DATA-ORIENTED ALGORITHM FOR ROUTE CHOICE SET GENERATION IN A METROPOLITAN AREA WITH MOBILE PHONE GPS DATA

T. Nakamura [a,] *, Y. Sekimoto [a], T. Usui [a], R. Shibasaki [a]

[a] Center for Spatial Information Science, University of Tokyo, Tokyo, Japan – (ki_ki_gu, sekimoto, usui, shiba)@csis.u-tokyo.ac.jp

Commission II, WG II/5

KEY WORDS: GIS, Generation, Spatial, Mobile, GPS, Urban

ABSTRACT:

Nowadays, for the estimation of traffic demand or people flow, modelling route choice activity in road networks is an important task and many algorithms have been developed to generate route choice sets. However, developing an algorithm based on a small amount of data that can be applied generally within a metropolitan area is difficult. This is because the characteristics of road networks vary widely.
On the other hand, recently, the collection of people movement data has lately become much easier, especially through mobile phones. Lately, most mobile phones include GPS functionality. Given this background, we propose a data-oriented algorithm to generate route choice sets using mobile phone GPS data. GPS data contain a number of measurement errors; hence, they must be adjusted to account for these errors before use in advanced people movement analysis. However, this is time-consuming and expensive, because an enormous amount of daily data can be obtained. Hence, the objective of this study is to develop an algorithm that can easily manage GPS data.
Specifically, at first movement data from all GPS data are selected by calculating the speed. Next, the nearest roads in the road network are selected from the GPS location and count such data for each road. Then An algorithm based on the GSP (Gateway Shortest Path) algorithm is proposed, which searches the shortest path through a given gateway. In the proposed algorithm, the road for which the utilization volume calculated by GPS data is large is selected as the gateway. Thus, route choice sets that are based on trends in real GPS data are generated.
To evaluate the proposed method, GPS data from 0.7 million people a year in Japan and DRM (Digital Road Map) as the road network are used. DRM is one of the most detailed road networks in Japan. Route choice sets using the proposed algorithm are generated and the cover rate of the utilization volume of each road under evaluation is calculated. As a result, the proposed route generation algorithm and GPS data cleaning process work well and a huge variety of routes that have high potential to be used in the real world can be generated.

1. INTRODUCTION

1.1 Motivation and Aims

Modelling route choice activity in road networks is one of the most important challenges in the estimation of traffic demand or people flow. Recently, human activity has become very complicated. In concern with the development of machines such as vehicles and car navigation systems, the diversity of peoples' route choices has increased. Hence, predicting route choice is difficult. Many route choice algorithms have been developed to tackle this problem. Some are geometric algorithms; some involve use of the logit model. However, developing an algorithm that can be used universally in a metropolitan area, whether by geometry or by the logit model with a small amount of data, is difficult because road networks vary widely in their characteristics and each person has his/her own point to choose roads and routes.
With rapid development of information technologies, the collection of people movement data has become much easier, especially through mobile phones. Recently, most mobile phones come with GPS functionality. As peoples' route choices become more complicated, actual observational data should be used to reflect the huge variety of route choices. However, the ease of obtaining real data leads to an excess of data; hence, selecting useful data or cleaning the obtained data becomes necessary. GPS data contain measurement errors and these errors must be removed for use in advanced people movement analysis.
Based on this background, a data-oriented algorithm to generate route choice sets by using mobile phone GPS data is porposed. Vehicle routes are focused on because they involve more alternatives than routes by other traffic modes (e.g., railway, walking). This study has two main purposes. One is to propose a data-oriented method to generate route choice sets. This includes methods to process the observed data with little computational cost. Reduction of this cost is important for estimating people flow in a metropolitan area because there are many origin–destination combinations. Another purpose is to clean up mobile phone GPS data to be used for the generation of route choice sets. It must also be simple to reduce the computational cost.

1.2 Overview and References

First of all, the route choice problem is considered. There are two important steps in solving this problem. One is to generate a route choice set and the other is to select the best-fitting route. Although the former is focused on in this study, some algorithms such as those based on the logit model address the

* Corresponding author. Tel.: +81 471364307 fax: +81 471364292

two steps at once (Dial, 1971). These algorithms can generate some routes probabilistically; however, they have a number of disadvantages. For example, the parameter tuning is cumbersome, the computational cost is high, and many similar routes are generated.

Some algorithms separate the problem into a generation phase and a selection phase. There are many algorithms to generate route choice sets. The K-shortest Path algorithm is a simple and well-known method that generates the first k shortest paths. Link penalty and link elimination are two popular techniques that iteratively generate a shortest path by giving penalties to or eliminating links used as the shortest path (De la Barra, 1993). The labeling approach generates many types of route such as the shortest path, the minimum cost path, the minimum time path, and the maximum use of expressways path (Ben-Akiva, 1984). Some algorithms based on GSP algorithm have been developed. GSP algorithm is an algorithm that searches the shortest path passing through a given gateway (Lombard, K. 1993). However, some critical shortcomings are indicated (Akgum, V., 2000). The main one is that some of the gateway paths may contain loops. In this study, loop paths problem is avoided by extending GSP with a given link instead of a given gateway.

These algorithms attempt to extract trends in route choices to generate the route choice set. As road networks vary greatly in their characteristics and environmental features that cannot be quantified affect route choice behavior, their application to a large-scale area gives rise to many gaps between the generated and actual routes. Based on this problem, we propose a data-oriented algorithm to reflect the differences of road networks by using observed data and extending GSP algorithm.

Secondly, we consider how to handle GPS data. In particular, we distinguish the traffic mode (e.g., railway, car, feet). There are mainly two types of methodologies: rule-based logic and fuzzy logic. In rule-based logic, the conditions of traffic mode selection are determined by GIS data and information extracted from raw GPS data such as speed (Chung, 2005). In fuzzy logic, traffic modes are identified by fuzzy variables such as the median speed as well as the ninety-fifty percentiles of the speed and acceleration distributions (Schuessler, 2009). The logic used these studies is complicated because their purpose is to label all GPS data. However, in this study, not all GPS data are necessary. Therefore, we can focus on data clearly labeled by the traffic mode and other necessary information.

2. PROPOSED APPROACH

In this section, the proposed approach is described. The section is divided into two parts: route choice set generation and GPS data cleaning. We explain their details in sections 3 and 4.

For route choice set generation, an algorithm based on usage frequency of links calculated by GPS data is proposed, as shown in **Figure 1**. At first the target origin and destination are set to generate routes. Then, the GPS data-cleaning phase is conducted. Initially, data relating to movements from the origin to the destination are extracted by location and time information. Next, the usage frequency of all related links is calculated with the extracted GPS data. We use this measure in the route generation phase by selecting well-used links based on the usage frequency and generating routes by using each well-used link. In this way, we generate a number of routes that include well-used links and hence have high use potential.

The proposed method has three key advantages. The first is that by directly reflecting the observed data, the generated routes display a trend of links that is difficult to obtain. The second is that the method is little influenced by errors of GPS data. By using an adequate amount of data, we can reduce the negative influence of errors because the data are finally aggregated as the to-be-used usage frequency. Moreover, because the proposed method aggregates data, it can deal with large amounts of data and easily incorporate new data. Finally, this method can be applied to the collection of sparse information data. In order to generate routes in a large-scale area, measurement of rich information data with small GPS intervals, for example, is impractical owing to the battery of mobile phones and the data volume. The proposed method is largely uninfluenced by sparse information because it does not analyze each movement.

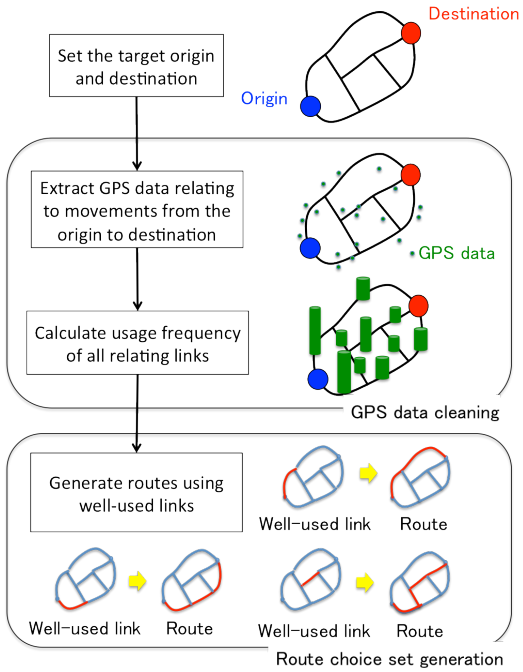

Figure 1. Schematic diagram of the proposed approach

3. GPS DATA CLEANING

3.1 Data

The proposed method is applied to GPS data obtained by mobile phones. **Table 1** shows the specification of the data. The data are collected by mobile phones whose owners agree to provide their GPS location data. GPS data are basically obtained while the owner is moving and its GPS interval is more than 5 min owing to the battery. The data are obtained from all areas of Japan. The number of observed mobile phones is about 0.7 million, and the period from August 1, 2010 to July 31, 2011.

For the road network, we use DRM (Digital Road Map), which is one of the most popular road networks in Japan. It is updated every year based on new topographical maps and information from road administrators nationwide.

Table 1. Specification of data

Location	All areas of Japan
GPS interval	More than 5 min Only while moving
Number	0.7 million mobile phones
Date	From August 1, 2010 to July 31, 2011.

3.2 Calculation of link usage frequency

In order to use this GPS data to generate a route choice set from an origin to a destination, extraction of the usable portion is necessary. **Figure 2** shows a flowchart of the data processing procedure. At first the data are separated into trips, each of which involves one movement. As the data used in this study are obtained while the owner is moving, they are separated into trips by cutting out data when the time between two consecutive GPS data is more than a threshold value. As the GPS interval is 5 min, the threshold value for trips is set as 60 min.

After separating the data into trips, the trips that pass through both the origin and destination are extracted. We consider trips that contain GPS data within a threshold value D from the origin and GPS data within D from the destination as trips that pass through both the origin and destination. In this study, D is set as 100 m because the observation error of GPS is less than this value.

Then, trips whose travel time from the origin to the destination is no more than a threshold value T_1 and no less than a threshold value T_2 are extracted to collect trips by vehicles. T_1 and T_2 represent the times when the vehicle speed is maximal and minimal, respectively. Specifically, T_1 and T_2 are calculated with the total length of the shortest path from the origin and the destination, as well as vehicle speeds of 30 km/h for T_1 and 80 km/h for T_2.

Next, GPS data around subway stations are deleted to remove trips by subway. This is because the travel time of trips by subway is similar to those of trips by vehicles. In addition, almost all GPS locations of trips by subway are around subway stations because the subway line is usually underground and GPS data cannot be obtained when the owner is in such an environment. 100 m is set as the threshold value.

Finally, the usage frequency of each link is calculated by counting the nearest link of each GPS datum extracted by the data processing.

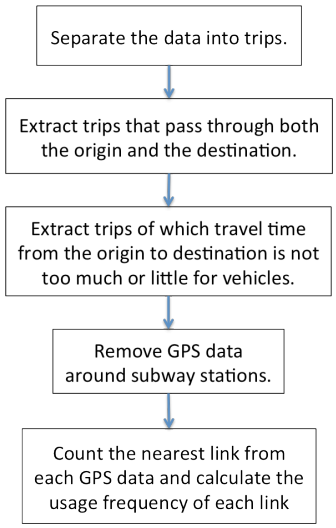

Figure 2. Calculation of link usage frequency

4. ROUTE CHOICE SET GENERATION

4.1 GSP Algorithm

In this paper, the algorithm to generate a route choice set based on the GSP (Gateway Shortest Path) algorithm is proposed. **Figure 3** shows an example of the GSP algorithm. The lower left circle is the origin and the upper right circle is the destination. The bold polyline on the left is the shortest path from the origin to the destination. The GSP method gives a gateway similar to the figure and generates the shortest path from the origin to the gateway ("Shortest Path #1"), and the shortest path from the gateway to the destination ("Shortest Path #2"). As a result, a new route connecting the two shortest paths is generated.

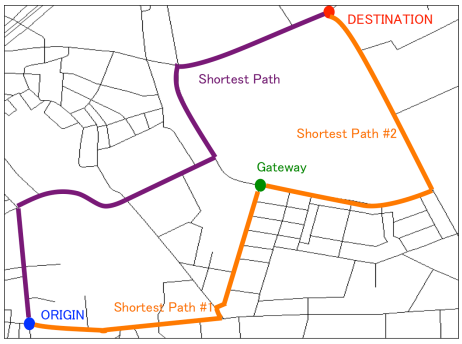

Figure 3. Example of GSP algorithm

In the proposed method, a link is given instead of a node. Then the shortest path including the given link is generated. Given links are selected based on the observed data. As a link has two nodes, there are two shortest paths from the origin to the given link. Likewise, there are two shortest paths from the given link to the destination (**Figure 4**). Therefore, the GSP algorithm generates two routes including the given link: the combination of "Shortest Path #1-1" and "Shortest Path #2-2," as well as the combination of "Shortest Path #1-2" and "Shortest Path #2-1". Finally, in order to select one of the routes, we compare the total route lengths and select the route with less length.

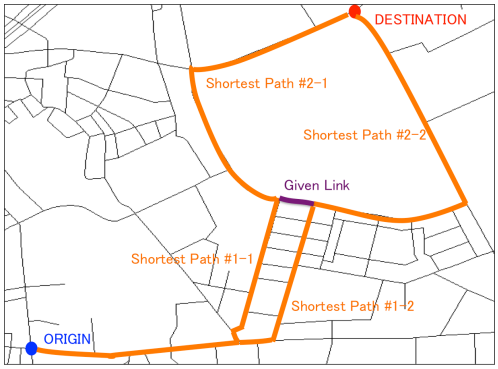

Figure 4. Application of GSP to the proposed method

4.2 Proposed data-oriented algorithm

Actual observed data is used to select a given link. As a given link should be well used to generate a route whose choice probability is high, the usage frequency of each link is calculated by the observed data.

Figure 5 is a flowchart of the proposed method to generate N routes. At first set S is prepared. Initially, S includes all links with the potential to be used to move from the origin to the destination. After this preparation, the best-used link l_i from S is selected as a given link. Then, the shortest path involving l_i is generated from all links, and the links included in it are eliminated from S. This process constitutes one cycle, which is iterated until the number of generated routes becomes N.

Figure 6 is an example of the first two iterations of this process. Initially, S includes all links in the figure, and the best-used link

is selected from S. Then, the first route with the best-used link as the given link is generated as in the upper figures. After the first route is generated, links included in it are eliminated from S, the next best-used link is searched, and the second route is generated as in the lower figures. This cycle is iterated until the N-th route is generated.

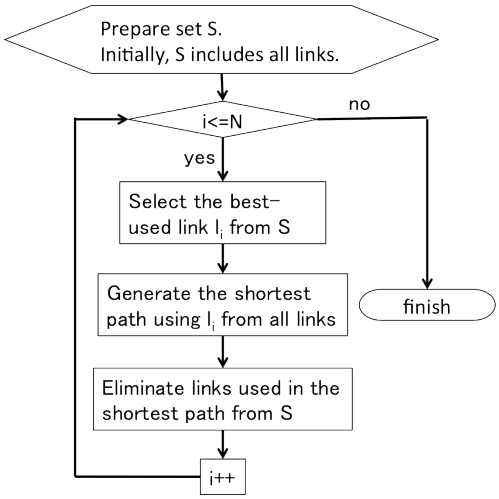

Figure 5. Flowchart of the proposed route generation algorithm

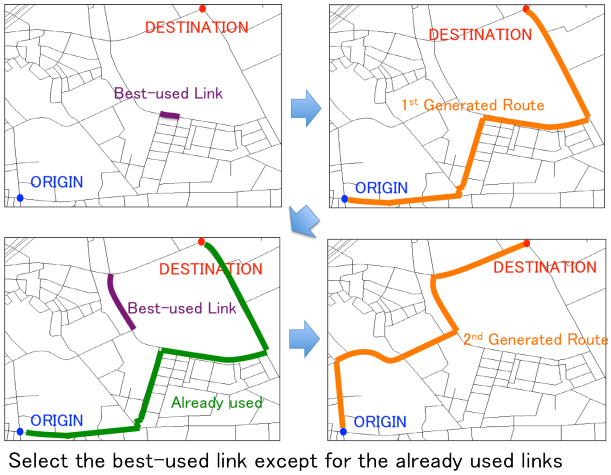

Figure 6. Example of the proposed method

5. EXPERIMENTAL STUDY

5.1 Result of GPS data cleaning

First, we describe a result of GPS data cleaning. The target area is the center of Tokyo, which includes Tokyo station and Shinjuku station; these stations are set as the origin and destination, respectively, as shown in **Figure 7**. The background lines in **Figure 7** show the road network of DRM, along with the locations of Tokyo station and Shinjuku station shown as red circles.

A total of 65965 mobile phones are detected in this target area, as shown in **Figure 8**. Many points are on the road network; these points represent GPS data during a single day and each trip has a separate color.

After extraction by the proposed data processing method, we can collect 23 trips in this case, as shown in **Figure 9**. The lines in **Figure 9** connect the time-consecutive GPS data for each trip.

As shown in the figure, we can extract the usable data to some extent. However, some of the GPS data are outliers. This is one of the most difficult problems when using GPS data. In addition, as the time intervals of the GPS data used in this study are more than 5 min, finding an outlier and removing it is a difficult task. Therefore, it cannot be avoided. On the other hand, the proposed algorithm of route choice set generation has an advantage to outliers. Even if the extracted data contain some outliers, we can scale back their influence by using large amounts of raw data because outliers do not gather at the same location. If an adequate amount of data is used, well-used links are selected as the given links and the nearest links from outliers are not selected.

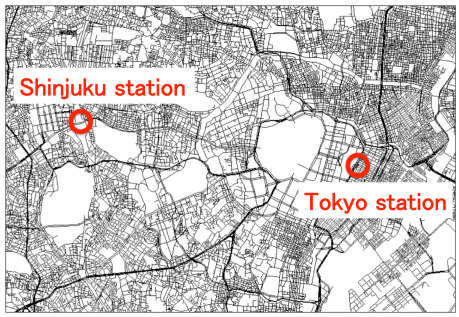

Figure 7. DRM road network in the Tokyo and Shinjuku Area

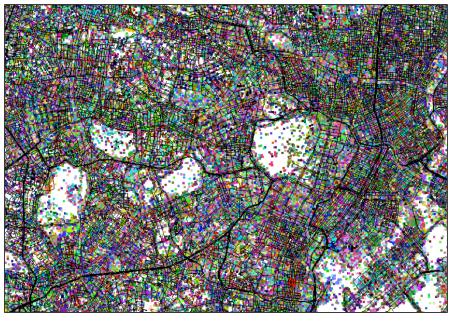

Figure 8. Single-day GPS data in the Tokyo and Shinjuku Area

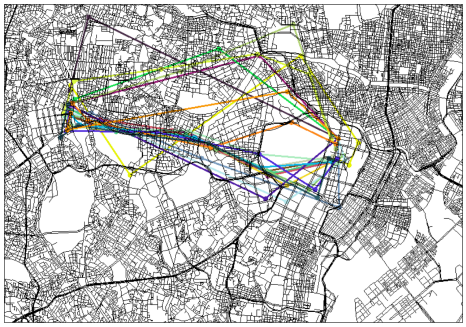

Figure 9. Data extracted by proposed data processing method

5.2 Result of route choice set generation

In this section, we describe the results of route choice set generation. The proposed method is applied to two areas: the same area in Tokyo as in the previous section, and in Chiba. The Tokyo area is the most congested in Japan, with many subway lines and stations, a complicated road network, and much traffic. As Chiba is adjacent to Tokyo, a certain amount of traffic exists. In contrast, as there are no subway stations in this

area of Chiba, GPS data cleaning is easier than the area in Tokyo.

We first consider the results of the area in Chiba. All GPS data for which the period is one year, after cleaning up the data, 1053 trips remain, are used. **Figure 10** shows the usage frequency of each link in the target area because of calculations by the extracted data. Here, the best-used 50 links are red, the next 50 links are green, and the next 50 links are blue. In this experiment, we generate three routes. As shown in **Figure 11**, the proposed method with the calculated usage frequency can generate a multiple types of route. **Figure 12** shows a result of route generation by the K-shortest path algorithm. Although three routes are generated, they are almost all identical and few links are changed. A similar result often occurs when algorithms that do not incorporate real data generate routes. Even if these similar routes are chosen in reality, different links between the three routes also have high usage frequency and the proposed method can generate them.

Table 2 shows the cover rate of the generated routes and the average of the total length. The cover rate is calculated as follows.

$$CoverRate = \frac{\sum_{l_i \in S} a_{l_i}}{\sum_{l_j \in S_{all}} a_{l_j}} \quad (1)$$

where S = set of links included in the generated routes
S_{all} = set of all links
l_i = link included in S
l_j = link included in S_{all}
a_l = usage frequency of link l

As a result, the cover rate of the proposed method is about 15% better than that of the K-shortest algorithm, and the average of the total length is about 10% longer. It can be said that the proposed method reflects the real data more than the K-shortest algorithm without an excessive increase in total length.

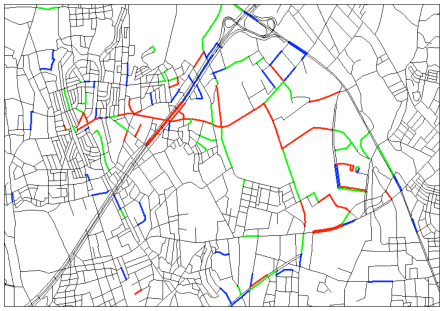

Figure 10. Usage frequency of each link in Chiba

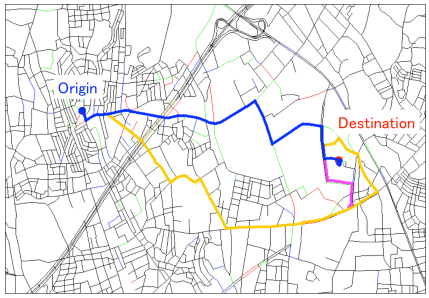

Figure 11. Routes generated by the proposed method in Chiba

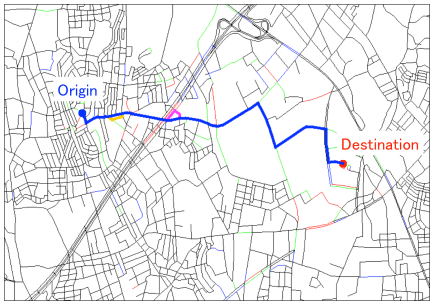

Figure 12. K-shortest routes (K=3) in Chiba

Table 2. Cover rate and average of total length of the result in Chiba

	Proposed method	K-shortest algorithm
Cover rate	55.43%	40.49%
Average of total length	5675 m	4620 m

Secondly, the proposed method is applied to the Tokyo area, where there are far more roads than Chiba and many subway stations, as shown in **Figure 13**. Therefore, the GPS data used to generate routes contain more errors. In Tokyo, we use data from a time span of 2 months, from August 1 to September 31, 2010. This is because we can collect an adequate amount of extracted data from 2-months worth of raw data. In this case, the number of extracted trips is 2299.

Figure 14 shows the result of the proposed method. As is the case in Chiba, we can generate multiple types of route, although the road network becomes more complicated and errors increase. Moreover, the routes generated by the K-shortest path algorithm are about the same just as in the previous case.

In terms of the cover rate, the proposed method is better than the K-shortest path algorithm and the average of the total length does not substantially increase. That the cover rate is lower in the latter than in the proposed algorithm demonstrates that the K-shortest path algorithm has more difficulty generating routes than the proposed method. Hence, we can generate acceptable routes. However, there is some room for improvement, and the two routes given by the proposed method are similar. This is because the method finally generates the shortest route. Therefore, it is worth adding the usage frequency to the cost function to generate the minimum cost route.

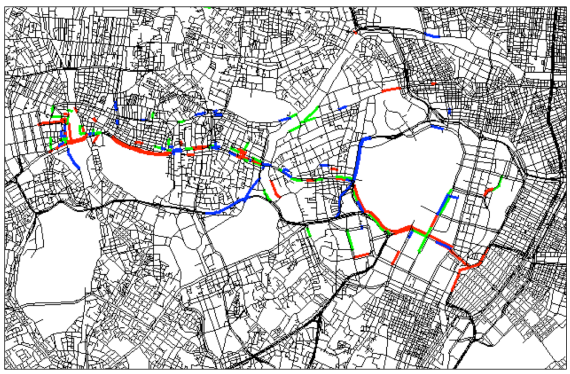

Figure 13. Usage frequency of each link in Tokyo

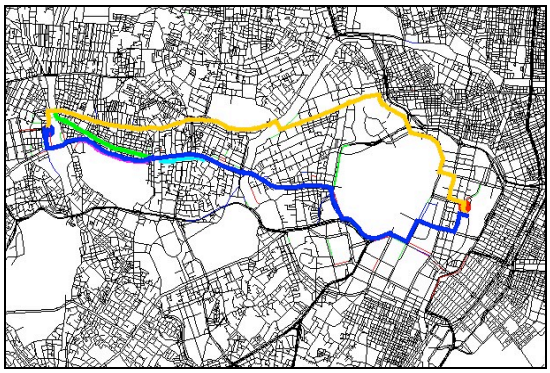

Figure 14. Routes generated by the proposed method in Tokyo

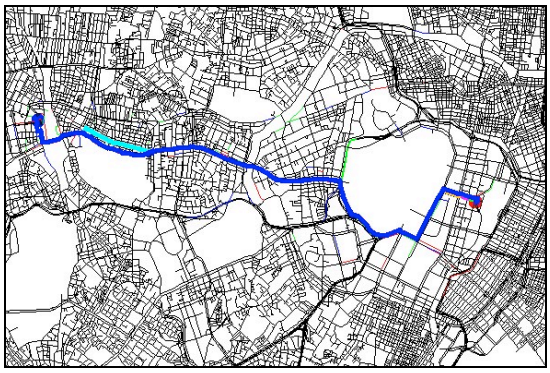

Figure 15. K-shortest routes (K = 3) in Tokyo

Table 3. Cover rate and average of total length of the result in Tokyo

	Proposed method	K-shortest algorithm
Cover rate	21.70%	12.35%
Average of total length	7837 m	7725 m

At last, we validate the improvement of GPS measurement errors by using large amounts of raw data. **Figure 16** shows the cover rate of generated routes with every 10% use of extracted data. Until 40% use of extracted data, the cover rate is getting better linearly. On the other hand, it is stable more than 40% use of them. Therefore, it can be said that the proposed system including GPS data cleaning process and route generation algorithm is robust over GPS measurement errors by using an adequate amount of observed data.

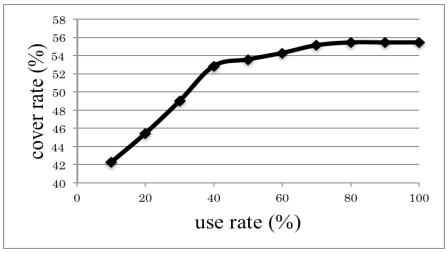

Figure 16. Cover rate by every 10% use of extracted data

6. CONCLUSIONS

We have proposed a system that generates a route choice set via a generation algorithm and a GPS data-cleaning process. By the proposed system, we can generate routes reflecting the observed GPS data of mobile phones.

The experimental results show that the GPS data-cleaning method works well and can extract useful data from a large amount of raw GPS data. Although some errors may remain in the extracted data, they are eliminated by the route generation phase because the proposed route generation algorithm is robust over errors with an adequate amount of observed data.

This route generation algorithm also works well. By this method, we can generate a huge variety of routes that have high potential to be used in the real world.

In future work, we intend to attempt additional validations for the proposed system, not only with GPS data for route generation, but also with real data of routes that can be used for comparing the generated routes. Moreover, in order to generate routes in all areas of Japan for such purposes as navigation and people flow simulations, it is necessary to consider the database architecture in terms of computational cost. In addition, for application of the proposed method to simulations of people flow, we should carefully consider how to calculate the choice probability of each generated route.

References:

Ben-Akiva, M., M.J. Bergman, A.J. Daly, and R. Ramaswamy, 1984. Modelling Inter Urban Route Choice Behaviour. *Proceedings of the 9th International Symposium on Transportation and Traffic Theory,* VNU Press, Utrecht, pp. 299–330.

Chung, E.-H. and Shalaby, A., 2005. A Trip Reconstruction Tool for GPS-based Personal Travel Surveys. *Transportation Planning and Technology*, 28(5), pp. 381–401.

De la Barra, T., B. Perez, and J. Anez, 1993. Multidimensional Path Search and Assignment. *Proceedings of the 21st PTRC Summer Meeting,* pp. 307–319.

Dial, R.B., 1971. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research*, 5(2), pp. 83–111.

DRM (Digital Road Map) http://www.drm.jp/english/drm/e_index.htm (January 12, 2012).

Lombard, K., and Church, R.L., 1993. The gateway shortest path problem: Generating alternative routes for a corridor location problem. *Geographical Systems 1*, pp. 25-45.

Schuessler, N., and Kay W Axhausen, 2009. Identifying trips and activities and their characteristics from GPS raw data without further information. *Transportation Research Record Journal of the Transportation Research Board No 2105 Transportation Research Board of the National Academies*, p. 1–28.

Akgun, V., Erkut, E., and Batta, R., 2000. On finding dissimilar paths, *European Journal of Operational research 121*, pp. 232-246.