Front. Comput. Sci.
https://doi.org/10.1007/s11704-017-6464-3

RESEARCH ARTICLE

Integrating GPS trajectory and topics from Twitter stream for human mobility estimation

Satoshi MIYAZAWA (✉)[1], Xuan SONG[2], Tianqi XIA[1], Ryosuke SHIBASAKI[2], Hodaka KANEDA[3]

1 Department of Socio-Cultural Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8563, Japan
2 Center for Spatial Information Science, The University of Tokyo, Kashiwa 277-8568, Japan
3 Zenrin DataCom Co'Ltd, Tokyo 108-6206, Japan

Abstract Understanding urban dynamics and large-scale human mobility will play a vital role in building smart cities and sustainable urbanization. Existing research in this domain mainly focuses on a single data source (e.g., GPS data, CDR data, etc.). In this study, we collect big and heterogeneous data and aim to investigate and discover the relationship between spatiotemporal topics found in geo-tagged tweets and GPS traces from smartphones. We employ Latent Dirichlet Allocation-based topic modeling on geo-tagged tweets to extract and classify the topics. Then the extracted topics from tweets and temporal population distribution from GPS traces are jointly used to model urban dynamics and human crowd flow. The experimental results and validations demonstrate the efficiency of our approach and suggest that the fusion of cross-domain data for urban dynamics modeling is more practical than previously thought.

Keywords GPS trajectory, human mobility, SNS, location-based social network (LBSN), topic modeling, data mining, spatiotemporal topic

1 Introduction

In smart-city research, understanding urban dynamics and large-scale human mobility has become one of the major modern challenges [1]. Devices and web services are becoming increasingly interconnected in social and community intelligence infrastructures [2] and data vitalization applications are constantly being developed to communicate with community and end users [3]. Especially, in modern cities with heterogeneous life patterns and high dependence on public infrastructure, city- or nation-scale understanding of human mobility is crucial for urban planning, transportation planning, and emergency response.

Recently, increasingly accurate and precise location data such as Call Detail Records (CDR) and Global Positioning System (GPS) logs have become available and have opened a broad range of research possibilities including understanding regional mobility [4,5] and mobility prediction during a large-scale disaster [6]. However, to fully utilize the data in many estimation, prediction, or classification models, supplemental information with high accuracy and precision is crucial.

One possible solution to this issue is to use data from the Location-based Social Network (LBSN) services. Some users of services such as Twitter, Sina Weibo, and Swarm post content with location tags (geo-tags) usually use GPS devices on their mobile phones. The data are accurate and precise, and given a large amount of data from the same users, the data can be considered trajectories of the users' mobility. Each post represents users' opinions and responses to environments; therefore, through an appropriate technique of opinion min-

Received September 23, 2016; accepted April 13, 2017

E-mail: koitaroh@csis.u-tokyo.ac.jp

ing [7], the textual content of the data can be considered as contextual and mobility information. Thus, data have large information potential, such as topical inference of the trajectories of tourists [8] and event detection and inference [9]. However, compared to the GPS datasets, the trajectory from LBSN is usually sparse and highly biased. To model and understand the city-scale mobility pattern, we need a comprehensive dataset while utilizing rich contextual information from real-time and high-resolution LBSN data. *Therefore, in this research, we aim to assimilate big mobility data from GPS and contextual information from LBSN data in regression models for large-scale human mobility estimation.*

In this study, we collect big and heterogeneous data and jointly use them for regression models using (1) GPS logs of approximately 440 thousand anonymized mobile phone users throughout Japan from July 25, 2013 to July 31, 2013 and (2) Geo-tagged tweets from Twitter during the same period. We employ latent Dirichlet allocation (LDA) topics modeling to discover latent topics from tweets. Then the extracted topics from tweets and temporal population distribution from GPS traces are jointly used to train a regression model to discover and model their relationships.

The key idea of this work is summarized in Fig. 1 and this work has the following key characteristics that make it unique compared to previous research.

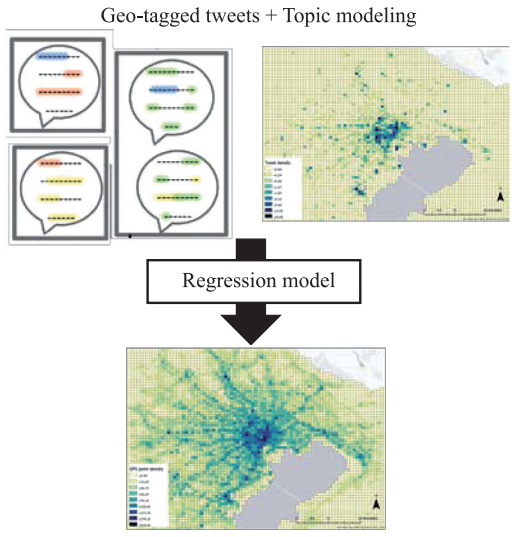


Fig. 1 The key idea of this work (Can we use cross-domain data for human mobility modeling and estimation? We employ topic modeling to discover latent topics from geo-tagged tweets, and estimate population distribution from GPS traces. Credit: "Konzatsu-Tokei®" ©ZENRIN DataCom CO., LTD)

1) Big and heterogeneous data: our research is based on a big and heterogeneous data source. We used the GPS records of 440 thousand users over one week and the related Twitter data.
2) Fusion of cross-domain data: We propose a novel pipeline using state-of-the-art methods to discover the relationships between human mobility data and location-based social network data and apply it to human mobility estimation.

The rest of this paper is organized as follows. We discuss related work in Section 2. The data are introduced in Section 3 and the formal definitions and model specifications are summarized in Sections 4 and 5. In Section 6, we experimentally evaluate our methods using real data, and finally Section 7 concludes the paper.

2 Related work

There have been several studies with location logs and trajectories starting from survey-based [10,11] to large-scale mobile phone data for city-scale or nation-scale mobility [4,5,12]. Traffic analysis is also a recent prominent subject using big data [13–18]. In addition, with supplemental information, application-focused models including disaster evacuation behavior [6,19], residential life patterns [20], public safety [21,22], and regions of different functions [23] have become possible.

As Guo et al. [24] proposed the concept of mobile crowd sensing and computing that leverage both mobile sensing of devices and cognitive intelligence of humans, LBSN data have also been used for understanding urban dynamics. For example, even though less than 1% of tweets in Twitter are geo-tagged [25], Twitter data have been widely used in the fields of Computer Science, GIScience, and transportation engineering [26]. Cheng and Wicks conducted a study to first extract clusters in geo-tagged tweets and then applied LDA to interpret the topics of the clusters [27]. Sakaki et al. estimated locations of earthquakes and typhoons using geo-tagged tweets [28]. Abbasi et al. extracted tourists' mobility patterns from other users [29]. Other studies include understanding topical features from trajectories based on geo-tagged text messages [8], emergency event location estimation [30], application development for improving emergency awareness [31], and urban land use estimation [32]. Some studies focus on the trajectory of the LBSN data to investigate mobility [33], such as cross-border mobility estimation [34] and global mobility patterns [35], suggesting geo-tagged tweets sufficiently reflect global and regional human mobil-

ity.

By combining mobility and social activity data, Pan et al. [36] developed a model to detect traffic anomalies and employed a tf-idf based approach on tweets to discover representative words for each anomaly; however, such application is rare as the number of studies combining multi-modal data for smart city research remains limited.

3 Big and heterogeneous data source

• **GPS data** The GPS trajectories used in this study are from approximately 440 thousand anonymized mobile phone users throughout Japan from June1, 2013 to July 31, 2013, which are processed by NTT DOCOMO, INC[1)]. An app on the phones sends the location of the users every five minutes, provided a GPS signal is available. One important drawback of the data is that the number of points is biased toward daytime and downtown areas, where people usually spend the daytime. This is because the data were collected when mobile phone users were usually on the move.

• **Geo-tagged tweets** The geo-tagged tweets from June 1, 2013 to July 31, 2013 were collected using the Twitter API. To extract tweets concerning mobility and social activity, only tweets posted from check-in services (e.g., Swarm) are used for the following experiment. Figure 2 compares the number of GPS points and the number of tweets at the same locations and shows low correlations, suggesting that estimation of human mobility from geo-tagged tweets is a very challenging task.

4 Hypothesis and preliminaries

4.1 Hypothesis

Nowadays, residents in urban cities move regularly using different transportation infrastructures. Some of them enjoy Twitter or other LBSNs with their mobile devices. They sometimes post their "check-ins" during their commutes and travels. Each post represents their sentiments, lifestyle, or events they observe or participate in. Especially, when they encounter anomalies such as traffic accidents or seasonal events, they tweet more heavily. Collectively, while we assume the collection of tweets represents peoples' sentiments, lifestyles, or events, we expect to observe significantly more tweets about unprecedented events.

What and how people tweet is heavily biased depending on the topics. For example, under normal circumstances, people tweet about their personal thoughts or post responses to other tweets. However, when they observe unexpected events, they tend to be more descriptive of the events. We assume this is the key to understanding the background of human mobility.

4.2 Preliminaries

Human mobility data and social media posts are different modal data and require separate preliminary data processing steps. For human mobility data, GPS trajectories from mobile phones are used for this study. The GPS trajectory is structured as a 4-tuple:

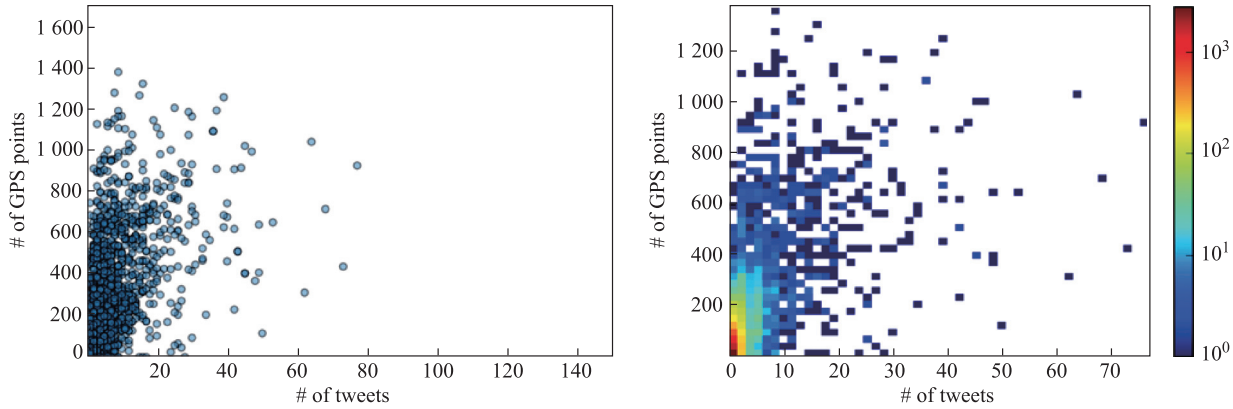


Fig. 2 The number of GPS points and tweets (Low correlation can be observed between the number of GPS points and the number of tweets, suggesting more supplemental features are required for better estimation. Credit: "Konzatsu-Tokei®" ©ZENRINDataCom CO., LTD)

[1)] "Konzatsu-Tokei (R)" Data refers to people flow data collected by individual location data sent from mobile phones with enabled AUTO-GPS function under users' consent, through the "docomo map navi" service provided by NTT DOCOMO, INC. Those data are processed collectively and statistically to conceal private information. Original location data is GPS data (latitude, longitude) transmitted every five minutes and does not include information to specify individuals, such as gender or age

$$X = \{(u, \tau, lat, lon)\}, \tag{1}$$

where u, τ, lat, and lon are the device ID, timestamp, latitude, and longitude, respectively.

For social media posts, geo-tagged tweets are used for this study. The tweet dataset is structured also as a table containing the user id, latitude, longitude, timestamp, and raw tweet. Supplemental components of original tweets such as URLs, hashtags, usernames, and location names in "check-in" tweets are removed. For further processing, we use the following equations to define tweets:

$$t_i = \{(u_i, \tau_i, lat_i, lon_i, \mathbf{w_i})\} \in T, \tag{2}$$

and the text in the tweets.

$$\mathbf{w_i} = \{w_{i,1}, w_{i,2}, \ldots, w_{i,N_i}\} \in V. \tag{3}$$

A tweet t_i is a 5-tuple where u_i, τ, lat, and lon are device ID, timestamp, latitude, and longitude by a user with corresponding timestamps, and $\mathbf{w}$ is a bag-of-words which contains N_i words. Let the vocabulary $V = \{1, 2, \ldots, V\}$ be a set of word IDs so that each word appears in the collection of words V at least once.

Our goal is to train a model $Y \approx f(X, B)$. Given X as a combined feature with the number of tweets and the result of topic modeling, our model can infer and estimate the number of GPS points Y.

5 Models

5.1 Overview of models

The entire process consists of three parts (Fig. 3): temporal distribution estimation, topic modeling, and regression. First, the grid-based temporal distributions of GPS points and tweets are calculated. Second, using the tweets, topic modeling is conducted to discover latent topics. Last, by utilizing the temporal distribution of tweets and the results of topic modeling, regression models are trained to predict the temporal distribution of GPS points. As tweets, and potential GPS trajectories, are meant to be collected in real-time, we prioritize model simplicity and processing time for possible applications.

5.2 Temporal distribution estimation from GPS data and tweets

Each dataset is discretized to produce a human mobility tensor and social activity tensor. We selected one hour as the time interval Δd_t. In addition, we defined a grid with Δd_{lat} and Δd_{lon}. We experimentally selected Δd_{lat} and Δd_{lon}= 1km, calculated by the Vincenty distance as the grid size to ensure uniformity throughout Japan.

The human mobility tensor $\mathbf{M} = \{m_{i,j,t}\}$ is a three-dimensional array containing the number of GPS data points

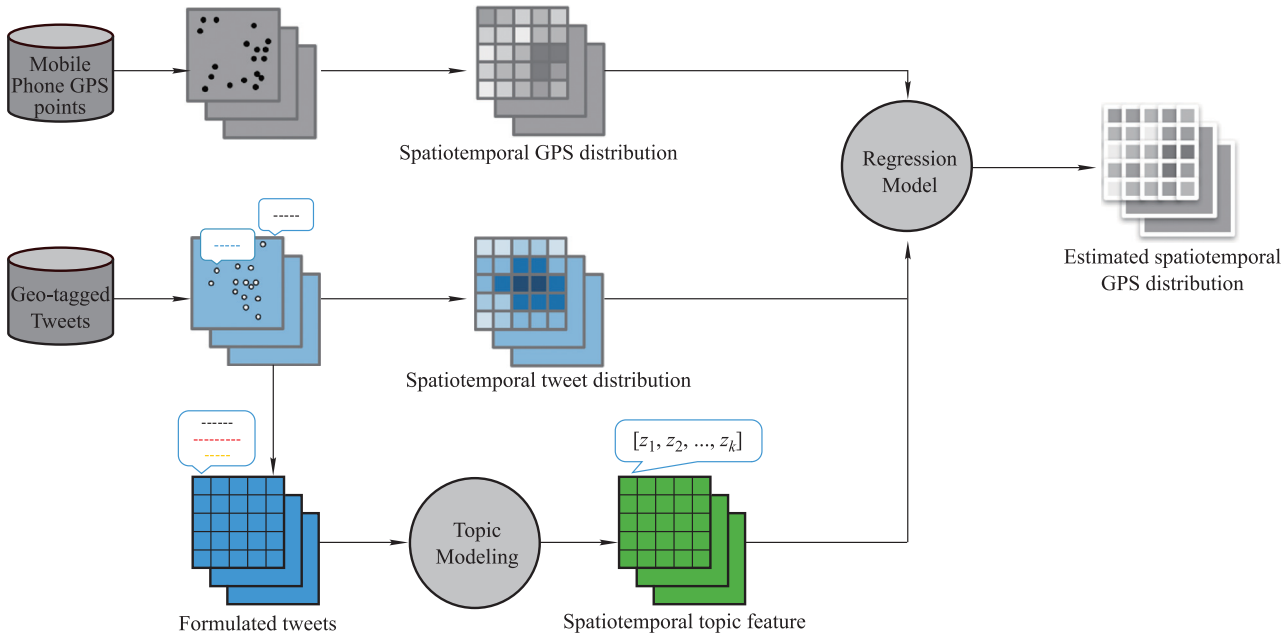


Fig. 3 Model overview

m in the ith latitude index, jth longitude index, and tth temporal index. We have removed duplicate points by the same users from sets of spatiotemporal indices. Similarly, the tensor $\mathbf{S} = \{s_{i,j,t}\}$ is a three-dimensional array containing the number of geo-tagged tweets s in the ith latitude index, jth longitude index, and tth temporal index. For subsequent topic modeling, the tensor $\mathbf{W} = \{\boldsymbol{w}_{i,j,t}\}$ containing the bag-of-words representation of tweets under the same indices is also defined. For example, if $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ for tweets t_1 and t_2 fall under the same indices, these bag-of-words representations are concatenated. The vocabulary $\boldsymbol{V}$ is defined for the entire dataset and is shared among the tensors for the following models.

5.3 Topic modeling from twitter data

Three methods are applied on the geo-tagged tweets. First, we use Latent Semantic Analysis (LSA) based on an "online incremental streamed distributed training algorithm" [37]. LSA takes the term frequency-inverse term frequency matrix as the input and computes a low-rank approximation of the input matrix using singular value decomposition:

$$X_{t\times d} = U_{t\times k}\Sigma_{k\times k}{V_{d\times k}}^{\mathrm{T}}, \tag{4}$$

where X is the term-document frequency matrix, U and V are orthogonal matrices and Σ is a diagonal matrix. t is the number of terms, d is the number of documents, and k is the dimension size (the number of topics). The ith column in X represents a vector corresponding to the ith document in relation to each term, while the ith column in $V(d_i)$ becomes the vector corresponding to the ith document in the low directional space where the number of topics = k. $\Sigma_k d_i$ for each document will be saved and used with the regression models.

Second, LDA is performed. LDA is a probabilistic extension of LSA [38]. LDA assumes that a set of documents are derived from k topics through a generative process where each topic has a multinomial distribution $\beta_k \sim \mathrm{Dirichlet}(\eta)$ over the vocabulary. For each document d, the distribution over topics $\theta_d \sim \mathrm{Dirichlet}(\alpha)$ is drawn followed by topic index $z_{di} \in \{1, 2, \ldots, K\}$ and topic weights $z_{di} \sim \theta_d$, and finally word w_{di} is drawn from the selected topic $w_{di} \sim \beta_{zdi}$. In this study, a variation with faster online implementation [39] is used. Each probability corresponding to the topics for each document will be saved and used with the regression models.

Finally, the topic tensor $\mathbf{T} = \{z_{k,i,j,t}\}$ is defined using the result of the topic models. It contains topic weights z of topic k on a collection of tweets falling under the ith latitude index, jth longitude index, and tth temporal index. We experimentally set the number of topics k as 10.

5.4 Regression models for data fusion

To understand the relationship between GPS trajectory and spatiotemporal topics from topic modeling, we develop the regression model, $Y \approx f(X, B)$. Under three scenarios (# of tweets → GPS, # of tweets + LSA → GPS, and # of tweets + LDA → GPS), Y is the mobility tensor $\mathbf{M}$ reshaped into a column vector. For the first scenario (# of tweets → GPS), X is tensor $\mathbf{S}$ reshaped into a column vector. For the second and third scenarios (# of tweets + LSA → GPS and # of tweets + LDA → GPS), X is tensor $\mathbf{S}$ or $\mathbf{T}$, reshaped and stacked to construct a matrix with 1+k rows with $\mathbf{T}$ from LSA or LDA, respectively.

In this study, the Support Vector Regression (SVR) model is employed, and we also use the Radial Basis Function (RBF) and Gradient Boosting (GB). The RBF kernel function is defined as follows:

$$\exp(-\gamma|x - x'|^2), \tag{5}$$

where γ represents kernel parameters in the form of a kernel coefficient.

Gradient Boostinghas high predictive power with heterogeneous features and takes the form of an ensemble of decision tree models. The loss function is optimized using least squares regression, least absolute deviation, and the combination of the two.The evaluation of the model performance is based on k-fold cross validation with $k = 3$. The parameters for each model are optimized based on grid searches.

6 Experiments

We selected the area of interest as the Greater Tokyo Area (138.72 to 140.87 in longitude, 34.9 to 36.28 in latitude) and the timespan from July 25, 2013 to July 31, 2013. The timespan is that for which we could prepare both GPS data and Twitter data and in this timeframe, we anticipated a yearly fireworks festival (*Sumidagawa Fireworks Festival*) usually held on the last Saturday in July (the 27th in 2013). However, in 2013, the festival was canceled during the event for the first time in the festival's history due to heavy rain. Spectators of the event are known to gather along Sumidagawa-river, which is one of the largest rivers near the city center of Tokyo.

6.1 Results of topic modeling

As observed in Tables 1 and 2, some topics can be considered strongly related to mobility and transportation infrastructure. While LSA tends to produce topics on generic activities, LDA produces topics potentially about specific events.

LDA also produces some topics mentioning certain locations (e.g., Tokyo and Shinjuku stations, as both are popular transportation hubs in the area).When comparing the shares of probability for each topic in the LDA model, temporal fluctuation can be observed (Fig. 4). Every morning, the shares of certain topics (e.g., topic 3 “Food” and 4 “Travel”) drop and shares of other topics (e.g., topic 7 “Routine”) increase. Geographically, topic 4 “Travel” is significant around transportation hubs and tourist spots. Topic 7 “Travel, Events” is clustered around the fireworks festival venue during the fireworks festival.This suggests that the model successfully captures the dominant visiting purpose aroundan area.

Table 1 Selected topics and words with high association (LSA)

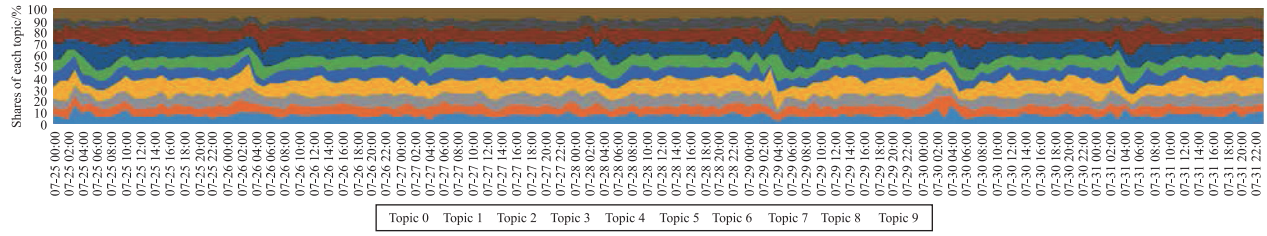
Topics	Words
1. Daily life	+投稿(post) +写真(picture) +駅(station) +動画(video) +店(store) +来(come) +今日(today) +乗り換え(transfer) +ランチ(lunch)
2. “morning, commute”	+おはよう(good morning) +今日(today) +駅(station) +朝(morning) +ありがとう(thank you) +雨(rain) +来(come) +日(day) +行っ(go)
8. “evening, commute”	+帰宅(go home) −来(come) −久しぶり(long time) +いま(now) +ただいま(came home) +休憩(rest) −通過(pass) +する(do)

Table 2 Selected topics and words with high association (LDA)

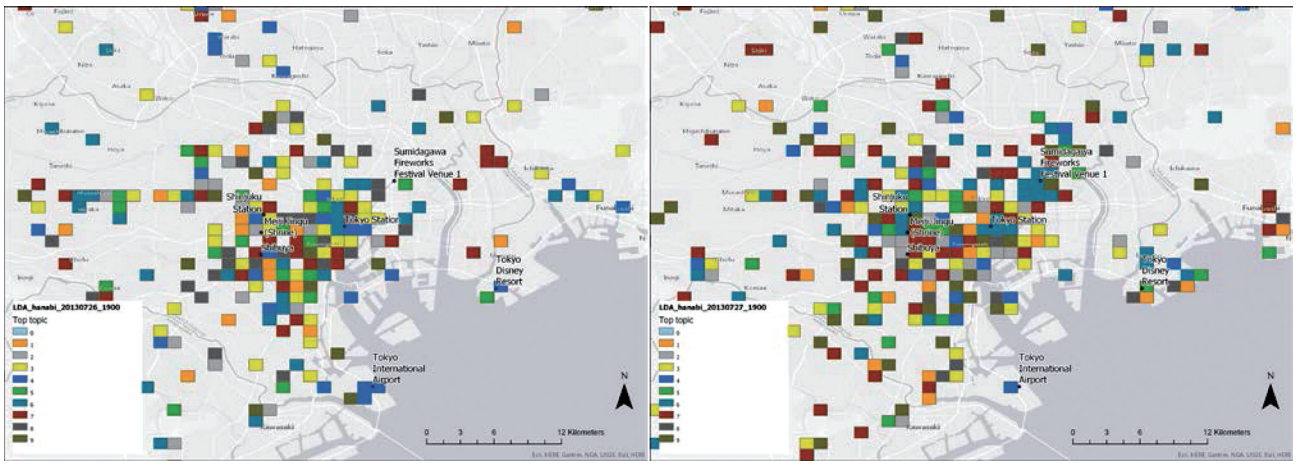
Topics	Words
3. Food	店(store) 麺(noodle) ラーメン(noodle) 食べ(eat) 今日(today) ビール(beer) 人(people) そば(noodle) 円(circle or yen)
4. Travel	Tokyo in 東京(Tokyo) ocean 来(come) 前(past) 行っ(go) 休憩(a break)
6. Travel, Events	駅(station) 店(store) 線(line) 新宿(Shinjuku) here 東京(Tokyo) 花火(fireworks) JR(Japan Railways) 今日(today)
7. Routine	ござい(greeting) おはよう(Good Morning) 会(meet) 今日(today) 初(first time) 日(day) 飲み(drink)

6.2 Regression analysis for data fusion

For model performance evaluation, the Root Mean Square Error (RMSE: in number of GPS points) is calculated and used for performance comparison (Table 3). Overall, the combined feature with the number of tweets and LSA with Gradient Boosting achieved the best performance. Figure 5 shows the average distribution of GPS points, tweets, and estimated GPS points among all temporal units in the time span. Generally, when the average number of tweets in the spatiotemporal

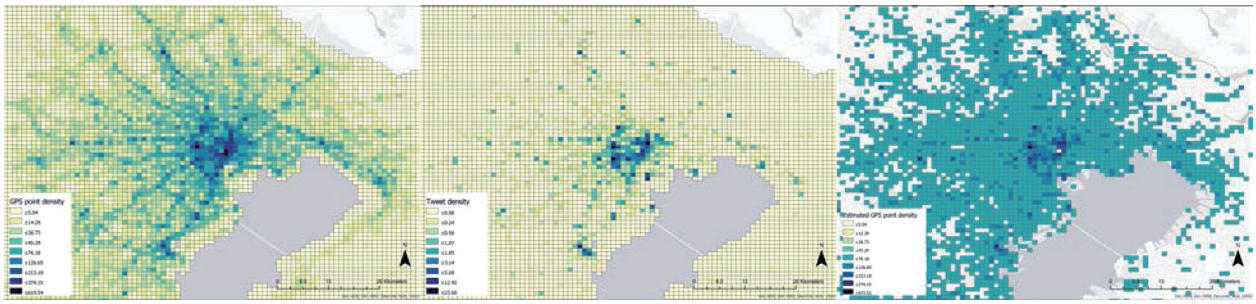


(a)



(b) (c)

Fig. 4 (a) Average shares of each topic (LDA) in each temporal unit; (b) top topic from LDA model during 07/26/2013 (Friday) 19:00 – 20:00; (c) top topic from LDA model during 07/27/2013 (Saturday, Fireworks festival) 19:00 – 20:00 (On average, Topics 3 and 7 take the greatest shares in each temporal unit. Topic 4 (Travel) is dominant in transportation hubs and tourist spots (Tokyo station, Airport, Tokyo Disney Resort) on Friday. On Saturday, Topic 6 (Travel, Event) is significant around the fireworks festival venue)



(a) (b) (c)

Fig. 5 Distribution of (a) GPS points, (b) tweets, and (c) estimated distribution of GPS points (Credit: "Konzatsu-Tokei®" ©ZENRIN DataCom CO., LTD)

unit is below three, the model overestimates the number of GPS points. This is reasonable, but is one of the most significant limitations of the model.

Table 3 Regression model performance (RMSE)

	SVR	GB
# of tweets	102.95	99.21
# of tweets + LSA	99.55	93.64
# of tweets + LDA	100.89	96.93

Figure 6 shows the spatial distribution of RMSE values (top left) and the estimation error (ground truth – estimated value) for 19:00 – 20:00 each day of the week. Overall, the accuracy around tourist spots (Tokyo Disney Resort, Meiji Jingu, and an international airport) is higher than average; however, in the areas around transportation hubs (Tokyo and Shinjuku stations), the accuracy is lower than average. This may be because of intensive daily transportation mixed with commute and leisure.The model tends to underestimate on weekdays around major transportation hubs (Tokyo and Shinjuku stations) and the accuracy is higher during weekends.

Figure 7 shows the RMSE in each temporal unit and RMSE compared to the number of tweets in the spatiotemporal unit. The accuracy is generally higher during the weekend (07/27 and 07/28). However, the accuracy is lower during rush hour on weekdays (06:00 – 10:00 and 17:00 – 20:00). Compared to the number of tweets, as the number of tweets increases, the accuracy generally increases, even though there are clear fluctuations in accuracy in some places with larger numbers of tweets.

These two analyses suggest that the intensive transportation due to the commute during rush hour around transportation hubs makes the estimation difficult. We think this is one of the major limitations of the model.

6.3 Parameter evaluation

To evaluate the parameter selection, we conducted the experiment using different sets of parameters: time interval Δd_t, spatial grid intervals Δd_{lat} and Δd_{lon}, the number of topics k, and the time span for the experiment. Table 4 summarizes the results.

First, increasing the number of topics does not significantly improve the model performance. Even though a higher number of topics theoretically helps the model to distinguish more particular topics, this city-scale human mobility estimation application appears to be dominated by generic daily mobility patterns.

As time interval Δd_t or spatial grid intervals Δd_{lat} and Δd_{lon} become smaller (higher resolution), R^2 decreases, but the RMSE also decreases.With higher resolution, the error tends to become smaller(better RMSE), but noise hurts the model's fitting ability (worse R^2). This parameter selection involves a tradeoff between accuracy (RMSE) and performance (R^2), and computation cost. In addition to the basic limitation mentioned in Section 6.2, the recommendation is to use higher resolution by securing the average number of tweets in the spatiotemporal unit as at least three.

Table 4 Model performance evaluation with different parameter sets

	RMSE			R^2		
# of topics	10	20	30	10	20	30
# of tweets + LSA (GB)	93.64	92.96	92.94	0.41	0.41	0.41
# of tweets + LDA (GB)	96.93	95.06	95.66	0.36	0.38	0.38
Time interval /mins	30	60	90	30	60	90
# of tweets + LSA (GB)	66.45	93.64	116.04	0.33	0.41	0.44
# of tweets + LDA (GB)	67.41	96.93	119.16	0.31	0.36	0.4
Spatial grid interval /m	500	1000	1500	500	1000	1500
# of tweets + LSA (GB)	57.55	93.64	126.32	0.32	0.41	0.45
# of tweets + LDA (GB)	59.15	96.93	128.55	0.28	0.36	0.43
Time span /weeks	1	4	8	1	4	8
# of tweets + LSA (GB)	93.64	97.23	96.96	0.41	0.39	0.41
# of tweets + LDA (GB)	96.93	98.98	98.9	0.36	0.37	0.38

7 Conclusion

In this study, we combined big and heterogeneous GPS data

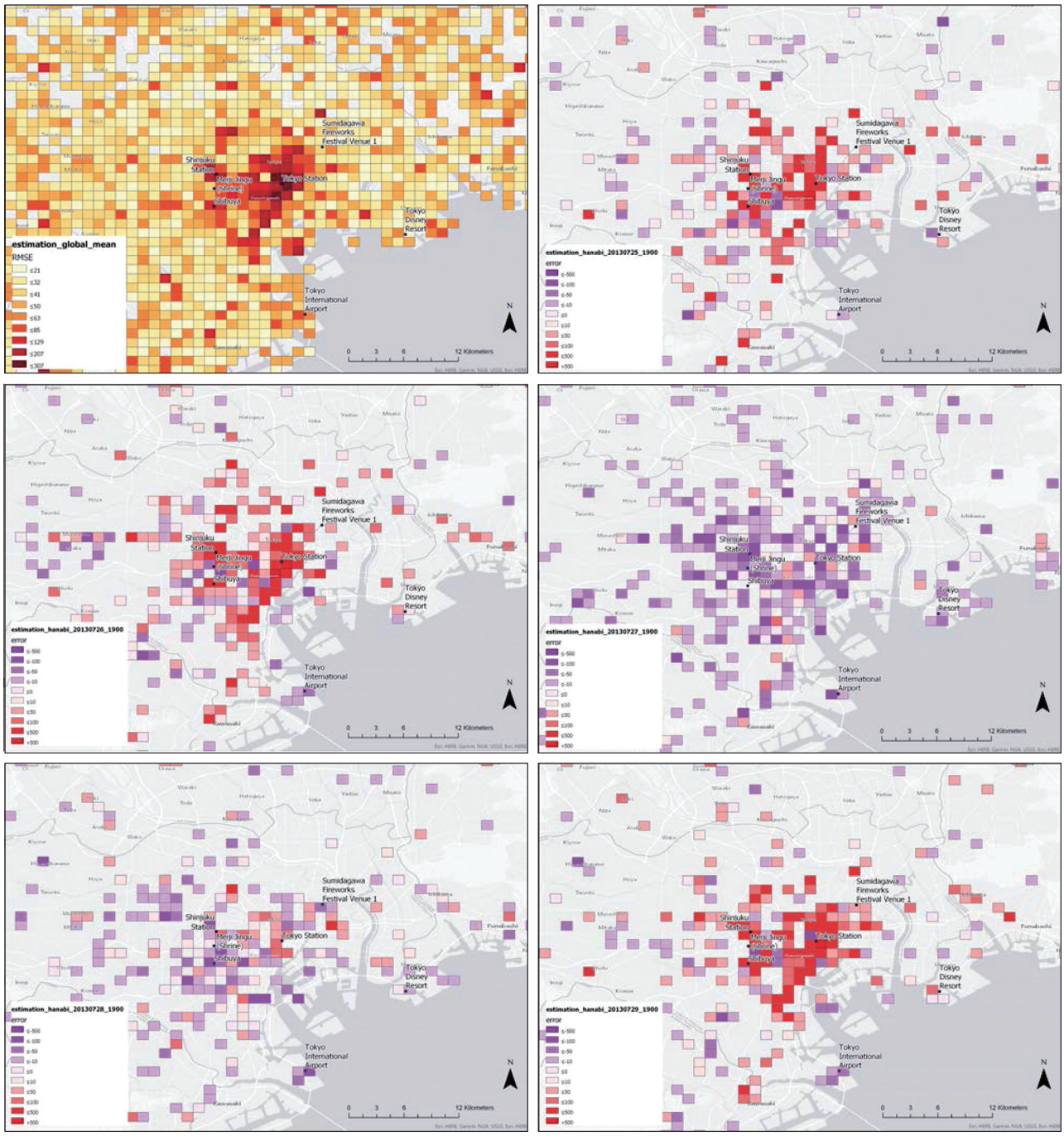


Fig. 6 Spatial distribution of RMSE (top left) and estimation error (ground truth – estimated value) in 19:00 – 20:00 on each day from 07/25/2013 (Thursday) to 07/29/2013 (Monday) (Credit: "Konzatsu-Tokei®" ©ZENRINDataCom CO., LTD)

with geo-tagged tweets. The application of topic modeling discovers several behaviors in urban life style including seasonal events. The regression model with GPS, tweets, and topics from tweets produces better results to infer human mobility, especially during unprecedented events. The result suggests that spatiotemporal topics from geo-tagged tweets are influential to evaluate human mobility. The analysis of the regression result suggests some key indicators for parameter choice. The model performs best on weekends and sometimes on weekdays, except during rush hour around transportation hubs. The model also works during large-scale events, where a sufficient number of tweets is expected, depending on the parameter choices.

As future work, we aim at cross-domain data fusion [40,41] using deep neural network models to explore the multi-feature representation of human mobility and social

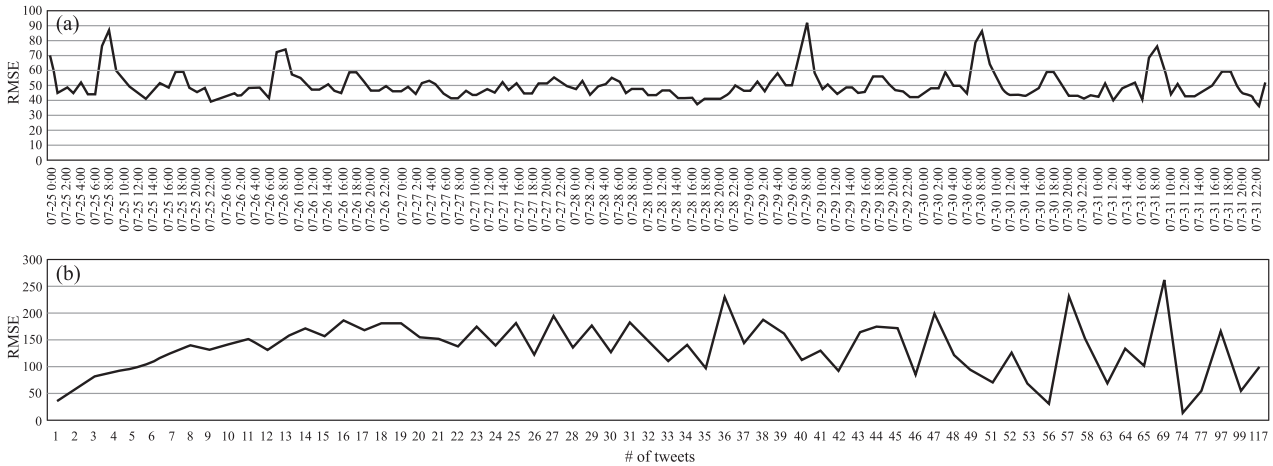


Fig. 7 (a) RMSE in each temporal unit and (b) RMSE compared to # of tweets in spatiotemporal unit

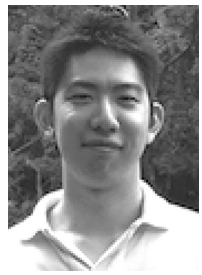
activities.

Acknowledgements This work was partially supported by JST, Strategic International Collaborative Research Program (SICORP); Grant in-Aid for Scientific Research B (17H01784) and Grant in-Aid for Young Scientists (26730113) of Japan's Ministry of Education, Culture, Sports, Science, and Technology (MEXT). We specially thank ZENRIN DataCom CO., LTD for the provision of GPS data and their support, and Nightley Inc. for geo-tagged tweets.

References

1. Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3): 38
2. Zhang D, Wang Z, Guo B, Yu Z. Social and community intelligence: technologies and trends. IEEE Software, 2012, 29(4): 88–92
3. Xiong Z, Zheng Y, Li C. Data vitalization's perspective towards smart city: a reference model for data service oriented architecture. In: Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. 2014, 865–874
4. Calabrese F, Diao M, Lorenzo G D, Ferreira J, Ratti C. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transportation Research Part C: Emerging Technologies, 2013, 26: 301–313
5. Kang C, Ma X, Tong D, Liu Y. Intra-urban human mobility patterns: an urban morphology perspective. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1702–1717
6. Song X, Zhang Q, Sekimoto Y, Shibasaki R. Prediction of human emergency behavior and their mobility following large-scale disaster. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014, 5–14
7. Zhai Z, Liu B, Wang J, Xu H, Jia P. Product feature grouping for opinion mining. IEEE Intelligent Systems, 2012, 27(4): 37–44
8. Kim Y, Han J, Yuan C. TOPTRAC: topical trajectory pattern mining. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015, 587–596
9. Cheng T, Wicks T. Event detection using Twitter: a spatio-temporal approach. PLoS One, 2014, 9(6): e97807
10. Grinberger Y, Shoval N. A temporal-contextual analysis of urban dynamics using location-based data. International Journal of Geographical Information Science, 2015, 29(11): 1969–1987.
11. Spaccapietra S, Parent C, Damiani M L, De Macedo J A, Porto F,Vangenot C. A conceptual view on trajectories. Data & knowledge engineering, 2008, 65(1): 126–146.
12. Sekimoto Y, Shibasaki R, Kanasugi H, Usui T, Shimazaki Y. PFlow: reconstructing people flow recycling large-scale social survey data. IEEE Pervasive Computing, 2011, 10(4), 27–35
13. Wang J, Gu Q, Wu J, Liu G, Xiong Z. Traffic speed prediction and congestion source exploration: a deep learning method. In: Proceedings of the 16th IEEE International Conference on Data Mining. 2016, 499–508
14. Wang J, Gao F, Cui P, Li C, Xiong Z. Discovering urban spatio-temporal structure from time-evolving traffic networks. In: Proceedings of the 16th Asia-Pacific Web Conference. 2014, 93–104
15. Dong W, Wang Y, Yu H. An identification model of urban critical links with macroscopic fundamental diagram theory. Frontiers of Computer Science, 2017, 11(1): 27–37
16. Chen L, Ma X, Pan G, Jakubowicz J. Understanding bike trip patterns leveraging bike sharing system open data. Frontiers of Computer Science, 2017, 11(1): 38–48
17. Wang J, Wang Y, Zhang D, Wang L, Chen C, Lee J W, He Y. Real-time and generic queue time estimation based on mobile crowdsensing. Frontiers of Computer Science, 2017, 11(1): 49–60
18. Chen C, Chen X, Wang Z, Wang Y, Zhang D. Scenicplanner: planning scenic travel routes leveraging heterogeneous user-generated digital footprints. Frontiers of Computer Science, 2017, 11(1): 61–74
19. Song X, Zhang Q, Sekimoto Y, Horanont T, Ueyama S, Shibasaki R. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013, 1231–1239
20. Wang J, Chen C, Wu J, Xiong Z. No longer sleeping with a bomb: A duet system for protecting urban safety from dangerous goods. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017, 1673–1681

21. Wang J, Lin Y, Wu J, Wang Z, Xiong Z. Coupling implicit and explicit knowledge for customer volume prediction. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017, 1569–1575
22. Fan Z, Song X, Shibasaki R. CitySpectrum: anon-negative tensor factorization approach. In: Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014, 213–223
23. Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012, 186–194
24. Guo B, Wang Z, Yu Z, Wang Y, Yen N Y, Huang R, Zhou, X. Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm. ACM Computing Surveys, 2015, 48(1): 7
25. Morstatter F, Pfeffer J, Liu H, Carley K M. Is the Sample Good Enough? Comparing data from Twitter's streaming API with Twitter's firehose. In: Proceedings of ICWSM. 2013, 400–408
26. Steiger E, De Albuquerque J P, Zipf A. An advanced systematic literature review on spatiotemporal analyses of Twitter data. Transactions in GIS, 2015, 19(6): 809–834
27. Cheng T, Wicks T. Event detection using Twitter: a spatio-temporal approach. PLoS One, 2014 9(6), e97807
28. Sakaki T, Okazaki M, Matsuo Y. Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 919–931
29. Abbasi A, Rashidi T H, Maghrebi M, Waller S T. Utilising location based social media in travel survey methods: bringing Twitter data into the play. In: Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. 2015, 1–9
30. Ao J, Zhang P, Cao Y. Estimating the locations of emergency events from Twitter streams. Procedia Computer Science, 2014, 31: 731–739
31. Cameron M A, Power R, Robinson B, Yin J. Emergency situation awareness from twitter for crisis management. In: Proceedings of the 21st International Conference on World Wide Web. 2012, 695–698
32. Frias-Martinez V, Frias-Martinez E. Spectral clustering for sensing urban land use using Twitter activity. Engineering Applications of Artificial Intelligence, 2014, 35: 237–245
33. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D. Understanding human mobility from Twitter. PLoS One, 2015, 10(7): e0131469
34. Blanford J I, Huang Z, Savelyev A, MacEachren A M. Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. PLoS One, 2015, 10(6): e0129202
35. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 2014, 41(3): 260–271
36. Pan B, Zheng Y, Wilkie D, Shahabi C. Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2013, 334–343
37. Řehůřek R. Subspace tracking for latent semantic analysis. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. 2011, 289–300
38. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of machine Learning research, 2003, 3(4-5), 993–1022.
39. Hoffman M D, Blei D M, Bach F. Online learning for latent dirichlet allocation. In: Proceedings of the Neural Information Processing Systems Conference. 2010, 856–864
40. Zheng Y. Methodologies for cross-domain data fusion: an overview. IEEE Transactions on Big Data, 2015, 1(1): 16–34.
41. Wang J, He X, Wang Z, Wu J W, Yuan N J, Xie X, Xiong Z. CD-CNN: a partially supervised cross-domain deep learning model for urban resident recognition. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018



Satoshi Miyazawa is a PhD student of the Department of Socio-Cultural Environmental Studies at The University of Tokyo, Japan. His research interests include human mobility, LBSN, data mining, and machine learning.

Xuan Song received the BS degree in information engineering from the Jilin University, China in 2005 and PhD degree in signal and information processing from Peking University, China in 2010. From 2010 to 2012, he joined the Center for Spatial Information Science, The University of Tokyo, Japan as a postdoctoral researcher. In 2012 and 2015, he was promoted to project assistant professor and project associate professor at the same university. His research areas are mainly in artificial intelligence and data mining.

Tianqi Xia is a master student of the Department of Socio-Cultural Environmental Studies, The University of Tokyo, Japan. He received his BS degree in geographic information science from Wuhan University, China. His research interests include spatial data mining, data analysis and intelligent transportation systems.

Ryosuke Shibasaki is a professor at the Center for Spatial Information Science, The University of Tokyo, Japan. His research interests include satellite and airborne remote sensing, track- ing technologies, geospatial information gathering and integration among heterogeneous systems, and common service platforms for geospatial information.

Hodaka Kaneda is an employee of ZENRIN-Datacom CO., LTD, Japan. His work is to deal with GPS data and to supply “Konzatsu-Tokei (R)" Data.