GPS Trajectory Data Enrichment based on a Latent Statistical Model

Akira Kinoshita¹, Atsuhiro Takasu², Kenro Aihara², Jun Ishii³, Hisashi Kurasawa³, Hiroshi Sato³, Motonori Nakamura³ and Jun Adachi²

¹The University of Tokyo, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
²National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
³NTT Network Innovation Laboratories, 3-9-11 Midoricho, Musashino, Tokyo, Japan

Keywords: GPS Trajectory Data, Interpolation, Latent Statistical Model, Moving Mode Estimation.

Abstract: This paper proposes a latent statistical model for analyzing global positioning system (GPS) trajectory data. Because of the rapid spread of GPS-equipped devices, numerous GPS trajectories have become available, and they are useful for various location-aware systems. To better utilize GPS data, a number of sensor data mining techniques have been developed. This paper discusses the application of a latent statistical model to two closely related problems, namely, moving mode estimation and interpolation of the GPS observation. The proposed model estimates a latent mode of moving objects and represents moving patterns according to the mode by exploiting a large GPS trajectory dataset. We evaluate the effectiveness of the model through experiments using the GeoLife GPS Trajectories dataset and show that more than three-quarters of covered locations were correctly reproduced by interpolation at a fine granularity.

1 INTRODUCTION

Because of the rapid spread of mobile devices equipped with a global positioning system (GPS), location information is combined with a wide variety of data and effectively exploited to realize new location-aware systems as well as to make existing systems smarter. For example, recommender systems utilize GPS data in several tasks, such as location-aware shopping recommendations (Yang et al., 2008) and tourism recommendations (Cao et al., 2010). Location information is key information for intelligent transportation systems such as traffic monitoring (Schnitzler et al., 2014) and incident detection (Kinoshita et al., 2015).

To utilize location information effectively, a number of sensor data mining methods have been proposed, with the moving mode estimation method often discussed in the literature. When analyzing user behavior, the means of traveling are useful. However, most GPS data do not contain such information. Zheng et al. (Zheng et al., 2010a) proposed a moving mode prediction method based on an extended hidden Markov model (HMM) where a moving mode is represented by a hidden state of the HMM. They assumed that the modes represent purposes and means of traveling such as driving, shopping, etc.

There has also been a growing interest in trajectory pattern mining. Giannotti et al. (Giannotti et al., 2007) proposed a frequency-based method, where they found popular areas and frequent moving patterns from trajectories. Monreale et al. (Monreale et al., 2009) extended this study for location prediction. They extracted the moving pattern represented by tree-structured data called a T-pattern tree from the training data, and then predicted the position based on the moving patterns.

Although the amount of GPS data is extremely large, we still need to enrich the data in various aspects. For example, the sampling rate is limited for saving the consumption of energy, which causes a sparsity problem for some analyses. Sampling every few seconds, for instance, is not sufficient for identifying the route of a car that is moving fast. In addition, some sensing data could be missing because of transmission failure.

This paper discusses two GPS data enrichment...
problems, namely, interpolation of GPS trajectories and traveling mode estimation.

When the sampling frequency is not high enough for analysis, GPS trajectories are interpolated in space and time. From the spatial point of view, there are several approaches to estimate locations and paths of GPS data. Many of them form a trajectory curve from discrete GPS position data. Brunsdon (Brunsdon, 2007) applied a principal curve detection technique (Biau and Fischer, 2012) to trace paths from GPS data. Sankararaman et al. (Sankararaman et al., 2013) extracted trajectory curve segments from trajectories, where a frequent portion of the trajectories is extracted by the dynamic time warping-based similarity. When moving objects are supposed to be on the road, map matching is useful for interpolation, and many studies have investigated the map-matching problem. Feng and Timmermans (Feng and Timmermans, 2013) proposed a map-matching method of GPS data based on the Bayesian belief network. Karagiorgou and Pfoser (Karagiorgou and Pfoser, 2012) proposed a map generation method where they detected intersections and then made a road network by connecting them. Hao et al. (Hao et al., 2014) proposed a probabilistic model to estimate the vehicle driving state, such as idling and acceleration, to estimate precisely the location at any time.

From the temporal point of view, Yang et al. (Yang et al., 2013) proposed the extended Gaussian mixture model (GMM) to estimate the traveling time of vehicles, where GMM is used to represent the probability density function of traveling time. Wang et al. (Wang et al., 2014) proposed a tensor-based method of traveling time.

The present paper proposes a statistical model for interpolating GPS sensing points. It introduces traveling modes to describe a movement behavior that varies according to the transportation means, with the expectation of improvement in interpolation accuracy. To exploit the training trajectories labeled with traveling modes, we applied the semi-supervised learning technique to obtain an effective model and evaluated the model effectiveness using real data.

2 MODEL FOR GPS TRAJECTORY ENRICHMENT

2.1 Trajectory

As moving objects generally travel on a road, observed GPS points are often mapped onto the road by a map-matching technique (e.g., Goh et al., 2012; Wei et al., 2012). However, the observed location is usually erroneous and the map-matching result is not always correct. In addition, people sometimes get out of the road network such as in a park. Therefore, we describe the location of moving objects by a grid. We first partition a 2-dimensional space into cells each of which represents an equal-sized and mutually excluded rectangle. Let G denote the grid, i.e., the set of cells in which objects move. We represent the location of a moving object by a cell \( g \in G \), meaning that the object is somewhere in the cell.

Given a grid \( G \), the movement of a moving object is described by the cells it passes through and by the traveling time for each cell. Let \( g_i \) be the \( i \)-th cell that the object passes through. Once the object enters a cell \( g_i \), it travels in it for time \( t_i \), then moves to the next cell \( g_{i+1} \). Therefore, the trajectory of a moving object is defined as a pair \( (g, t) \), where \( g = (g_1, g_2, \ldots, g_l) \) is a location sequence, \( t = (t_1, t_2, \ldots, t_l) \) is a traveling time sequence, and \( l \) is the length of both sequences.

2.2 Traveling Mode

Moving objects, particularly people, change location by walking or by various means of transportation such as a vehicle. Even when an object moves by the same means, its behavior may be different according to its location. For example, people tend to walk quickly in a business district to go to work, whereas they tend to walk more slowly in a commercial district. We introduce a set \( M \) of modes to distinguish the behavior patterns. Note that the mode is latent because we cannot observe it explicitly. A moving object may change its traveling mode at any time while traveling, but it makes the model too complicated. Therefore, in this paper, we assume that the moving object travels with the same mode in a cell. The traveling mode depends on the location. For example, the “train” mode is likely to be chosen on a railway, while the “car” mode is likely to be chosen on an expressway. Therefore, for each cell \( g \), we introduce a multinomial probability distribution with parameter \( \theta_m = (\theta_{gm})_{m \in M} \). The probability of the traveling mode \( m \in M \) of an object in a cell \( g \) is:

\[
p(m \mid g) := \theta_{gm}.
\] (1)

2.3 Traveling Time of Moving Objects

The traveling time varies according to the traveling mode and location as well as the individual characteristics. To avoid the sparsity problem in parameter estimation, we ignore the differences between individuals. For each mode \( m \) in a cell \( g \), we describe the distribution of traveling time \( t \) in terms of a univari-
ate Gaussian distribution with mean $\mu_{gm}$ and variance $\sigma_{gm}^2$, i.e.,
\[
p(t \mid g, m) := \mathcal{N}(t; \mu_{gm}, \sigma_{gm}^2) = \frac{1}{\sqrt{2\pi\sigma_{gm}^2}} \exp\left(-\frac{(t - \mu_{gm})^2}{2\sigma_{gm}^2}\right).
\]

Taking a marginal distribution, the traveling time $t$ to pass through $g$ follows a Gaussian mixture distribution:
\[
p(t \mid g) = \sum_{m \in M} \theta_{gm} \mathcal{N}(t; \mu_{gm}, \sigma_{gm}^2).
\]

### 2.4 Moving Direction

To predict the future location of a moving object, we introduce a probability distribution for the next cell to move into.

Suppose a moving object is in a cell $g$ and consider the probability distribution over the adjacent cells. The probability distribution should depend on the cell itself because of traffic constraints such as “no right turn” and attractors such as popular shops. It should also depend on the traveling mode. If a moving object is traveling by train, it tends to go straight to the next cell. On the other hand, if the moving object is walking, it may move to various directions.

From this observation, we introduce the probability distribution of the direction of an adjacent cell to which a moving object moves. Let $D$ be a set of directions of adjacent cells: \{north, east, south, west\}. When a moving object is in a cell $g$, we assume that its moving direction $d \in D$ follows a multinomial distribution with parameter $\phi_{gm} := (\phi_{gmd})_{d \in D}$, namely,
\[
p(d \mid g, m) := \phi_{gmd}.
\]

Now, the trajectory of a moving object is redefined as a triple $(g, t, d)$, where $d := (d_1, d_2, \ldots, d_l)$ is the moving direction sequence and $d_l$ is the moving direction from the $i$-th cell $g_i$.

### 2.5 Likelihood of a Trajectory

Let $x := (g, t, d)$ be a trajectory. The moving object takes only one of the traveling modes for each cell, although they are latent. Let $m_l$ be the traveling mode in the $i$-th cell and $y := (m_1, m_2, \ldots, m_l)$ be the mode sequence for the trajectory $x$. Then, the complete-data likelihood of the model is given as follows:
\[
p(x, y) = \prod_{i=1}^{l} p(t_i \mid g_i, m_i) \cdot p(d_i \mid g_i, m_i) \cdot p(m_i \mid g_i) = \prod_{i=1}^{l} \mathcal{N}(t_i; \mu_{gmi}, \sigma_{gmi}^2) \cdot \theta_{gmi} \cdot \phi_{gmi,d}.
\]

### 3 PARAMETER ESTIMATION

We adopt a maximum a posteriori (MAP) estimation for learning the prediction model. Let us first introduce the conjugate priors for each probability distribution. The symmetric Dirichlet distribution with the parameter $\alpha$ (respectively $\beta$) is used for the multinomial distributions for the mode (respectively the moving direction), whereas the Gaussian-gamma distribution with the parameters $\nu$, $\eta$, $a$, and $b$ is used for the traveling time distribution. Now the generative process of the model parameters is to choose them as follows:

1. $\theta_g \sim \text{Dir}(\alpha)$ for each $g \in G$,
2. $\phi_{gm} \sim \text{Dir}(\beta)$ for each $g \in G$ and $m \in M$,
3. $(\mu_{gm}, (\sigma_{gm}^2)^{-1}) \sim \text{GaussianGamma}(\nu, \eta, a, b)$ for each $g \in G$ and $m \in M$.

The observed trajectory data are considered to be generated under these parameters. Most of them are unlabeled, i.e., their mode is unknown. Let $X_u$ denote a set of unlabeled trajectories. The generative process of $X_u$ is as follows.

4. For each observation in each trajectory in $X_u$, (a) $m \sim \text{Multi}(\theta_g)$, (b) $d \sim \text{Multi}(\phi_{gm})$, (c) $t \sim \mathcal{N}(\mu_{gm}, \sigma_{gm}^2)$.

On the other hand, we can obtain a portion of labeled data where the mode of observation is known as in the GeoLife dataset (Zheng et al., 2009). Let $X_l$ denote a set of labeled trajectories. The generative process of $X_l$ is as follows.

5. For each observation in each trajectory in $X_l$, (a) $d \sim \text{Multi}(\phi_{gm})$, (b) $t \sim \mathcal{N}(\mu_{gm}, \sigma_{gm}^2)$, where $m$ is the labeled mode.

For simplicity, we denote the set of parameters used in our model by $\Theta$:
\[
\Theta := \{\{\theta_g\}_{g \in G}, \{\phi_{gm}, \mu_{gm}, \sigma_{gm}^2\}_{g \in G, m \in M}\}.
\]

Using both labeled and unlabeled data, $\Theta$ can be estimated by solving the following formula including two weight parameters $\lambda_l$ and $\lambda_u$ that control the effect of labeled and unlabeled data, respectively (Grönroos et al., 2014):
\[
\arg \max_{\Theta} \left[ \log p(\Theta) + \lambda_l \log p(X_l \mid \Theta) + \lambda_u \log p(X_u \mid \Theta) \right].
\]

Although we are in a semi-supervised situation, the estimate of $\Theta$ can be computed by an expectation-maximization (EM) algorithm. In the remainder of
this section, we derive the MAP estimator, concentrating on differences from the ordinary textbook treatments (Bishop, 2006; Zhu and Goldberg, 2009) because of lack of space.

The \( Q(\Theta, \hat{\Theta}) \) function for our model is given by

\[
Q(\Theta, \hat{\Theta}) = \sum_{x \in X_u} \sum_{i=1}^{N_g} p(m | x_i, \Theta) \ln p(x_i, m | \Theta),
\]

where

\[
p(m | x_i, \Theta) \propto \hat{\Theta}_g m \cdot \phi_{gmd}, \mathcal{N}(t_i | \mu_g m, \sigma_g m^2),
\]

\[
p(x_i, m | \Theta) = \theta_{g m}, \phi_{gmd}, \mathcal{N}(t_i | \mu_g m, \sigma_g m^2),
\]

and \( \hat{\Theta} \) refers to the parameters estimated in the previous EM iteration. The E step computes Equation (9) for each observation \( x_i := (g_i, t_i, d_i) \) in the unlabeled training dataset \( X_u \) for each mode \( m \in M \).

According to Equation (7), the objective function to be maximized in the M step is given by

\[
F(\Theta) := \ln p(\Theta) + \lambda_l \ln p(X_l | \Theta) + \lambda_u Q(\Theta, \hat{\Theta}).
\]

The first term is rewritten using the priors we introduced above, as follows:

\[
\ln p(\Theta) = \sum_{g \in G} \ln p(\theta_g | \alpha) + \sum_{g \in G, m \in M} \ln p(\theta_{g m} | \beta)
\]

\[
+ \sum_{g \in G, m \in M} \ln p \left( \mu_{g m}, (\sigma_{g m}^2)^{-1} \right) \mathcal{N}(t | \mu_{g m}, \sigma_{g m}^2).
\]

The second term is the weighted log-likelihood of the labeled training data. Using Equation (5), we obtain

\[
\ln p(X_l | \Theta) = \sum_{g \in G, m \in M} N_{g m} \ln \theta_{g m}
\]

\[
+ \sum_{g \in G, m \in M, d \in D} N_{g m d} \ln \phi_{g m d}
\]

\[
+ \sum_{g \in G, m \in M} \sum_{j=1}^{N_{g m}} \ln \mathcal{N}(t_j | \mu_{g m}, \sigma_{g m}^2),
\]

where \( N_{g m} \) is the number of labeled observations in the cell \( g \) with the mode label \( m \), \( N_{g m d} \) is the number of labeled observations whose direction is \( d \) in the cell \( g \) with the mode label \( m \), and \( t_j \) is the \( j \)-th labeled observation value of the travel time in the cell \( g \) with the mode label \( m \). The third term is the weighted \( Q \) function, which can be rewritten as follows:

\[
Q(\Theta, \hat{\Theta}) = \sum_{g \in G, d \in D} \sum_{j=1}^{N_{g m d}} \gamma_{g m d j} \left[ \ln \theta_{g m} + \ln \phi_{g m d} + \ln \mathcal{N}(t_j | \mu_{g m}, \sigma_{g m}^2) \right],
\]

where \( N_{g m d} \) is the number of unlabeled observations in the cell \( g \) whose direction is \( d \), \( x_j := (g, t_j, d) \) is the \( j \)-th unlabeled observation value in the cell \( g \) with the direction \( d \), and \( \gamma_{g m d j} := p(m | x_j, \Theta) \). Because the parameters \( \theta_g \) and \( \phi_{g m} \) have a constraint, respectively, Equation (11) is maximized by introducing Lagrange multipliers and setting its partial derivative to zero. The update equations are derived as follows:

\[
\theta_{g m} \propto \alpha - 1 + \lambda_l N_{g m} + \lambda_u \sum_{d \in D} \gamma_{g m d j},
\]

\[
\phi_{g m} \propto \beta - 1 + \lambda_l N_{g m d} + \lambda_u \sum_{j=1}^{N_{g m d}} \gamma_{g m d j},
\]

\[
\mu_{g m} = \frac{\nu \lambda_l \sum_{j=1}^{N_{g m}} t_j + \lambda_u \sum_{d \in D} \sum_{j=1}^{N_{g m d}} \gamma_{g m d j} t_j}{\eta + \lambda_l N_{g m} + \lambda_u \sum_{d \in D} \sum_{j=1}^{N_{g m d}} \gamma_{g m d j}},
\]

\[
\sigma_{g m}^2 = \frac{S}{2\alpha - 1 + \lambda_l N_{g m} + \lambda_u \sum_{d \in D} \sum_{j=1}^{N_{g m d}} \gamma_{g m d j}},
\]

where \( S := 2b + \eta(\mu_{g m} - v)^2 + \lambda_l \sum_{j=1}^{N_{g m}} (t_j - \mu_{g m})^2 + \lambda_u \sum_{d \in D} \sum_{j=1}^{N_{g m d}} \gamma_{g m d j} (t_j - \mu_{g m})^2.\)

### 4 INTERPOLATION

Now assume that we have the total traveling time \( t_x := \sum_{j=1}^{N_{g m}} t_j \) instead of the traveling time sequence \( t \). Let \( x \) be a triple \( (g, t, d) \). As each \( t_j \) follows a Gaussian distribution \( \mathcal{N}(\mu_{g m}, \sigma_{g m}^2) \), the sum of normally distributed variables \( t_x \) obeys the Gaussian distribution \( \mathcal{N}(\mu_x, \sigma_x^2) \), where

\[
\mu_x = \sum_{j=1}^{N_{g m}} \mu_{g m}, \quad \sigma_x^2 = \sum_{j=1}^{N_{g m}} \sigma_{g m}^2.
\]

Therefore,

\[
p(x, y) = p(t_x | g, y) \prod_{i=1}^{N_{g m}} p(d_i | g_i, m_i) \cdot p(m_i | g_i)
\]

\[
= \mathcal{N}(\mu_x; \mu_y, \sigma_y^2) \cdot \prod_{i=1}^{N_{g m}} \phi_{g m d i}.
\]

Using two distant GPS observations, the total traveling time \( t_x \) and the first and last cells of the location sequence \( g \) can be calculated directly. Given a set of possible location sequences \( \{g\} \), a corresponding set of possible moving direction sequences \( \{d\} \), a set of possible traveling mode sequences \( \{y\} \), we can obtain the maximum-likelihood trajectory, i.e., the most
probable g, d, and y, using Equation (21), whereby
the observations are interpolated. On the assumption
that the trajectory travels $n_{ew}$ cells along the east–west
direction and $n_{ns}$ cells along the north–south direc-
tion via the shortest path, the cardinality of the search
space is
\[
\frac{(n_{ew} + n_{ns})!}{n_{ew}! n_{ns}!} |M|^{n_{ew} + n_{ns}}.
\] (22)

5 EXPERIMENTAL RESULTS

5.1 Experimental Setup

We used GeoLife GPS Trajectories Version 1.3 (Zheng et al., 2008; Zheng et al., 2010b; Zheng et al.,
2009) for evaluating the proposed latent model. This
dataset consisted of trajectories of 182 users. The tra-
jectories of 69 users were associated with a traveling
mode in walk, run, bike, bus, taxi, car, subway, train,
airplane. Therefore, the number of modes $|M|$ was 9.

We chose data inside the area of central Beijing
shown in Figure 1, where observed GPS data were
not sparse. We generated trajectories by concatenat-
ing observations in chronological order whenever the
time gap between two consecutive observations was
10 s or less.

Next, we converted the original GeoLife trajectory
data into our trajectory form described in Section 2.1.
Figure 2 illustrates the method of conversion. We first
split the area into cells with a width of 0.0006° of lon-
gitude and height of 0.0005° of latitude. Each cell $g$
is a rectangle with side lengths of about 50 m. An original
GeoLife trajectory is a sequence of spatiotemporal
points. We converted the trajectory by finding all the
cells through which it passes and calculating the trav-
eling time for each cell by linear interpolation. We
used 90% of the converted dataset for training and the
residual 10% of the data was used for the test. The
training dataset included nine-tenths of consecutive
trajectories for each user.

5.2 Model Parameter Estimation

We estimated the model parameters by MAP esti-
mation. There were 4,462,614 observations from
186,617 cells in the training dataset. We used prior
parameters for each mode, as shown in Table 1, which
we chose arbitrarily. The value of $\nu$ was determined
by dividing 50 m, which is equal to the length of a
side of a cell, by the mean travel speed we assumed
(Table 1). We also used prior and weight parameters:
$\alpha = \beta = 2.0$, $\lambda_{a} = 0.5$, $\lambda_{l} = 1.0$.

We implemented the EM algorithm described
above using OpenMP for multiprocessing. The EM
algorithm has iterated the E step and the M step until
the improvement in log-likelihood fell below 0.01%.
The estimation was executed on our 32-core Xeon
computer. The EM algorithm was finished in nine it-
erations taking 11.3 s.

Table 1: Transportation modes and their priors.

<table>
<thead>
<tr>
<th>mode</th>
<th>mean speed $\nu$ [s]</th>
<th>$\eta$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>walk</td>
<td>1 km/h</td>
<td>180.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>run</td>
<td>4 km/h</td>
<td>45.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>bike</td>
<td>10 km/h</td>
<td>18.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>bus</td>
<td>15 km/h</td>
<td>12.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>taxi</td>
<td>20 km/h</td>
<td>9.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>car</td>
<td>30 km/h</td>
<td>6.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>subway</td>
<td>40 km/h</td>
<td>4.5</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>train</td>
<td>60 km/h</td>
<td>3.0</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
<tr>
<td>airplane</td>
<td>900 km/h</td>
<td>0.2</td>
<td>$10^{-6}$</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3 shows the estimated parameters; only
two of nine modes are shown because of space limitations. As can be seen, there are regional differences of traveling mode tendencies: slower modes tend to appear around local streets, while faster modes are likely to appear on arteries or railways. Figure 4 is an enlarged view, showing the differences in moving direction and mean travel time between two different modes. The moving direction is also learned, so that a trajectory travels on the right side of wide roads and that it takes different routes depending on the mode.

5.3 Interpolation and Traveling Mode Estimation

We evaluated the performance of our interpolation method. As our algorithm has high complexity, we prepared a 3x5 dataset by collecting all subtrajectories that travel three cells along a north–south or east–west direction and five cells along the orthogonal direction via the shortest path. For each subtrajectory in the 3x5 dataset, we estimated the intermediate cells given its first and last cells and its total travel time. The cardinality of the search space was 2,410,616,376. The interpolation was finished in 1 min for each subtrajectory using our 72-core Xeon computer and the OpenMP technology. We evaluated the interpolation performance by recall, i.e., the number of correctly interpolated cells divided by the number of the total intermediate cells that actually included observation data of the original GeoLife dataset.

In the test dataset, there were 8,276 subject subtrajectories and the recall was 78.8% (38,695 success/49,092 cells). Although we did not conduct any parameter tuning, more than three-quarters of the interpolations were successful. There is an ample room...
for improving performance by giving better priors and weight parameters when the model is learned. On the other hand, our interpolation method has high complexity and is not scalable. Further work is being carried out to improve the scalability of the algorithm.

Of the 3x5 subtrajectories in the test dataset, 309 have traveling mode labels. Our interpolation algorithm estimated the traveling mode for each cell at the same time as determining the cells. The accuracy of the traveling mode was 12.9% (184 correct/1,427 correctly interpolated cells). Figure 5 shows the confusion matrix.

One possible reason for the poor accuracy of traveling mode estimation is that the distribution of traveling time was similar for all traveling modes. Figure 6 shows box plots of traveling time over the whole dataset for each traveling mode. Here it can be seen that the distributions of some modes, car and train for example, were similar. This calls for further discussion on feature selection. Another possible problem is the labeling quality in the dataset. For example, a GeoLife trajectory shown in Figure 7 was labeled as “airplane” mode, but its movement was unnatural for an airplane because it traveled along expressways, arteries, and local streets. As the interpolation in this experiment was conducted only in the spatial aspect, it remains a challenge for future research to enrich trajectories in the temporal aspect. Improving the performance of traveling mode estimation would assist this kind of trajectory enrichment.

6 CONCLUSION

We have studied the problem of GPS trajectory enrichment, namely, interpolation and traveling mode estimation. We developed a statistical model where the traveling time and the moving direction depended on both the location and the latent traveling mode, whereas the mode also depended on the location. We derived formulas to estimate the MAP parameters of the model using GPS observation data which can include some observations with traveling mode labels. Our method was applied to the GeoLife dataset. The results showed that our model could describe the characteristics of movements depending on location and traveling mode and that more than three-quarters of covered locations were correctly reproduced by interpolation at a fine granularity. Future work will include the development of a more computationally efficient interpolation algorithm, optimization of the set of traveling modes, and feature selection.

ACKNOWLEDGEMENTS

This work was supported by CPS-IIP Project in the research promotion program for national-level challenges “Research and development for the realization of next-generation IT platforms” by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Figure 5: Confusion matrix of traveling mode estimation.

Figure 6: Distribution of traveling time for each mode.

Figure 7: A GeoLife trajectory labeled as “airplane” mode (map tiles ©OpenStreetMap contributors, CC BY-SA 2.0).
REFERENCES


