Spatiotemporal Route Estimation Consistent with Human Mobility Using Cellular Network Data

Hiroshi Kanasugi[1], Yoshihide Sekimoto[1], Mori Kurokawa[2],
Takafumi Watanabe[2], Shigeki Muramatsu[2] and Ryosuke Shibasaki[1]
[1]The University of Tokyo, Chiba, Japan
[2]KDDI R&D Laboratories Inc., Saitama, Japan
yok@csis.u-tokyo.ac.jp, sekimoto@csis.u-tokyo.ac.jp, mo-kurokawa@kddilabs.jp,
tk-watanabe@kddilabs.jp, mura@kddilabs.jp, shiba@csis.u-tokyo.ac.jp

Abstract— **Continuous personal position information has been attracting attention in a variety of service and research areas. In recent years, many studies have applied the telecommunication histories of mobile phones (CDRs: call detail records) to position acquisition. Although large-scale and long-term data are accumulated from CDRs through everyday use of mobile phones, the spatial resolution of CDRs is lower than that of existing positioning technologies. Therefore, interpolating spatiotemporal positions of such sparse CDRs in accordance with human behavior models will facilitate services and researches. In this paper, we propose a new method to compensate for CDR drawbacks in tracking positions. We generate as many candidate routes as possible in the spatiotemporal domain using trip patterns interpolated using road and railway networks and select the most likely route from them. Trip patterns are feasible combinations between stay places that are detected from individual location histories in CDRs. The most likely route could be estimated through comparing candidate routes to observed CDRs during a target day. We also show the assessment of our method using CDRs and GPS logs obtained in the experimental survey.**

Keywords-spaitotemporal route estimation; personal mobility; CDRs

I. INTRODUCTION

Continuous personal position information has been attracting attention in various service and research areas. Although embedding GPS functionality in mobile phones has become a common means of acquiring the positions of their users, issues such as power consumption and indoor positioning remain. As a solution to the aforementioned issues, in recent years, the telecommunication histories of mobile phones (CDRs: call detail records) have been applied to position acquisition and personal behavior analysis. CDRs are recorded in the existing infrastructures of mobile phone carriers; therefore, they can be used to eliminate the additional workload for mobile phones during data acquisition and to acquire large-scale and long-term data on all telecommunication users. Mobile phone carriers should release some CDRs even if there are certain restrictions and if agreement to the terms of usage from each mobile phone user must be obtained; nevertheless, the use of CDRs is attractive and expected in some research fields relevant to human mobility, such as transportation, disaster reduction, and urban development.

Originally, the purpose of recording CDRs has been to discover problems with telecommunication infrastructures and resolve them expeditiously. And also CDRs have been fundamental for generating billing data. When making a call, sending a text, or browsing the internet using a mobile phone, timestamp and position information of the connected base station are recorded as CDR data. Therefore, the temporal resolution of CDRs differs for each person according to the mobile phone communication pattern. In addition, the spatial resolution of CDRs is lower than that of existing positioning devices such as GPS because they depend on the coverage area of the base stations. Accordingly, interpolating the spatiotemporal positions of such sparse CDRs in accordance with actual human behavior will facilitate relevant services and researches.

In this study, we attempt to estimate a user's personal route over an entire day based on spatiotemporal similarity using CDRs as spatially sparse personal footprints. As the basic principle, the estimation is conducted by generating as many candidate routes as possible using trip patterns interpolated using the shortest path of road and railway networks, and selecting the most likely route from them. In particular, candidate routes in the spatial domain are exhaustively generated based on stay places detected from individual location histories in CDRs and trip patterns consisting of the shortest paths between them. The spatial route is then determined by selecting the nearest candidate route based on the shape from the trajectory of CDRs on the target day of estimation.

When generating the candidate route, we only employ the shortest paths and exclude the individual diversity of route selection and traffic information. Additionally, to consider the diversity of the occurrence time of the trips in days, the temporal patterns of the spatial route are generated by shifting the occurrence time of the trips. Finally, the most likely route in the spatiotemporal domain could be estimated by comparing the likelihood values of the observed CDRs on the target day. The proposed estimation method is assessed by CDRs and GPS logs obtained from the experimental survey that is conducted for 25 days with 184 examinees.

The remainder of the paper is as follows. In section II, we describe related work based on CDRs in particular, and in

978-1-4244-9529-0/13/$31.00 ©2013 IEEE

section III, we explain the proposed methods. In section IV, we present the results of the experiment. Finally, in section V, we summarize the results and suggest future directions for research.

II. RELATED WORKS

Personal position information, if obtained continuously and accumulated regionally, helps grasp personal mobility characteristics and gain an overview of the time-varying population distribution. Ashbrook et al. obtained long-term GPS data on several people and predicted their mobility patterns by estimating significant locations from GPS data [3]. However, these data did not involve temporal characteristics or specific trajectories. Although Froehlich et al. also acquired long-term GPS data from hundreds of participants, they attempted to predict driving routes by clustering individual trips extracted from GPS logs excluding temporal characteristics [4]. Because there are numerous existing studies related to mobility estimation using GPS logs, some studies would be applicable to CDR-based estimation by regarding CDRs as spare position data.

Given that CDRs already cover a widespread area globally and that telecommunications data on individuals have been continuously recorded, many studies have employed CDRs for mobility analysis. Some studies have provided overviews of time-varying population distributions by summarizing and estimating the number of people staying around specific areas [6][10]. On the other hand, by regarding CDRs as footprints of personal movements, other studies have attempted to predict people's mobility by extracting significant locations such as the origins and destinations of trips [5][7][8][9]. Although most of the studies usually organize CDRs into statistics around certain areas or base stations, there are few studies attempting to estimate consistent spatiotemporal routes from CDRs.

III. METHODS

In this section, as well as through Figure 1, we describe methods for spatiotemporal route estimation as follows. In subsection A, we explain the method to detect stay places from CDRs, and in subsection B, we describe how candidate routes are generated from stay places and how the most appropriate route is identified from them. Subsequently, in subsection C, we present the method of temporal pattern estimation through reallocation of the duration of stay. Finally, the validation method for the estimated route with GPS logs is summarized in subsection D.

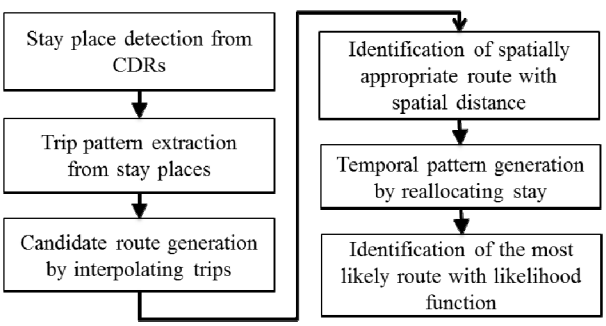

Figure 1. Outline of route estimation procedure

A. Stay Place Detection from CDRs

Here, we describe methods to identify stay places and time intervals of each user based on individual location histories $\{P_C(t):t = t_0,\ldots,t_n\}$, where $P_C(t)$ represents a two-dimensional spatial point at time t in CDR data. First, to investigate whether the user stayed or moved within each time segment, we employed a uniform approach for time segmentation using sliding windows with a width T and a shift S. In other words, let the initial time be t_0. Then, the first time segment is $[t_0, t_0+T)$, the second time segment is $[t_0 + S, t_0 + T + S)$, etc. The mean-shift procedure [1] is then applied to each point belonging to each time segment. The mean-shift procedure is used to seek the mode of the density of spatial points. Every iteration of the procedure calculates the mean $m(P_C(t))$ of nearby points of $P_C(t)$ within a window determined by a window function $K(\cdot)$ and shifts points $P_C(t)$ to $m(P_C(t))$ until they converge.

$$m(P_C(t)) = \frac{\sum_{P_C(t')} K(P_C(t') - P_C(t)) \cdot P_C(t)}{\sum_{P_C(t')} K(P_C(t') - P_C(t))} \quad (1)$$

As for the window function $K(\cdot)$, we use a rectangular window defined as follows: $K(P_C(t)) = 1$ if $\| x \| < h1$, and 0 otherwise, where the parameter $h1$ corresponds to the bandwidth of the density estimation. Convergence is checked by evaluating the difference in the mean points $\| m(P_C(t)) - P_C(t) \| < h2$. We then determine that a user stayed in a time segment if the resulting mean points of the segment are concentrated in the range of a circle with a radius of $h2$.

Subsequently, we cluster individual location histories in the spatiotemporal domain to obtain stay places and time intervals through the following two procedures. First, to obtain stay places, the spatial points are clustered by reapplying the mean-shift procedure to the resulting mean points only within stay time segments. Second, we determine a series of time segments, in which the first one and the last one belong to the same spatial cluster and the intermediate ones do not belong to different spatial clusters, as the stay time intervals.

B. Spatial Route Estimation Based on Stay Places and Trip Patterns

Here, we explain how candidate routes are generated from individual stay places detected from location histories in CDRs and how the most appropriate route is identified from them. Because our target of estimating the trip pattern in this study is the ordinary mobility route that a user usually takes (i.e., he leaves home in the morning and returns at night), we initially determine the location of each users' home. Here, assuming that users stay home most often, the mode of stay places should simply be home.

Subsequently, we connect two consecutive stay places in a trip, and organize time-ordered trips in a day to a trip pattern. According to individual trip patterns, possible patterns starting and arriving at the home place are

exhaustively enumerated regardless of the occurrence probability of each trip. At this time, we restrict the number of trips between 2 and N_{trip}. In addition, stay places can appear in trip patterns twice or more; for example, a salesman often returns to the office after visiting some places. Then, to determine the spatial routes of possible trip patterns, we interpolate each trip with the shortest path on road and railway networks by the proposed method in [7]. Here, we only employ the shortest paths and exclude individual diversity of route selection and traffic information to simplify the generation of candidate routes; however, such additional information should be integrated for further practical estimation. As the result, we can prepare candidate routes in the spatial domain. Figure 2 shows the procedure involved in preparing spatial routes.

Next, we identify the most appropriate route for a target day of estimation through selecting the nearest route in spatial distance D_S between the CDR trajectory in the target day and candidate routes, both of which are point arrays representing trajectories over an entire day (Figure 3). The spatial distance D_S consists of the following parameters: length of the CDR trajectory N_C, length of the candidate route N_C, CDR position P_C, position in the candidate route P_R, and ellipsoidal distance $dist(P_C(i),P_R(j))$.

$$D_S = \frac{1}{2N_C}\sum_{i=1}^{N_C} dist(P_C(i), P_R(\left[\frac{i \cdot N_C}{N_R}\right])) + \frac{1}{2N_R}\sum_{i=1}^{N_R} dist(P_C(\left[\frac{i \cdot N_R}{N_C}\right]), P_R(i)) \quad (2)$$

Assuming that both point arrays have the same length owing to expressing the trajectory over an entire day, the spatial distance can be calculated as the average distance between points in approximately the same order. In addition, if consecutive points are recorded in CDRs, they should be preliminarily removed for eliminating bias and for obtaining an accurate spatial distance. However, the spatial distance becomes a high value in some cases out of our estimation target, such as routes where the user does not come back home or he/she stays in a place all day. We then define the threshold value T_d for spatial distance to eliminate exceptions.

C. *Temporal Pattern Estimation Through Reallocating Duration of Stay*

Although the selected candidate route is spatially appropriate, it does not contain the occurrence time of trips and duration of stays. The duration of stay is not always the same even if the spatial route and duration of the trip are identical. That is, the duration of stay should be diversified to correspond to the occurrence time of the trips. Therefore, we generate temporal patterns for the selected candidate route by reallocating the duration of each stay place with the unit time duration T_S. We then obtain the most likely route as an estimation result based on the likelihood function. By assuming that the CDR is recorded when a user is within radius R from a base station, the likelihood function L is

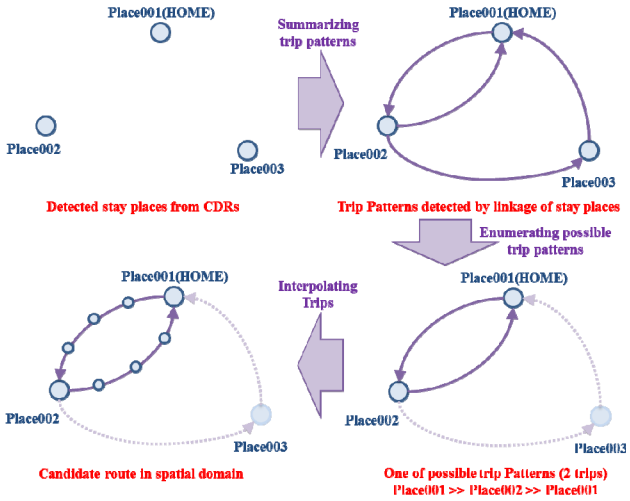

Figure 2. Procedure of generating feasible routes

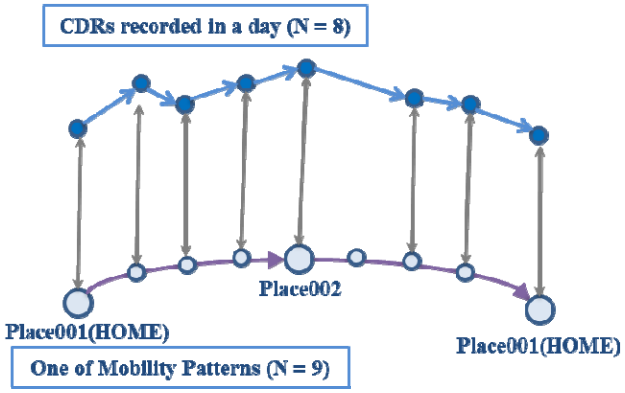

Figure 3. Simplified image of spatial distance calculation

defined by the uniform distribution with R as follows. The ε value denotes a value sufficiently lower than 1.

$$L = \sum_{i=1}^{N_C} \log f(P_C(t_i), P_R(t_i)) \quad (3)$$

$$f(P_C(i), P_R(j)) = \begin{cases} \dfrac{1-\varepsilon}{\pi R^2} & dist(P_C(i), P_R(j)) \le R \\ \varepsilon & otherwise \end{cases} \quad (4)$$

Likelihood can be calculated by comparing candidate routes diversified temporally with observed CDRs in the target day of estimation.

IV. EXPERIMENTS

In this section, we explain the experimental results for assessing the proposed methods with practical CDRs observed in an experimental survey and the consideration about both results with high and low accuracy.

A. *Summary of Experimental Data*

From November 28, 2011, to December 22, 2011, we conducted the experimental survey with 184 examinees to obtain activity data: CDRs, GPS logs, activity status data from a web diary, and personal attributes via a questionnaire. Examinees consented to the privacy policy and terms of the experiment. In the survey, the average number of accumulated CDRs in a day was 454.9 for each examinee. That is, one telecommunication event occurs every 3 min. Although the Android application for GPS logging in every 5 min generates a telecommunication event every 15 min to send logs, more frequent telecommunication events were generated during ordinary activities.

For validating the proposed estimation methods mentioned in the previous section, we employ CDRs and GPS logs in the experimental data. Subsequently, as the baseline data to generate candidate routes for route estimation, we extract 1 week of CDR data from November 28 to December 22 and apply the proposed methods to a day arbitrarily selected from the survey period. For the validations, the optional parameters are set as follows: $T = 20$ min, $S = 5$ min, $h1 = 4{,}000$ m, $h2 = 1{,}000$ m, $N_{trip} = 5$ trips, $T_d = 10{,}000$ m, $T_S = 30$ min, $R = 3{,}000$ m, $\varepsilon = \exp(-10)$.

B. *Validation Method for Estimation Result*

Here, we describe the method for validating the estimated result obtained with the methods proposed in the above subsections. To measure the accuracy by GPS logs obtained together with CDRs, we defined the following validation function. The validation function calculates the average distance E between GPS points $\{P_G(t):t = t_1,\ldots t_m\}$ and points in the estimated result at the same time slice.

$$E = \frac{1}{m}\sum_{i=1}^{m} dist(P_R(t_i), P_G(t_i)) \qquad (5)$$

C. *Estimation Results and Consideration*

After applying the estimation methods to individual data, we can obtain available results about 129 examinees. The route estimation for the remaining examinees cannot operate correctly because of some exceptions; there are no CDRs or GPS logs in the target day of estimation, or no trip patterns are generated owing to an insufficient number of stay places. Therefore, we describe the estimation results and consideration about 129 examinees as follows.

First, in the detection of stay places from CDRs, the average number of stay places per day for each examinee is approximately 2.9, and the average number of total stay places within the baseline period is approximately 4.4, as shown in Figure 4. As well as home and office, where ordinary people usually visit, a few additional stay places are detected from activities within a week. However, because we configured the bandwidth of density estimation to 4,000 m, close range movements such as a short stay at a convenience store nearby the home are concentrated into a single stay place. Considering the effective range of the base station,

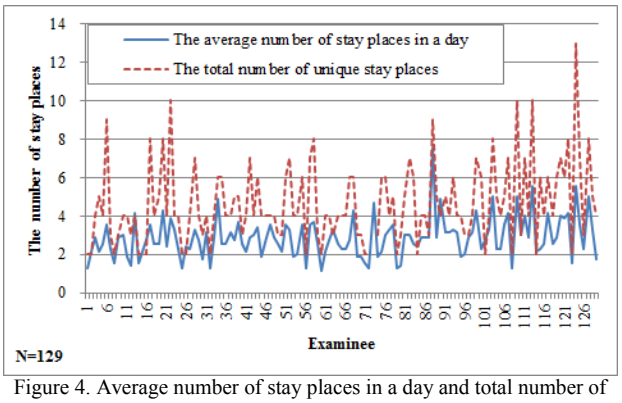

Figure 4. Average number of stay places in a day and total number of stay places during baseline period for each examinee

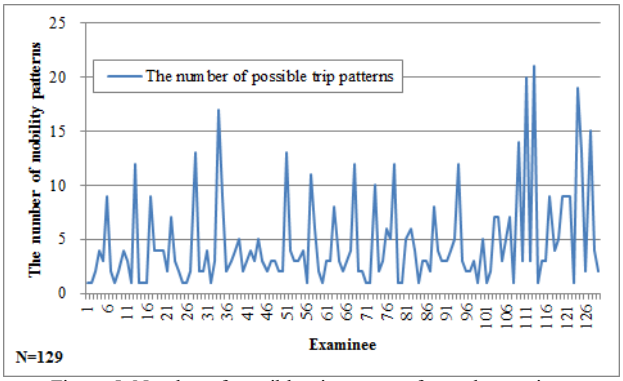

Figure 5. Number of possible trip patterns for each examinee

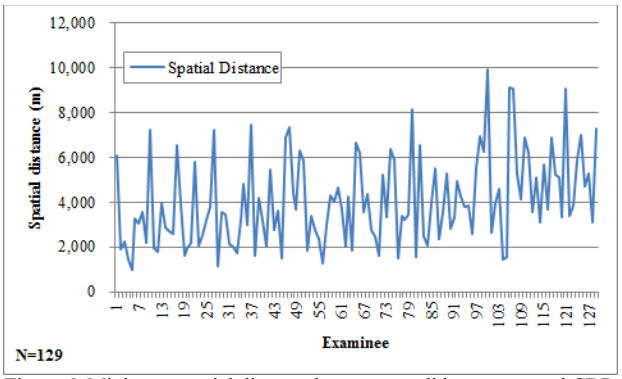

Figure 6. Minimum spatial distance between candidate routes and CDR locations in target day of estimation

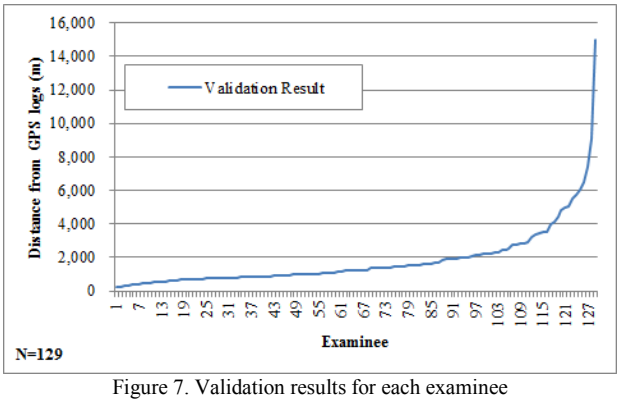

Figure 7. Validation results for each examinee

further detailed detection is difficult with the proposed methods.

Figure 5 shows the number of possible trip patterns generated from stay places for each examinee; there are approximately 4.9 possible trip patterns per examinee. In comparison with the total number of stay places, the number of exhaustive trip patterns within five trips is small. We consider that trip patterns to visit certain stay places might be determined in general.

Figure 6 shows the spatial distance between the most appropriate route interpolated by the shortest path and location histories of CDRs in the target day. The average distance per each examinee is about 4 km; however, the distances between examinees are inhomogeneous. We consider that interpolating trips with the shortest path causes this result because mobility routes are not always the shortest path and depend on personal situations such as transportation cost. Subsequently, the validation result shown in Figure 7 denotes that the average distance per examinee is approximately 1.8 km. The result indicates that the estimated routes with the proposed methods are relatively close to actual trajectories.

The individual details of the accurate result are shown in Figure 8 and Table 1. According to Figure 8, the estimated route in red almost overlaps the GPS logs in yellow except near home, which is represented by the left-upper stay place in white. The examinee seems to use the railway for commuting; however, the estimated route selects a different railway station from the actual one on the shortest path interpolation. Although the number of detected stay places is only three, as represented in Figure 8, a more realistic route than the CDR trajectory can be estimated with the proposed methods. On the other hand, Figure 9 and Table 2 show the other inaccurate result of the route estimation. Because the proposed methods employ CDR histories for baseline data to generate candidate routes, mobility routes outside of the CDR histories are unavailable for our estimation methods. In addition, in the spatial distance calculation of candidate routes, because both the CDR trajectory and candidate routes are considered to represent movement over the entire day, the difference in the trajectory length causes the estimation result to worsen. The remaining issues lie in the generation of a wide variety of candidate routes, and not only in the shortest path, and in the improvement of the method for determining the spatial distance.

V. Conclusion and Future Tasks

A. Conclusion

In this study, we attempt to estimate a personal mobility route with spatiotemporal consistency over an entire day based on CDR data. As for the estimation methods, candidate routes in the spatial domain are exhaustively generated based on stay places detected from individual location histories in CDRs and trip patterns with the shortest path between them. The spatial route is then determined by identifying the nearest candidate route based on shape from the trajectory of CDRs during the target day of estimation. Additionally, to consider the diversity of the occurrence time

Table 1. Parameters obtained in the estimation procedure for an accurate validation result

Content	Value
Estimated route type	Railway
The number of possible trip patterns	2
The number of CDRs in the target day	1,002
Spatial distance	6,284 m
Log-likelihood	-18,707
The number of GPS points in the day	119
Validation result (distance from GPS logs)	982 m

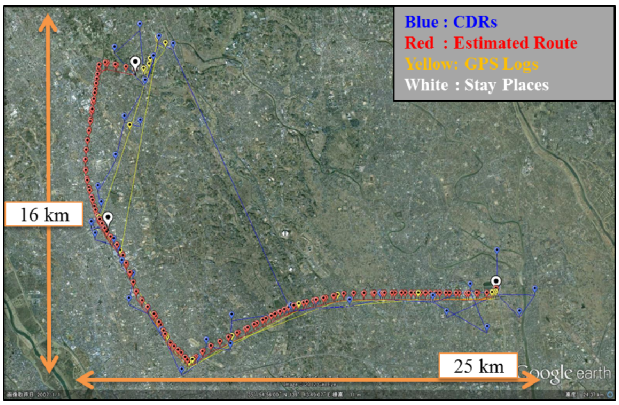

Figure 8. Estimated route, CDRs, and GPS logs for examinee of accurate validation result

Table 2. Parameters obtained in the estimation procedure for an inaccurate validation result

Content	Value
Estimated route type	Road
The number of possible trip patterns	3
The number of CDRs in the target day	2,370
Spatial distance	6,889 m
Log-likelihood	-46,994
The number of GPS points in the day	208
Validation result (distance from GPS logs)	2,789 m

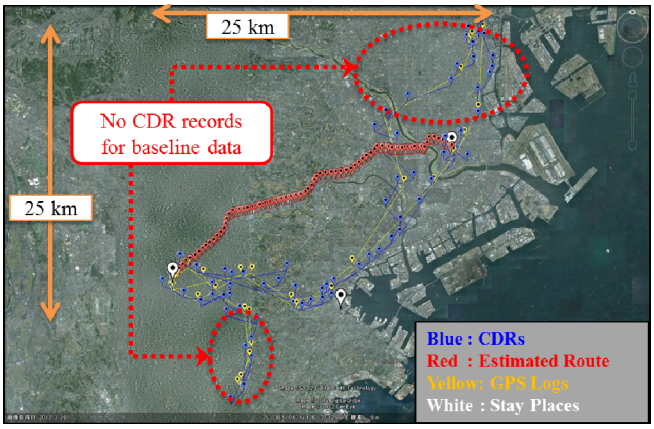

Figure 9. Estimated route, CDRs, and GPS logs for examinee of inaccurate validation result

of the trips on different days, the temporal patterns of the spatial route are generated by reallocating the duration of stays. Finally, the most likely route in the spatiotemporal domain could be estimated by comparing the likelihood to the observed CDRs in the target day. The proposed methods are assessed by CDRs and GPS logs obtained by the experimental survey, with the result that the average distance between the estimated routes and GPS logs per examinee is approximately 1.8 km.

B. Future Tasks

Although we employ the shortest path for route interpolation of trip patterns, another interpolation for generating a wide variety of candidate routes is necessary. Moreover, in addition to improving the method for determining the spatial distance between candidate routes, additional methods for improving the estimation using CDRs in the target day such as trip segmentation for simplifying the spatial distance evaluation should be considered.

ACKNOWLEDGMENT

This work was supported by the GRENE (Environmental Information) project of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and partially funded by a Grant-in-Aid for Young Scientists from MEXT.

REFERENCES

[1] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", IEEE Trans. Pattern Anal. Mach. Intell, Vol. 17(8), pp. 790-799, 1995

[2] R. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Vasharsvky, and C. Volinsky, "Route classification using cellular handoff patterns", Proc. of the 13th ACM International Conference on Ubiquitous Computing, pp. 123-132, 2011

[3] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users", Personal and Ubiquitous Computing, Vol. 7, 275-286, Oct 2003, DOI= http://dx.doi.org/10.1007/s00779-003-0240-0

[4] J. Froehlich and J. Krumm, "Route Prediction from Trip Observations", Society of Automotive Engineers (SAE) 2008 World Congress, paper 2008-01-0201, 2008

[5] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti, "Estimating Origin-Destination Flows Using Mobile Phone Location Data", IEEE Pervasive Computing, Vol. 10, No. 4, pp. 36-44, 2011, DOI= http://dx.doi.org/10.1109/MPRV.2011.41

[6] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A Tale of One City: Using Cellular Network Data for Urban Planning", IEEE Pervasive Computing, Vol. 10, No. 4, pp. 18-26, 2011, DOI= http://dx.doi.org/10.1109/MPRV.2011.44

[7] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui and Y. Shimazaki, "PFlow: Reconstructing People Flow Recycling Large-Scale Social Survey Data", IEEE Pervasive Computing, Vol. 10, No. 4, pp. 27-35, 2011, DOI= http://dx.doi.org/10.1109/MPRV.2011.43

[8] S. Isaacman, R. Becker, R. Caceres, S. G. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying Important Places in People's Lives from Cellular Network Data", Proc. of the 9th International Conference on Pervasive Computing, pp. 133-151, 2011.

[9] M. A. Bayir, M. Demirbas, and N. Eagle, "Mobility profiler: A framework for discovering mobility profiles of cell phone users", Proc. of the International Conference on Pervasive and Mobile Computing, Vol. 6, No. 4, pp. 435-454, 2010

[10] T. Horanont and R. Shibasaki, "An Implementation of Mobile Sensing For Large-Scale Urban Monitoring", UrbanSense08, 2008