Real-time passenger location estimation using CDRs and train objects generated from crowdsourced timetables

Takuya Kanno
The University of Tokyo,
Komaba, Tokyo 153-8508,
JAPAN
kanno@csis.u-tokyo.ac.jp

Hiroshi Kanasugi
The University of Tokyo,
Komaba, Tokyo 153-8508
JAPAN
yok@iis.u-tokyo.ac.jp

Yoshihide Sekimoto
The University of Tokyo,
Komaba, Tokyo 153-8508,
JAPAN
sekimoto@iis.u-tokyo.ac.jp

Ryosuke Shibasaki
The University of Tokyo,
Komaba, Tokyo 153-8508,
JAPAN
shiba@csis.u-tokyo.ac.jp

Ubicomp/ISWC'15 Adjunct , September 07-11, 2015, Osaka, Japan
ACM 978-1-4503-3575-1/15/09.
http://dx.doi.org/10.1145/2800835.2800984

Abstract

This paper shows that cell phone Call Detail Records (CDRs) and train objects in GIS, generated from crowdsourced timetable information, can be used to estimate the train on which a specific passenger is riding. Passenger train ride likelihood calculation is first conducted using CDRs and targeted train objects. Then, the results obtained are compared with the trains specified in GPS logs. Empirical results obtained contain both good cases such as trains being estimated from CDRs and train objects corresponding to the train specified in GPS logs and bad cases such as the estimated train not corresponding to the actual train on which the passenger is riding. Therefore, analysis of each case is also carried out. The application of train objects and cell phone CDRs can facilitate analysis of how passengers ride trains, identification of origin-destination rail routes, and congestion/delay reduction.

Author Keywords

Call Detail Records; train objects; crowdsourcing train diagrams; likelihood estimation

ACM Classification Keywords

H.2.8. DATABASE MANAGEMENT: Database Applications: Spatial databases and GIS

1. Introduction

A variety of social issues has been promoted and resolved by analyzing Call Detail Records (CDRs) data, which are now widely available as a result of the pervasiveness of mobile phones.

Because CDR data are obtained continuously, much analysis related with movement pattern has been performed. UN Global Pulse utilizes such data for social issues, such as disaster management, spread of infectious disease, socio-economics, and transportation [10]. In addition, CDR data have been used for behavioral pattern predictions [8] [9], extraction of people’s stays [7], route estimation, and road traffic volume estimation [4]. Recently, origin–destination trips have also been detected using CDR [2].

In addition, CDR's temporal density has increased as a result of the pervasiveness of smartphones in recent times.

Another type of cell phone data called location area update (LAU) can also be utilized. LAU shows which cell tower area a user is in so that calls can be transferred to that user’s cell phone. LAU has been utilized to measure congestion, delay, and passenger flows in the Paris underground transit system using underground cell towers [1].

However, because of the low spatiotemporal resolution of CDR the amount of in-depth research conducted into behavioral patterns at detailed levels, such as studies using GPS is low [3].

Ishizuka et al. [5] detected train commuters using CDRs and GIS information. This paper builds on their research to determine which train a passenger is riding from train timetables. This makes it possible to measure the railway demand at the time of commuting in large cities such as Tokyo, where quite a lot of people commute by train. The results obtained may help to reduce congestion.

In concrete terms, we estimate the maximum likelihood of real-time passenger train rides from crowdsourcing type train objects and smartphone CDRs to obtain complete information about moving data such as routes and trains ridden.

The remainder of this paper is organized as follows. Section 2 describes how train objects are generated from timetable information created via crowdsourcing. Section 3 outlines the proposed method for detecting the movement of commuters from CDRs. Section 4 describes the likelihood method of calculating train objects for the commuting CDRs. Section 5 outlines the experiments conducted and analyzes the results obtained.

2. Train objects from crowdsourcing type timetable

In this section, the method by which train objects are generated from timetables created via crowdsourcing is explained.

A crowdsourcing type timetable database called “Eki.locky” already exists in Japan (“Eki” is the Japanese word for

Cell phone data: Call Detail Records (CDRs)

Call Detail Records (CDRs) are large-scale infrastructure data associated with the operations of mobile networks. CDRs typically comprise information on when call/data telecommunication was conducted and which cell tower was used. CDRs have been attracting increased attention recently because no additional burden is placed on users to acquire information, in contrast to GPS logs.

"Station.") Eki.Locky is an iOS/Android train timetable application that displays the time for the departure of the next train. Timetable information is created via a register and is shared with all users. If the required station timetable has not yet been registered, anyone can create and upload the timetable file. Further, all users are able to access the timetable provided in the application. This is the method used to obtain the timetable data. The timetable information covers 91% of the stations in Japan (9,988 stations). The rest is mainly expired timetables in local areas. Timetable data are stored in NextTrain format. This service's web page URL is http://eki.locky.jp/site/top.

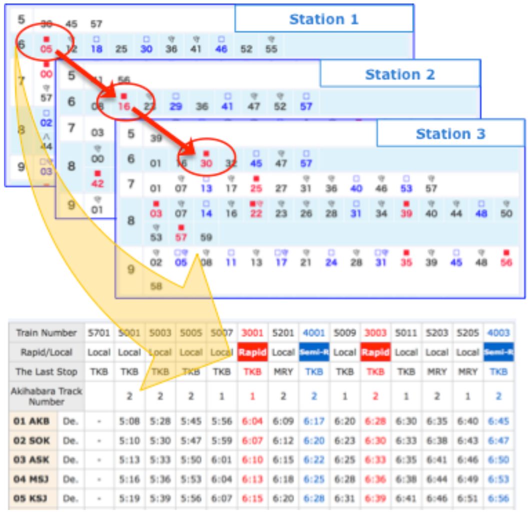
Timetables for each train are generated by reconstructing the departure time of the same train from the timetable information of each station. The station location is geocoded because the timetable does not include the station location on GIS, only the station name. The latitude and longitude of trains are interpolated every minute between the stations using the shape data of the routes, based on the assumption that the train is moving at a constant speed.

In this way, the data [time every minute, latitude, longitude] for each train object are generated from the timetable.

3. Call detail records

CDR is explained in the sidebar on the left. In addition, the details of CDR surveys used in this study are given in the sidebar on the left of the ensuing page. The temporal density using CDRs is especially high because of smartphone telecommunication.

Detecting the trajectory of the commuting CDRs:



Train Number		5701	5001	5003	5005	5007	3001	5201	4001	5009	3003	5011	5203	5205	4003
Rapid/Local		Local	Local	Local	Local	Local	Rapid	Local	Semi-R	Local	Rapid	Local	Local	Local	Semi-R
The Last Stop		TKB	TKB	TKB	TKB	TKB	TKB	MRY	TKB	TKB	TKB	TKB	MRY	MRY	TKB
Akihabara Track Number			2	2	2	1	1	2	2	1	2	1	2	1	2
01 AKB	De.	-	5:08	5:28	5:45	5:56	6:04	6:09	6:17	6:20	6:28	6:30	6:35	6:40	6:45
02 SOK	De.	-	5:10	5:30	5:47	5:59	6:07	6:12	6:20	6:23	6:30	6:33	6:38	6:43	6:47
03 ASK	De.	-	5:13	5:33	5:50	6:01	6:10	6:15	6:22	6:25	6:33	6:35	6:41	6:46	6:50
04 MSJ	De.	-	5:16	5:36	5:53	6:04	6:13	6:18	6:25	6:28	6:36	6:38	6:44	6:49	6:53
05 KSJ	De.	-	5:19	5:39	5:56	6:07	6:15	6:20	6:28	6:31	6:39	6:41	6:46	6:51	6:56

Figure 1: Timetables for each train are generated by reconstructing the departure time of the same train from the timetable information of each station.

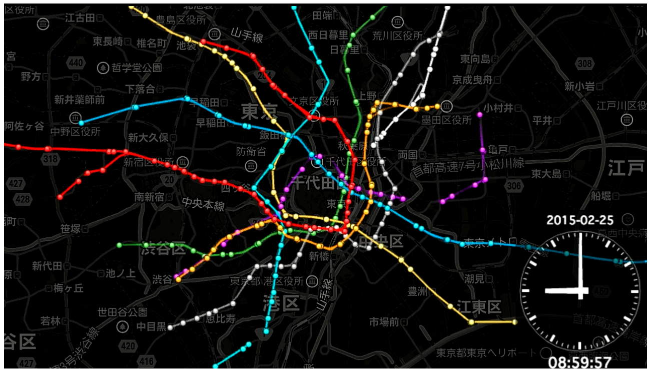


Figure 2: Train objects identified at a particular moment. They are generated by interpolating the timetable of each train. Each color represents a different railway. Approximately 20 train objects are running at the same time on each route.

CDRs survey used in this study

In this study, only the CDRs of survey subjects who consented were used. The information included in the CDR used in this paper was [user ID, communication start date and time, latitude and longitude of the start of communication base stations, communication end date and time of the communication end, and base station latitude and longitude]. Distinct information that could be used to identify the call and communication was not included.

·Survey period: February 1, 2013 to March 1, 2013.

·Acquired data: Both CDS and GPS logs.

·Number of subjects: 48 persons

·Conditions of subjects: Persons using 3G Android smartphones and commuting by car or train.

·Average communication interval from CDRs: 2.9 min

·Number of cell towers in the CDRs: 19,528 (entire country).

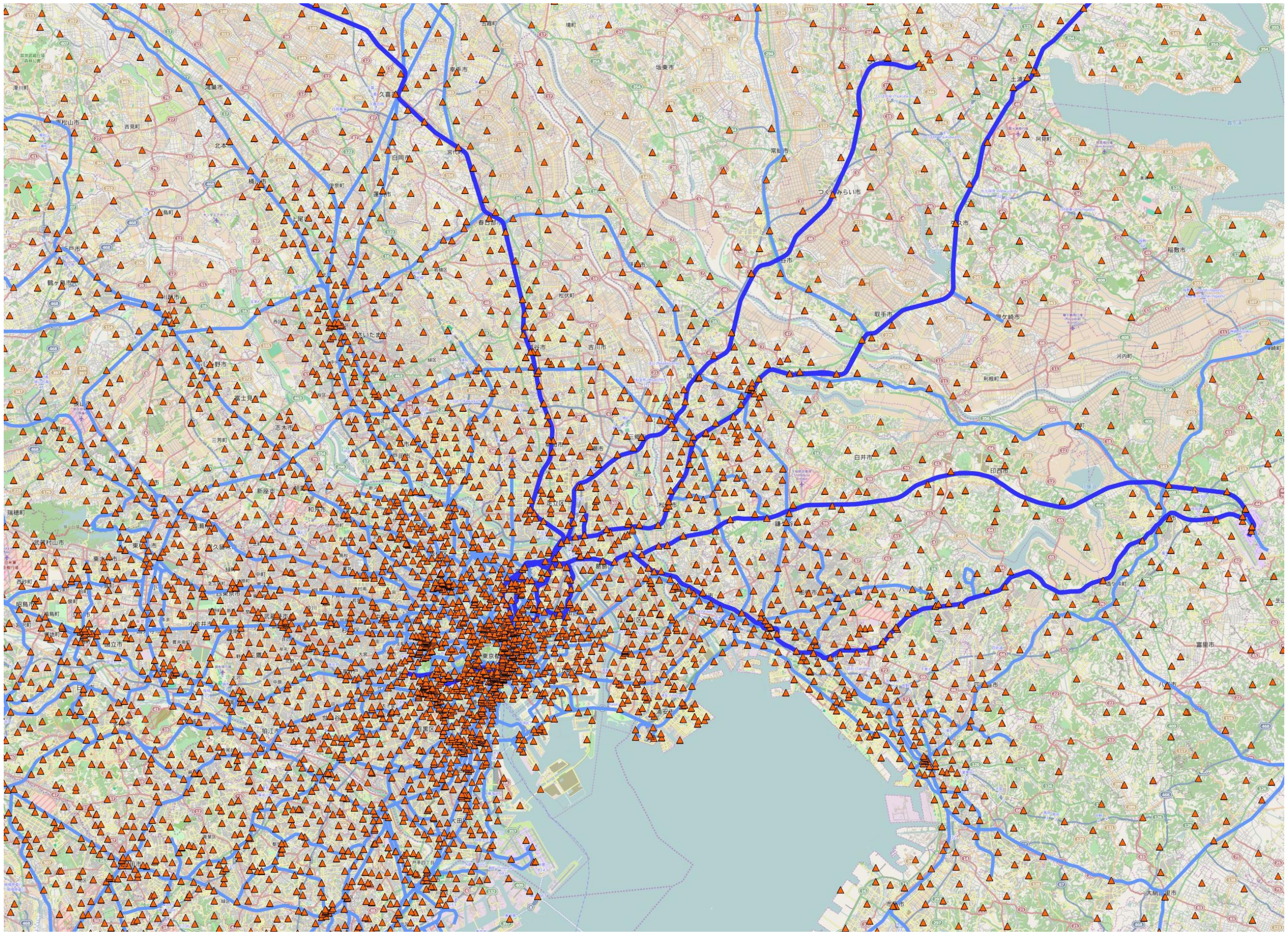


Figure 3: The Tokyo metropolitan railway network (blue lines) and mobile phone base station (orange triangles). Bold blue railways are the lines targeted in this study. Please note that these base stations are included in only this survey's CDRs, not all base stations are in this area.

The stay point of CDRs is detected by clustering the space-time information, as proposed by Kanasugi et al. [6]. To set the trip during the stay, the procedure is as follows. Extract a trip that is moving at 8:00 as the commuting trip. The point sequence of CDRs included in this trip is then set to the commuting CDRs trajectory.

4. Estimation method using the riding of train object from CDRs

In this section, the methods used to obtain the most likely train object as an estimation result based on the likelihood function is described. The likelihood function *L* consists of the following parameters: 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, length of the CDR trajectory N_C, CDR location P_C, and train object location P_T.

By assuming that the CDR is recorded when a user is within radius R from a cell tower, the likelihood function L includes the distance between train objects and the cell tower of the CDR at the point when the CDR is observed. L is defined as

$$L=\sum_{i=1}^{Nc}\log f(\mathrm{P_C}(\mathrm{t_i}),\mathrm{P_T}(\mathrm{t_i})) \qquad (1)$$

$$f(P_C(i),P_T(j))=\begin{cases}\dfrac{1}{\pi R^2} & dist(P_C(i),P_T(j))\le R\\ 0 & otherwise\end{cases} \qquad (2)$$

For sensitivity analysis, calculate the likelihood by changing the radius R (2 km, 3 km, 4 km).

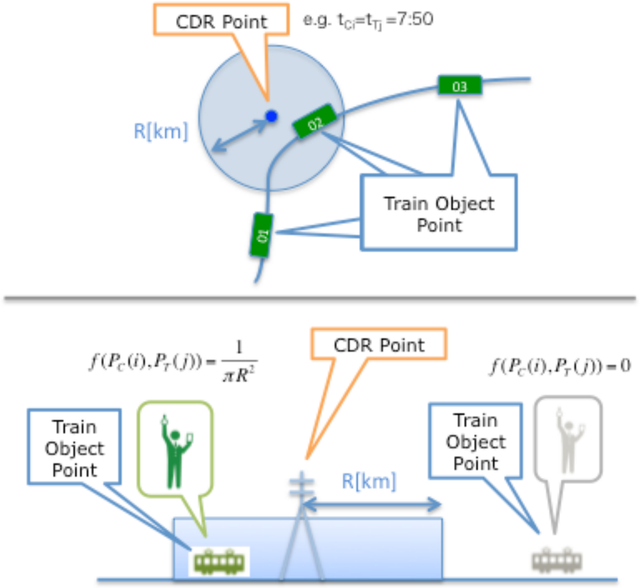


Figure 4: At the point when a CDR is observed, calculate the distance between the train objects and the cell tower of the CDR. The probability of cell phone existence is defined as a uniform distribution within concentric circles from the cell tower. Then, likelihood calculation of the passenger riding the train is conducted. For the sensitivity analysis, calculate the likelihood by changing the radius R (2 km, 3 km, 4 km).

5. Experimental evaluation

5-1. Estimation target

The target railway lines were those on which the most commuting CDRs were detected. These lines were the Joban line, Joban rapid line, Joban local train, Hokuso line, Keisei line, Chiyoda line, Tobu Isezaki line, and Tsukuba express line. These routes are directed toward the center of Tokyo from its north-eastern end. Further, a total of 685 train objects are on the selected railway routes. Four persons met the criteria of using one of the targeted eight railways and commuting CDR trajectory. To compare with the actual train being ridden, we imposed a criterion that they should be included in the GPS logs of the trip. As a result, the number of persons was reduced to three.

5-2. Estimation result

We succeeded in identifying the actual train on which the persons were riding on some trips.

5-2-1. Corresponding results

In this section, successful estimation results are shown. Fig. 5 shows the likelihood L of each radius. There were 85 CDR points.

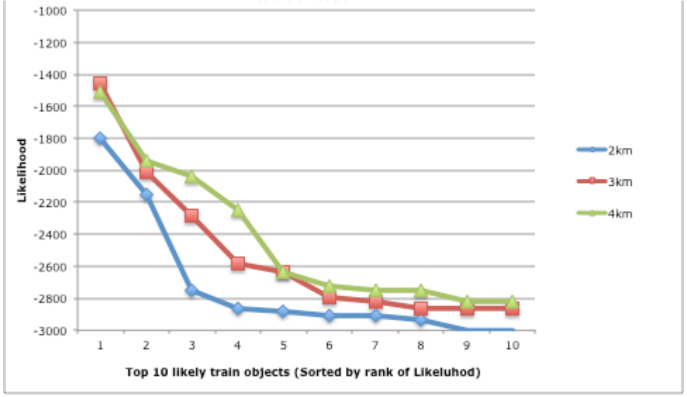


Figure 5: Top 10 likelihood values for the train object. A large difference in the likelihood values exists in the top five train objects. The difference between the remaining objects is small.

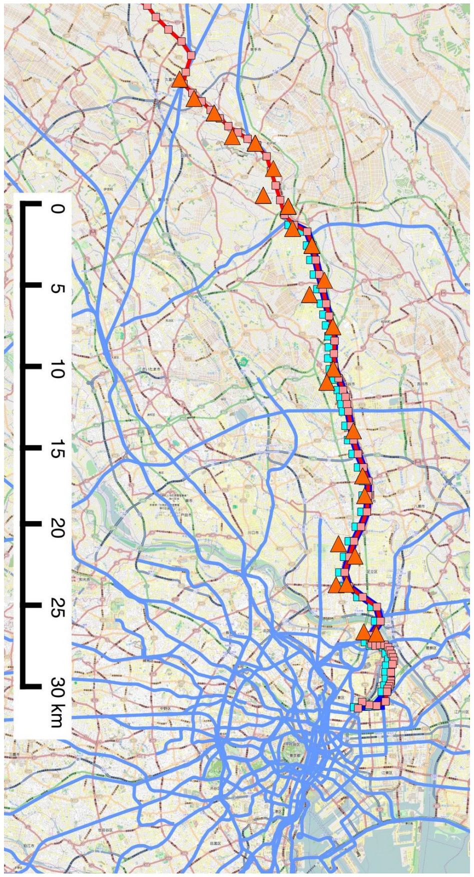


Figure 6: Orange triangles show commuting trajectory of CDR. Light blue squares and deep blue lines show the location per minute and trajectory of the train most likely being ridden. The pink squares and deep red line show the second most likely train's locations and trajectory in experiment A. In the northern area, the highest likelihood train's trajectory and the trajectory of CDR exists, but second most likely train's trajectory does not exist. The presence of a sequence of points in this area increases the likelihood value of the maximum likelihood estimated train.

Table 1: Top three highest likelihood train objects within each radius. The actual train being ridden, as specified by GPS logs, is **TI54**. In this case, the maximum likelihood estimation is correct for each radius

	Train object ID according to likelihood rank to observed passenger's CDRs		
Radius in formula (2)	**Highest likelihood**	**Second highest likelihood**	**Third highest likelihood**
2 km	**TI54**	TI85	TI96
3 km	**TI54**	TI85	TI81
4 km	**TI54**	TI81	TI85

The object ID of the train being ridden as identified from the GPS log is TI54. This is the same as the estimated train object at all radiuses. The distribution point of the CDR is over a wide range of approximately 50 km and may uniformly increase the likelihood value. The CDR trajectory appears to have sufficient points, the distance for which the CDRs exist in the section appears to be sufficiently long, and the distribution appears to be sufficiently uniform.

5-2-2. Result with insufficient likelihood

In this section, an estimation result that is correct only at R = 4 km is presented. Fig. 7 shows the likelihood L of each radius. There are 24 CDRs points.

The object ID of the train being ridden is identified from the GPS log as JF27. This is the same as the estimated train object at only a radius of 4 km. The distribution point of the CDR is over a wide range of approximately 30 km. In the cases where R = 2 km and 3 km, it could be considered that the accuracy decreased because of the assumption that the train object is moving at constant speed did not match the actual movement. In addition, there is the possibility that the train was delayed and may not match the train object.

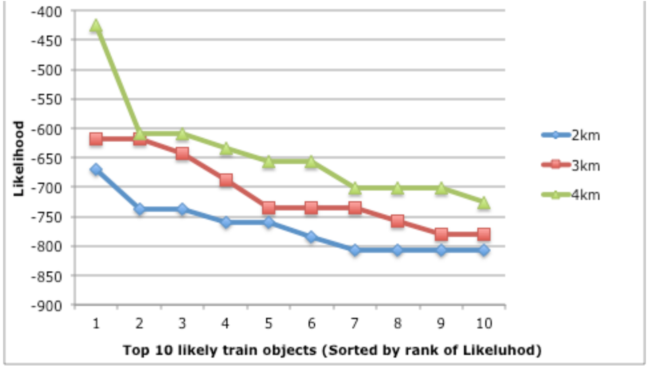


Figure 7: Top 10 likelihood values for the train object. The maximum likelihood in the case of 4 km is significantly higher than that of the case for two positions or less and radius is 2 km or 3 km.

Table 2: Top three highest likelihood train objects within each radius. The actual train being ridden, as specified by GPS logs, is **JF27**.

	Train object ID according to likelihood rank to observed passenger's CDRs		
Radius in formula (2)	**Highest likelihood**	**Second highest likelihood**	**Third highest likelihood**
2 km	JKs20	JKe40	JKe36
3 km	JKs20	JKe36	**JF27**
4 km	**JF27**	JKs20	JKe36

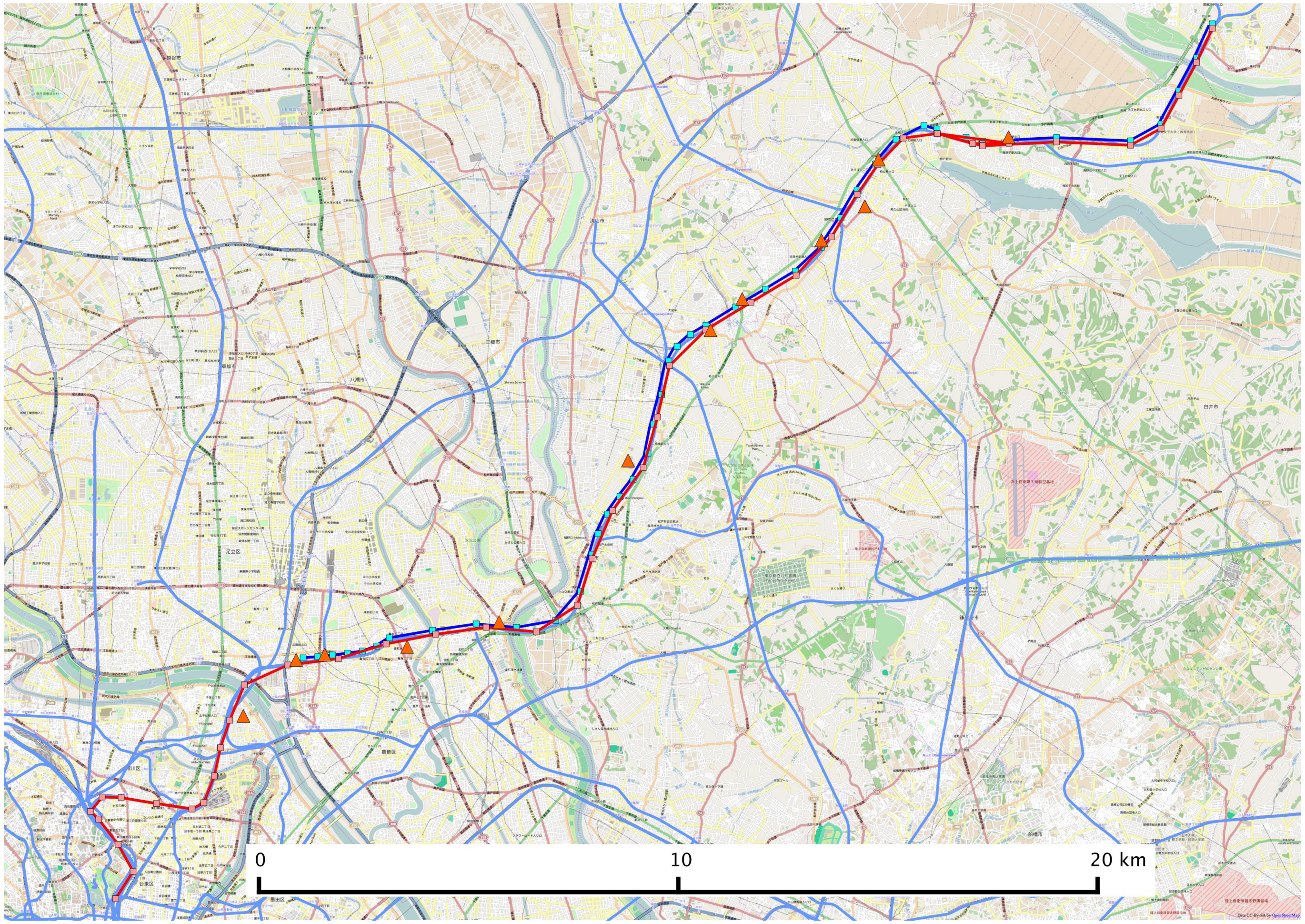


Figure 8: Estimation result is correct at only R = 4 km. The distribution point of the CDR is over a wide range of approximately 30 km. In the cases where R = 2 km and 3 km, we believe that the accuracy decreased because the assumption that the train object is at constant speed did not match the actual movement. In addition, there is the possibility that train was delayed and may not match the train object.

6. Conclusion and future work

6-1. Conclusion

Real-time estimation of the train being ridden by a passenger using crowdsourcing type train object and general CDRs was successful in cases where the CDRs had a wide ranging distribution and uniform distribution made accurate estimation possible.

Under the assumption of constant probability of occurrence of communication within the circle, the result of the likelihood calculation changed according to the radius distance.

6-2. Future work

The method used to detect the commuting CDRs trajectory has to be improved because only a few trips were detected. More extensive evaluation is also required. In addition, improvement of estimation accuracy is required to deal with train delays and passenger transfers.

Acknowledgements

We are grateful for the use of joint research data with KDDI R&D LABS, Eki.locky's train timetable, National Land Numerical Information's railway data (MLIT), and Open Street Map.

References

1. V. Aguiléra, S. Allio, V. Benezech, F. Combes, C. Million. 2013. Using cell-phone data to measure quality of service and passenger flows of Paris Transit System. *Transportation Research Part C: Emerging Technologies*. 43, 2: 198–211. http://dx.doi.org/10.1016/j.trc.2013.11.007

2. L. Alexandera, S. Jiangb, M. Murgaa, M. C. Gonzáleza. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies. http://dx.doi.org/10.1016/j.trc.2015.02.018

3. W. Apichon. 2013. *A study on human activity analysis with large scale GPS data of mobile phone using cloud computing platform*. Ph.D. Dissertation. The University of Tokyo, Tokyo, Japan.

4. Y. Hasegawa, Y. Sekimoto, T. Kashiyama, H. Kanasugi. 2014. Transportation Melting Pot Dhaka: Road-link based traffic volume estimation from sparse CDR data. The 1st International Conference on IoT in Urban Space (Urb-IoT 2014), Rome.

5. H. Ishizuka, N. Kobayashi, S. Muramatsu, C. Ono. 2015. Detecting train commuters using CDRs and GIS information. NetMob2015 book of abstracts::posters, pp.39-41, 2015

6. H. Kanasugi, Y. Sekimoto, M. Kurokawa, T. Watanabe, S. Muramatsu, R. Shibasaki. 2013. Spatiotemporal route estimation consistent with human mobility using cellular network data. In *Proceedings of PerMoby2013*, 267-272.

7. M. Kurokawa, T. Watanabe, S. Muramatsu, H. Kanasugi, Y. Sekimoto, R. Shibasaki. 2013. Extracting people's stays from cellular network data. NetMob2013, 2013.05.

8. M. Ozer, I. Keles, I. H. Toroslu, P. Karagoz. 2013. Predicting the change of location of mobile phone users. MobiGIS2013, pp. 43-50, 2013.

9. M. Ozer, I… Keles, I. H. Toroslu, P. Karagoz, S. Ergut. 2014. Predicting the next location change and time of change for mobile phone users. MobiGIS2014, pp. 51-59, 2014.

10. UN Global Pulse, Mobile Phone Network Data for Development, http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf, 2013.10.