GENERATING PROBABILISTIC PATH OBSERVATION FROM GPS DATA FOR ROUTE CHOICE MODELING

Jeffrey Newman, Jingmin Chen, Michel Bierlaire

Ecole Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory,
Station 18, CH-1015 Lausanne, Switzerland

Abstract: Map matching algorithms are the conventional way to generate path observations from GPS data for route choice models. The deterministic matching may introduce extra biases to parameters of route choice models if the matching is wrong. In this paper, a new methodology is proposed to probabilistically generate path representation from GPS location data and the underlying network. This methodology takes advantage of both spatial and temporal relationships existing in the location data and the network. The generated result includes a set of potential true paths, along with a probability of each proposed path to have been the actual path. An algorithm is designed and applied to a simulated trip.

Keywords: path probability, spatial temporal, route choice modelling, map matching

1 INTRODUCTION

The data describing travellers' travelled routes is an indispensable input for route choice models. Merits of using GPS devices to collect data for transport studies have been recognized by comparing the collected data against the data collected from conventional survey methods, such as travel diary and telephone retrieval (Murakami and Wagner 1999; Bellemans, Kochan et al. 2008). However, to serve as an input for the models, the discontinuous location data needs to be processed to generate the actual path.

Deterministic map matching algorithms are the conventional way to generate a unique true path from GPS data trace. Significant advancements have been made in this field and the advanced algorithms may perform well generally in some applications (White, Bernstein et al. 2000; Quddus, Ochieng et al. 2007). However, using deterministic map-matching algorithms in route choice modelling may introduce extra biases if the location data is matched to a wrong path. The wrong matching is unavoidable because neither the location data nor the network data is always accurate. Also some map-matching algorithms bring systematic errors because, for example, the algorithm prefers to match to main roads.

Route choice modelling frameworks have been adapted to accept a probabilistic representation of the actual path (Bierlaire and Frejinger 2008). An observation is no longer necessary to be just a unique path in this framework. It can be represented by a set of potential paths, along with a probability for each path to have been the actual one. Bierlaire, Chen et al. 2009 proposed a theoretical framework for calculating probabilistic network mapping of location data to paths. In this paper, we introduce detailed techniques of an algorithm proposed by Bierlaire, Newman et al. 2009, which uses the framework. This algorithm is different from probabilistic approach used in some map-matching algorithms, which still generate a unique result from observations (Ochieng, Quddus et al. 2003).

In the next section we will introduce the probabilistic measurement for spatial relationship between a single location observation and two kinds of network elements, location and arc. In section 3, the algorithm used for calculating the path probability will be introduced. This algorithm utilizes the spatial temporal relationship reflected in observations. In section 4, an application of the algorithm to a synthetic scenario will be demonstrated. Finally, we will discuss some conclusions and future works.

2 SPATIAL MEASUREMENTS

2.1 The Network Representation

Let $G=(N,A)$ denote a network a network, where N is the set of all nodes and A the set of all arcs. The horizontal position of each node $n \in N$ is represented by $x_n = \{lat, lon\}$, which is a pair of coordinates consisting of latitude and longitude. The shape of physical route of arc a is described by an application $\ell_a : [0:1] \to \mathbb{R}^2$. For a point x on the arc, its position is given by a unique ϵ between 0 and 1 such that $x = \ell_a(\epsilon)$. In particular, $\ell_a(0)$ is the coordinates of the up-node, and $\ell_a(1)$ is the coordinates of the down-node of arc a. If arc a is a straight line between node u and node d, we have

$$\ell_a(\epsilon) = (1 \cdot - \epsilon) \square x_u + \epsilon\, x_d \qquad (1)$$

2.2 The location data

Location data is recorded by devices which are carried by travellers when they are travelling in the transportation network. The device makes observations on various kinds of direct and indirect location information sources from its sensors, including GPS readings, GSM cell tower information, WLAN base stations, etc. These location data sources can be generalized to be an observation of noisy location information with non-noisy time stamp information: $\hat{g} \hat{\hat{=}} (\hat{t}, x, \sigma^x, \hat{v}, \hat{\sigma}^v)$, which is a tuple containing:

- $\hat{t}$, a time stamp ;

- $\hat{x}$, a coordinates and the standard deviation of the error in the measurement of that coordinates, $\hat{\sigma}^x$;
- $\hat{v}$, a speed measurement and the standard deviation of the error in that measurement, $\hat{\sigma}^v$.

2.3 Location Measurement Errors

The location measurement is recorded as a pair of coordinates in north and south directions with errors e_{lat} and e_{lon}. We assume that errors are independent on directions and they both follow a normal distribution with mean 0 and standard deviation σ. Then the error distance between the true location and the observed measurement is:

$$r = \sqrt{e_{lat}^2 + e^2{}_{lon}}\,, \tag{2}$$

which follows Rayleigh distribution. The probability of observing the measurement in a location is defined as the probability that the distance between the measured point and the location is less than the error distance:

$$\Lambda(\hat{x}, \overline{x}) = \Pr(r \geq \| x \| - \overline{x}\ ) = \exp(-2\frac{r^2}{(\hat{\sigma}^x)^2})\,, \tag{3}$$

in which $\hat{\sigma}^x$ is the root mean square of standard deviations of e_{lat} and e_{lon}. Note that this probability is monotonically decreasing when the distance between the observed $\hat{x}$ and the hypothesized true $\overline{x}$ increases.

2.4 Single Location Measurement

For $\forall \epsilon_a \in a$ in the transportation network, the probability density of recording a GPS observation $\hat{g}$ is given by

$$f_{\mathbf{g},\varepsilon_{\mathbf{a}}}(\hat{g}, \hat{\epsilon}_a) = f_{\mathbf{g},\mathbf{x}}(g, \ell_a(\epsilon)) = \Lambda(x, \ell_a(\epsilon)). \tag{4}$$

By Bayes, given that a traveller is on some arc in the transportation network when a location observation $\hat{g}$ is recorded, the probability density of the traveller's location can be expressed as

$$f_{\boldsymbol{\epsilon}_{\mathbf{a}}}(\epsilon_a \mid \hat{g}, a) = \frac{f_{\mathbf{g},\varepsilon_{\mathbf{a}}}(\hat{g}, \epsilon_a)}{\Pr(\hat{g}, a)}, \tag{5}$$

where

$$\begin{aligned}\Pr(\hat{g},a) &= \int_{x\in a} f_{\mathbf{g},\mathbf{x}}(g,x)dx \\ &= l_a \cdot \int_0^1 f_{\mathbf{g},\epsilon}(\hat{g}, \epsilon_a)\, d\epsilon\end{aligned} \tag{6}$$

Then Equation (5) becomes

$$f_\epsilon(\epsilon_a \mid \hat{g},a) = \frac{\Lambda(\hat{x},\ell_a(\epsilon))}{l_a \cdot \int_0^1 \Lambda(\hat{x},\ell_a(\epsilon))\, d\epsilon}. \tag{7}$$

The probability that the traveller is on arc a when $\hat{g}$ is recorded is given by

$$\Pr(a \mid \hat{g}) = \frac{\Pr(\hat{g},a)}{\sum_{b\in A} \Pr(\hat{g},b)} = \frac{l_a \cdot \int_0^1 \Lambda(\hat{x},\ell_a(\epsilon))\, d\epsilon}{\sum_{b\in A} l_b \cdot \int_0^1 \Lambda(\hat{x},\ell_b(\epsilon))\, d\epsilon}. \tag{8}$$

For those arcs which are far away from $\hat{g}$, the probability value approach zero. Hence, we define a domain of data relevance (D) for $\hat{g}$, which only includes those arcs which have probability value greater than a threshold.

3 THE PATH PROBABILITY ALGORITHM

In the last section, only spatial relationship between an observation and the network is accounted in the measurements for the single observation. However, for a trace of location data, there also exist temporal relationships among observations and network. Therefore, in this section, a method is presented to calculate the probability that the path having been the actual one by accounting for both spatial and temporal relationships.

3.1 The Framework

We denote a series of GPS location data observed in a trip as $\hat{G} = \{\hat{g}_1, \cdots, g_{k-1}, \hat{g}_k\}$, in which $\hat{g}_1$ is the first GPS point observed and $\hat{g}_k$ is the last. Along a path, observing a GPS point is dependent on the previous GPS observation. This dependency is considered in the derivation of the measurement equation for calculating the probability of making the observations on a path, and we have

$$\Pr(\hat{g}_j, \hat{g}_{j-1}, \cdots, g_1 \mid p) = \Pr(g_j \mid g_{j-1}, \cdots, g_1, p)\Pr(g_{j-1}, \cdots, g_1 \mid p). \tag{9}$$

Therefore the probability that the path is the actual path is given by

$$\Pr(p \mid \hat{g}_1,\ldots,g_k) = \frac{\Pr(\hat{g}_1,\ldots,g_k \mid p)}{\sum_{\forall p'} \Pr(\hat{g}_1,\ldots,g_k \mid p')}. \qquad (10)$$

Since a path is comprised by connecting arcs, we calculate $\Pr(\hat{g}_j \mid g_{j-1},\cdots,g_1,p)$ in the domain of $\hat{g}_j$:

$$\begin{aligned}\Pr(\hat{g}_j \mid \hat{g}_{j-1},\cdots,g_1,p) &= \sum_{a\in(D_j\cap p)} \Pr(g_j,a \mid g_{j-1},\cdots,g_1,p) \\ &= \sum_{a\in(D_j\cap p)} l_a \int_0^1 f_{\mathbf{g},\varepsilon^{\mathbf{j}}}(\hat{g}_j,\epsilon_a)\, f_{\epsilon^{\mathbf{j}}}(\epsilon_a \mid g_{j-1},\cdots,g_1,p)\, d\epsilon_a,\end{aligned} \qquad (11)$$

where $\epsilon^{\mathbf{j}}$ denotes the random variable of the true position on an arc where $\hat{g}_j$ is recorded. $f_{\epsilon^{\mathbf{j}}}(\epsilon_a \mid \hat{g}_{j-1},\cdots,g_1,p)$ is the probability density function for the distribution of the current position given the trace of previous GPS observations, which we term the "state function". At the first observation, there isn't a previous observation, so

$$\Pr(\hat{g}_1 \mid P) = \sum_{a\in(D_j\cap p)} \Pr(g_1,a) \qquad (12)$$

4.2 The State Function

The underlying dependency in Equation (11) is actually resulted from the traveller’s movement during the intervening time between the two observations $\hat{g}_{j-1}$ and $\hat{g}_j$. This movement can be regarded as the travel from the domain of $\hat{g}_{j-1}$ to the domain of $\hat{g}_j$. Since a domain consists of several relevant arcs, the state function can be written as

$$f_{\epsilon^{\mathbf{j}}}(\epsilon_a \mid \hat{g}_{j-1},\cdots,g_1,p) = \sum_{b\in(D_{j-1}\cap p)} f_{\epsilon^{\mathbf{j}}}(\epsilon_a \mid b, g_{j-1},\cdots,g_1,p)\, f_{\mathbf{a}}(b \mid g_{j-1},\cdots,g_1,p)\,, \qquad (13)$$

where the probability $f_{\mathbf{a}}(b \mid \hat{g}_{j-1},\cdots,g_1,p)$ is calculated by

$$f_{\mathbf{a}}(b \mid \hat{g}_{j-1},\cdots,g_1,p) = \frac{\Pr(\hat{g}_{j-1},b \mid g_{j-2},\cdots,g_1,p)}{\Pr(\hat{g}_{j-1} \mid g_{j-2},\cdots,g_1,p)}\,. \qquad (14)$$

In Equation (13), $f_{\epsilon^{\mathbf{j}}}(\epsilon_a \mid b,\hat{g}_{j-1},\cdots,g_1,p)$ represents conditional probability of being at position ϵ_a when $\hat{g}_j$ is recorded, given that condition that the trace of GPS points before $\hat{g}_j$ has been observed. For simplification, only the previous $\hat{g}_{j-1}$ is taken into account, and the probability simplifies to $f_{\epsilon^{j}}(\epsilon_a | b, \hat{g}_{j-1})$. By considering each possible position on arc b it becomes

$$f_{\epsilon^{j}}(\epsilon_a | b, \hat{\hat{\hat{g}}}_{j-1}) = \int_{\epsilon_b=0}^{1} f_{\epsilon^{j}}(\epsilon_a | \epsilon_b, g_{j-1}, b) f_{\epsilon^{j-1}}(\epsilon_b | g_{j-1}, b) d\epsilon_b \,, \tag{15}$$

in which $f_{\epsilon^{j-1}}(\epsilon_b | \hat{g}_{j-1}, b)$ is the spatial relevance measurement which is calculated by Equation (7) . And $f_{\epsilon^{j}}(\epsilon_a | \epsilon_b, \hat{g}_{j-1}, b)$ is the probability that being at position ϵ_a when $\hat{g}_j$ is recorded, given the true position of recording $\hat{g}_{j-1}$ is ϵ_b, which we term the “position transition probability”. It reflects the trajectory of travelling from ϵ_b to ϵ_a .

2.3 Trajectory Between Observations

We depict the trajectory of travelling from previous position ϵ_b to the current position ϵ_a in Figure (1). There are several segments in the trajectory, and the total travel time $\mathbf{t}_{b \to a}$ is the summation of travel times on all segments. For instance, if a and b are different arcs, we have

$$\mathbf{t}_{b \to a} = \quad (1 - \epsilon_{b}\cdot) \boxed{\mathbf{t}}_b + \sum_{c \in sp^{b \to a}} \mathbf{t}_c + \mathbf{t}_w^{b \to a} + \epsilon_a \, \mathbf{t}_a = \hat{\hat{t}}_j - t_{j-1} \tag{16}$$

where

- $sp^{b \to a}$ is the sub-path from the down-node of arc b to the up-node of arc a;
- t_c is time cost on an arc c, which is a component arc of $sp^{b \to a}$;
- t_w is the total waiting time at intersections or other transportation facilities which might cause stops.

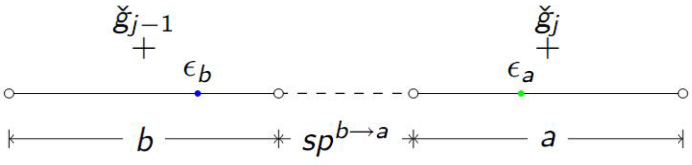

Figure 1 Position transition between adjacent domains

The position transition probability can be defined as the total travel time from ϵ_b to ϵ_a is the time difference observed:

$$f_{\boldsymbol{\epsilon}^{\mathbf{j}}}(\epsilon_a \mid \epsilon_b, \hat{g}_{j-1}, b) = f_{\mathbf{t}_{b\to a}}(\hat{t}_j - t_{j-1} \mid \epsilon_a, \epsilon_b, g_{j-1}, b) \tag{17}$$

In the following part of this section, the random variables in Equation (16) will be discussed.

Travel time cost on observing arcs a and b

We assume that the traveller keeps a constant speed when he is travelling on an arc. In the position transition probability measurement, a given condition is that a and b are the arcs where the GPS points are observed. So the observed speeds in $\hat{g}_{j-1}$ and $\hat{g}_j$ can be used to estimate the actual speeds of travelling on arcs a and b respectively. The normal distribution is a convenient and applicable assumption for the speed, and the mean $\hat{v}$ and standard deviation $\hat{\sigma}^v$ are given in the data. Since traveller's true speed is constrained by the capability of mean of transport which he is using, the speed distribution should be truncated within a continuous bound.

Travel time cost on arcs of intermediate sub-path $sp^{b\to a}$

The information about the travelling in $sp^{b\to a}$ is not explicitly observed. However, if the traveller's travelling pattern is stable, the speed data recorded in the nearly GPS points can be used to reveal the travelling pattern in the sub-path. Moreover, the travelling is also dependent on the underlying transportation network. In traffic theory, the free flow speed ratio reflects the traffic conditions. The inverse free flow speed ratio is

$$\varpi = \frac{\overline{v}}{v} \tag{18}$$

where $\overline{v}$ is the free flow speed or expected speed given in the network data, and v is the actual speed. Within a certain geographical area and time period, the traffic condition is assumed to be stable to some extent. We use normal distribution to depict ϖ, $\varpi \sim N(\overline{\varpi}, \delta_{\varpi}^2)$. At GPS observation $\hat{g}_j$, Θ_j is a set of GPS points which lie in a certain geographical area around $\hat{x}_j$ and a certain time period around $\hat{t}_j$. The inverse free flow speed ratio for $\hat{g}_j$ is calculated by

$$\hat{\varpi}_j \quad = \sum_{a \in D_j} \frac{\Pr(a \mid \hat{g}_j)}{\Pr(\hat{g}_j)} \frac{\bar{v}_a}{\hat{v}_j}. \tag{19}$$

The estimator for the mean of ϖ_j is

$$\bar{\varpi} = \frac{1}{n} \sum_{\hat{g}_i \in \Theta_j} \hat{\varpi}_i. \tag{20}$$

And estimator for the variance is

$$\delta_{\varpi}^2 = \frac{1}{n-1} \sum_{\hat{g}_i \in \Theta_j} (\hat{\varpi}_i - \bar{\varpi})^2. \tag{21}$$

For each $c \in sp^{b \to a}$,

$$\mathbf{t}_c = \frac{l_c}{\bar{v}_c} \cdot \varpi \tag{22}$$

follows normal distribution.

Waiting time caused by stops

During the interval time between adjacent observations, the traveller might be stopped due to traffic control devices existing in the network. The spots causing the stops are mostly intersections, where the traveller should wait for his green light. So if there are intersections between a and b, the possible waiting time should be captured. However, if the observation interval time is small enough, it is very possible that a very low speed GPS point is observed during the stop. For example, the interval time of recording a observation is set to be 10 seconds, then if the traveller has been waiting for 5 seconds, there is at least 50% possibility that a stop is observed, and 100% if 10 seconds. Hence, the meaning of the waiting time is introducing a penalty to those unlikely position transitions. The incorporation of GPS observations with very low speed is a topic for further research. Within this paper, we use uniform distribution to estimate the waiting time.

4 NUMERICAL EXAMPLES

We design a simulation scenario (see Figure (2)) to examine the capability of thealgorithm described in this paper. A synthetic network is constructed using two parallel horizontal lines, with vertical lines connecting them. Each horizontal line contains 10 arcs with length $94m$, and we change the length of vertical lines to different value in various scenarios to test the performance of the algorithm at various resolutions. All arcs in the network are bidirectional.

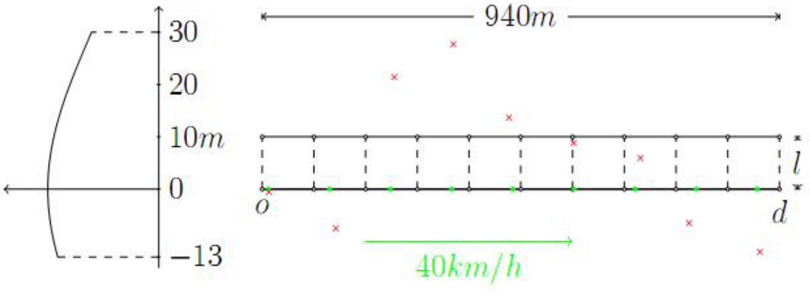

Figure 2 Simulated scenario

We simulate that a traveller drives a car at a constant speed $40km/h$, departing from node o, travelling along the bottom line, and arriving at his destination node d. Location observations are recorded in every $10s$. In Figure (2), the green solid points are those true locations where the observations are recorded. Errors are introduced to each observation.

- The errors of latitude and longitude are randomly drawn from normal distributions, which both have zero mean and different standard deviations randomly and independently selected from $\left[0,30\text{m}\right]$. The distributions of latitude and longitude errors are truncated in $[-13m,30m]$ and $[-20m,20m]$ respectively. And the standard deviation of latitude is multiplied by 1.5. These manipulations introduce systematic errors so as to make the simulation close to reality by. The root mean square of two standard deviations is calculated and recorded as $\hat{\sigma}^x$.
- Errors of speed is drawn from normal distribution as well, with zero mean, and a fixed standard deviation 6.

Among a large amount of simulated location measurements, we select a typical trace of location measurements of which the accuracy is bad. In Figure (2), the red x symbols are the locations observed. In this scenario, the human intuition can hardly recognise the true path. For comparison purpose, we run the algorithm with the length of vertical lines being $10m$, $15m$ and $20m$ respectively. For each scenario, only 6 most probable paths are presented. The two probability values under each path indicate the path probabilities calculated by different algorithms. The algorithm for the first value is the one describe above, which we term “spatial temporal algorithm”; while the second algorithm uses the similar method but without state function, and it basically just reflects the spatial relationship between the observations and the network, and we term it “spatial algorithm”.

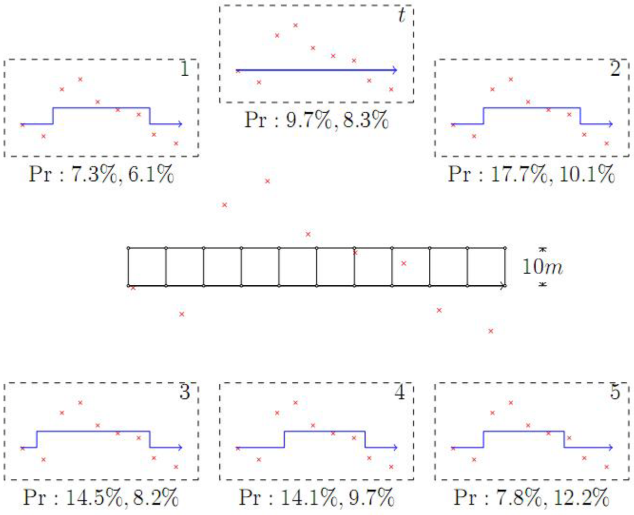

Figure 3 Result with $l = 10m$

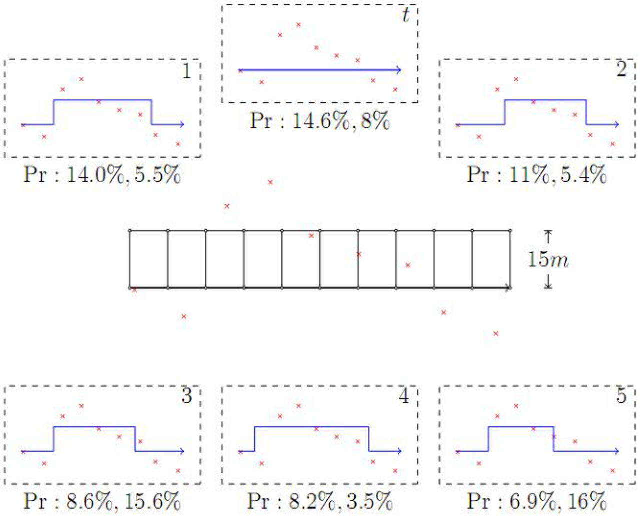

Figure 4 Result with $l = 15m$

The results in Figure (3) show that when $l = 10m$, both algorithms fail in giving the highest probability to the true path, although the differences between paths are not significant. When l is extended to be $15m$ (see Figure (4)), the spatial algorithm gives the highest probability values to the wrong paths (3 and 5). But the probability of the true path calculated by spatial temporal algorithm is the highest. Further, l is extended to be $20m$ (see Figure (5)). The true path gains the remarkable likelihood from spatial temporal algorithm, but the spatial algorithm fails again. The failure of the spatial temporal algorithm in the first case and the only marginally highest result in the second case also show that it is not a panacea.

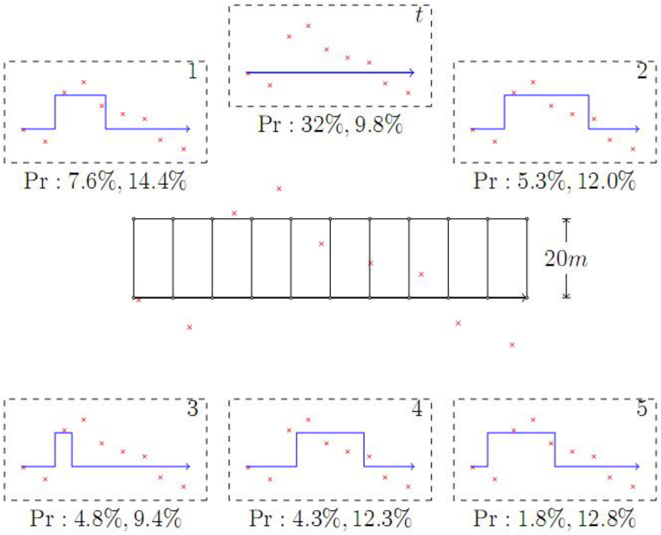

Figure 5 Result with $l = 15m$

5 CONCLUSIONS AND FUTURE WORKS

In this paper, an algorithm is presented to generate probabilistic representation of the path observation from the location data and the transportation network. Spatial relationships between single location observation and two basic network elements, location and arcs, are measured in a probabilistic fashion. For a series of location observations, the underlying spatial temporal relationships are taken into account in the calculation of the path likelihood which represents the probability that the path is the true path where the location observations are observed. Results from synthetic data show the viability of the algorithm in recognizing the true path.

A data collection campaign will be carried out to collect various kinds of data from sensors built in Nokia N95 smart phone, including GPS location data. This project is funded by, and collaboration with, the Nokia Research Center in Lausanne. About 100 N95 will be given out to respondents with pre-installed data recording software. Each respondent will use the device as her regular mobile phone, carrying it along with her throughout the day, while the software constantly records data and regularly sends the data to a server.

The algorithm will be improved by better utilizing low speed GPS observations. We sill compare it against the advanced map-matching algorithms. This algorithm will be used to generate path observations for route choice models.

REFERENCES

Bellemans, T., B. Kochan, et al. (2008). Field evaluation of personal digital assistant enabled by global positioning system impact on quality of activity and diary data. **Transportation Research Record:** 136-143.

Bierlaire, M., J. Chen, et al. (2009). Using location observations to observe routing for choice models, **Paper submitted to** Transportation Research Boarding Annual Meeting 2009.

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

Bierlaire, M., J. Newman, et al. (2009). A method of probabilistic map distribution of path likelihood. **Paper Submitted to** Swiss Transportation Research Conference 2009.

Murakami, E. and D. P. Wagner (1999). "Can using Global Positioning System (GPS) improve trip reporting?" **Transportation Research Part C: Emerging Technologies 7**(2-3): 149-165.

Ochieng, W., M. Quddus, et al. (2003). "Map-matching in complex urban road networks." **Brazilian Journal of Cartography** (Revista Brasileira de Cartografia) **55**(2): 1--18.

Quddus, M. A., W. Y. Ochieng, et al. (2007). "Current map-matching algorithms for transport applications: State-of-the art and future research directions." **Transportation Research Part C 15**(5): 312--328.

White, C. E., D. Bernstein, et al. (2000). "Some map matching algorithms for personal navigation assistants." **Transportation Research Part C: Emerging Technologies 8**: 91-108.