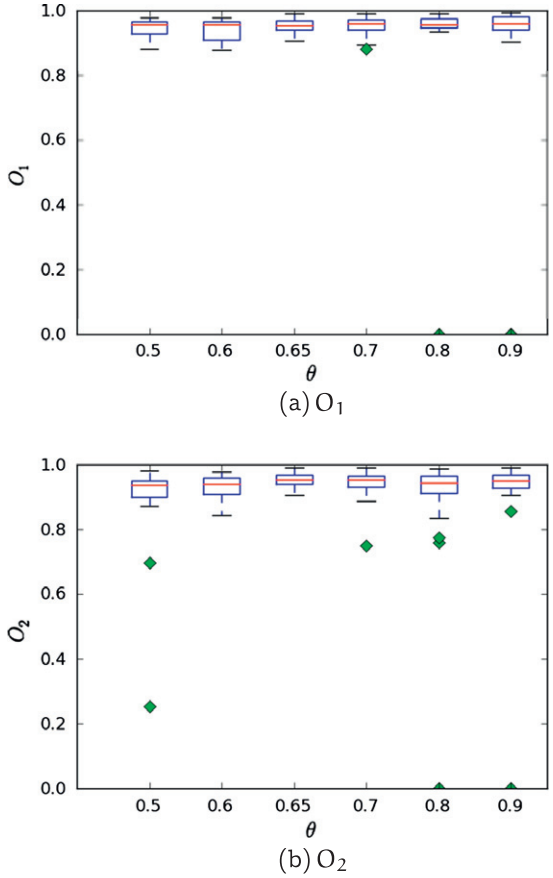


Fig. 14. Sensitivity analysis for θ parameter.

value is used, all the paths generated from the same trip should be, by the DDR definition, close to the GPS points, and hence, similar to each other. This is verified by Fig. 13b, in which high value of O_2 shows that all the path results are similar to $P_{\sigma_{\text{network}}=30}$. So, we can conclude that the probabilistic MM result is robust to σ_{network} variations.

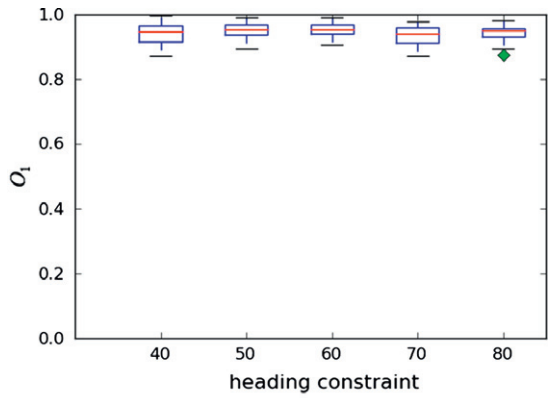
C.3. The DDR diameter

The parameter θ defines the diameter of a GPS point's DDR. Blunck et al. (2011) conduct an experiment that studies smartphone (Google Nexus One and Nokia N97 used) GPS data accuracy using a high performance dedicated GPS device as the benchmark device. They report that in open-sky urban conditions, in the worst case, at least 90% of the smartphone GPS points have the error distance less than 60 meters and 100% of them have error distance less than 100 meters. Therefore, in our experiment, we assume the DDR diameter to be 100m.

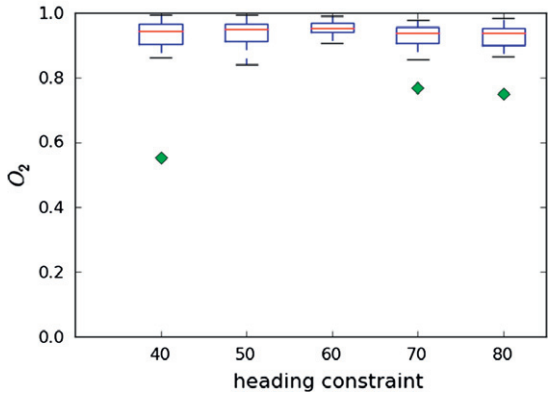
We perform a similar sensitivity analysis as in Section C.2 with θ to be 0.50 (123 m), 0.60 (106 m), 0.65 (97 m), 0.70 (88 m), 0.80 (70 m) and 0.90 (48 m) respectively (numbers in parentheses denote the corresponding diameters of the DDR), with 0.65 being the default setting. The O_1 and O_2 indicators are reported in Fig. 14. Generally high values of O_1 and O_2 suggest that the results are robust with respect to the variations of θ. However, from both graphs, we notice that some indicators are close to or equal 0 with θ being 0.50, 0.80 and 0.90. Actually, they correspond to the trip shown in Fig. 6d. This GPS trace is especially of low quality. It contains 15 stationary GPS points in the beginning of the trip (shown as a cluster in Fig. 6d), and there is a huge gap in the GPS trace. It is intentionally selected to analyze the robustness of the algorithm. With θ being 0.50, 0.80 and 0.90, the algorithm fails to produce any reasonable path. But the results with θ around 0.65 (0.6, 0.65 and 0.7) are reliable.

C.4. Heading constraint

In the total 1041 GPS points used in our experiment, 896 of them have speed greater than 10km/h. For these 896 GPS points, the mean of the recorded standard deviation of the heading error is 2.85 while the maximum is 36. So we safely



(a) O_1



(b) O_2

Fig. 15. Sensitivity analysis for heading constraint parameter.

assume that the heading error does not exceed 60° when the speed is greater than 10 km/h. It forms a rule for excluding unreasonable arcs from the DDR. At low speed status, heading measurements from the GPS are generally not reliable. Hence, for GPS points with speed less than 10 km/h, this heading constraint is not applied. Here, we analyze how much the variation of the heading constraint affects the result.

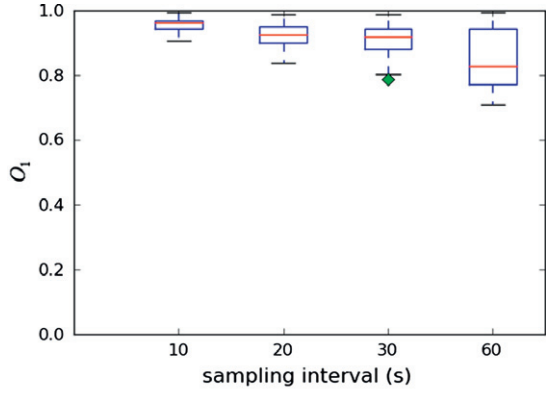
The analysis is performed on the values 40, 50, 60, 70, 80, where 60 is the default setting. The O_1 and O_2 indicators are reported in Fig. 15. As expected, the results are robust with respect to the parameter value variation, in the sense that both indicators for most of the cases are close to 1. In Fig. 15b, the outlier point in the boxplot for the constraint being 40 can be understood as an exceptional case when 40° is too tight for few GPS points. Overall, we can conclude that 60° is a suitable value for the heading constraint.

C.5. GPS sampling interval

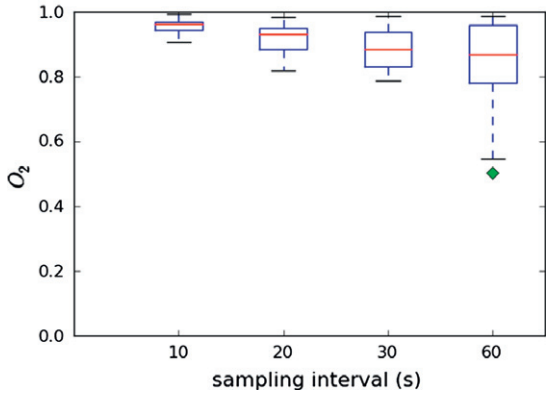
In this paper, the experiments are implemented for GPS data collected with sampling interval to be 10 s. However, we are also interested in applying the same method in more general situations. Therefore, we want to test the robustness of the method with respect to the data density. We artificially decrease the density of the data by manually increasing the GPS data interval to be $\kappa \in \{20, 30, 60\}$ s. The process is performed by selecting GPS points from the original data with following procedures:

1. select the first GPS point;
2. if the next GPS point is less than κ seconds later than the last selected GPS point, it is neglected; otherwise, it is selected;
3. repeat step 2 until the last GPS point.

We also perform a sensitivity analysis using O_1 and O_2 indicators. For the calculation of $O_2(P_{\text{interval}=\kappa}, P_{\text{interval}=10})$, the original data with interval 10*s* is truncated such that it has the same first and last GPS point as the processed data with interval κ. This is to guarantee that the GPS traces being compared have the same beginning and end, hence correspond to the same trip.



(a) O_1



(b) O_2

Fig. 16. Sensitivity analysis for GPS sampling interval.

First, we notice that the algorithm fails to proceeds at some GPS points in some cases (2 cases for $\kappa = 20$, 4 for $\kappa = 30$ and 4 for $\kappa = 60$). The temporary path set P_k produced at a certain iteration k is empty in these cases. Fig. 16 reports the O_1 and O_2 indicators for the successfully generated results. Fig. 16a shows a trend that the larger sampling interval, the higher uncertainty of the path result. This is consistent with the intuition that more GPS data brings more information, thus less uncertainty. Relatively high value of O_2 tells that results from different sampling interval settings are similar to the default setting. Since the paths generated from higher sampling interval are more heterogeneous; overall, they are less similar to the paths produced by the default setting, as reported by O_2 indicators. So, we conclude that, the performance of the proposed algorithm decrease with the density.

References

Barceló, J., Casas, J., 2005. Dynamic network simulation with aimsun. Simulation Approaches in Transportation Analysis, 57–98.

Ben-Akiva, M., Bierlaire, M., 2003. Discrete choice models with applications to departure time and route choice. In: Hall, R. (Ed.), Handbook of Transportation Science, second ed. Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 7–37.

Ben-Akiva, M., Bierlaire, M., Burton, D., Koutsopoulos, H., Mishalani, R., 2001. Network state estimation and prediction for real-time traffic management. Networks and Spatial Economics 1 (3), 293–318.

Ben-Akiva, M., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press Series in Transportation Studies. The MIT Press, Cambridge, MA.

Bierlaire, M., 2003. Biogeme: a free package for the estimation of discrete choice models. In: 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transportation Research Part C: Emerging Technologies 16 (2), 187–198.

Blewitt, G., Heflin, M.B., Webb, F.H., Lindqwister, U.J., Malla, R.P., 1992. Global coordinates with centimeter accuracy in the international terrestrial reference frame using GPS. Geophysical Research Letters 19 (9), 853–856.

Blunck, H., Kjærgaard, M., Toftegaard, T., 2011. Sensing and classifying impairments of GPS reception on mobile devices. In: Lyons, K., Hightower, J., Huang, E. (Eds.), Pervasive Computing, Lecture Notes in Computer Science, vol. 6696. Springer, Berlin/Heidelberg, pp. 350–367, chapter 22.

Ebendt, R., Sohr, A., Tcheumadjeu, L.C.T., Wagner, P., 2010. Utilizing historical and current travel times based on floating car data for management of an express truck fleet. In: 5th International Scientific Conference: Theoretical and Practical Issues in Transport.

Ehsani, R., Buchanon, S., Salyani, M., 2009. GPS Accuracy for Tree Scouting and other Horticultural Uses, Technical Report, University of Florida.

Flamm, M., Jemelin, C., Kaufmann, V., 2007. Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions. In: 7th Swiss Transport Research Conference, Ascona, Switzerland.

Flötteröd, G., Bierlaire, M., 2011. Metropolis-Hastings Sampling of Paths, Technical Report TRANSP-OR 110606, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Frejinger, E., 2008. Route Choice Analysis: Data, Models, Algorithms and Applications. PhD Thesis, EPFL.

Frejinger, E., Bierlaire, M., Ben-Akiva, M., 2009. Sampling of alternatives for route choice modeling. Transportation Research Part B: Methodological 43 (10), 984–994.

Greenfeld, J., 2002. Matching GPS observations to locations on a digital map. In: Proceedings of the 81th Annual Meeting of the Transportation Research Board, Washington, DC, USA.

Kiukkonen, N., Blom, J., Dousse, O., Laurila, J., 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. In: ICPS 2010: The 7th International Conference on Pervasive Services, Berlin.

Liao, L., Patterson, D.J., Fox, D., Kautz, H., 2007. Learning and inferring transportation routines. Artificial Intelligence 171 (5–6), 311–331.

Mahmassani, H., 2001. Dynamic network traffic assignment and simulation methodology for advanced system management applications. Networks and Spatial Economics 1 (3), 267–292.

Marchal, F., Hackney, J., Axhausen, K., 2005. Efficient map matching of large global positioning system data sets: tests on speed-monitoring experiment in zurich. Transportation Research Record: Journal of the Transportation Research Board 1935, 93–100.

Ochieng, W., Quddus, M., Noland, R., 2003. Map-matching in complex urban road networks. Brazilian Journal of Cartography (Revista Brasileira de Cartografia) 55 (2), 1–18.

OFROU, 2011. Route et trafic – chiffres et faits 2010, Technical report, Office fédéral des routes (OFROU).

Pyo, J., Shin, D., Tae-Kyung, S., 2001. Development of a map matching method using the multiple hypothesis technique. IEEE Proceedings on Intelligent Transportation Systems, 23–27.

Quddus, M.A., Noland, R.B., Ochieng, W.Y., 2005. Validation of map matching algorithms using high precision positioning with GPS. The Journal of Navigation 58 (02), 257–271.

Quddus, M.A., Ochieng, W.Y., 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.

Schuessler, N., Axhausen, K., 2009a. Map-Matching of GPS Traces on High-Resolution Navigation Networks using the Multiple Hypothesis Technique, Working Paper.

Schuessler, N., Axhausen, K., 2009b. Processing raw data from global positioning systems without additional information. Transportation Research Record: Journal of the Transportation Research Board 2105, 28–36.

Stopher, P.R., 2008. Collecting and processing data from mobile technologies. In: 8th International Conference on Survey Methods in Transport, Annecy, France.

Swisstopo, 2011. Cartes nationales de la Suisse: Symboles de nos cartes Symboles de nos cartes, Office fédéral de topographie swisstopo.

van Diggelen, F., 1998. GPS accuracy: lies, damn lies, and statistics. GPS World 9 (1), 41–45.

Yang, Q., Koutsopoulos, H.N., 1996. A microscopic traffic simulator for evaluation of dynamic traffic management systems. Transportation Research Part C: Emerging Technologies 4 (3), 113–129.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Huang, Y., 2010. T-drive: Driving Directions Based on Taxi Trajectories. ACM SIGSPATIAL GIS 2010. Association for Computing Machinery, Inc..