ERSA2019@Lyon, Aug. 29 to 31

Application of geographical topic model to metropolitan area in Japan

Makoto TSUKAI (Hiroshima University)
mtukai@hiroshima-u.ac.jp
Satoko OHONO (Kochi Prefecture)
sa06.ohstk@gmail.com
Yuta TSUKANO (Kansai Airport Company)
m171467@hiroshima-u.ac.jp

Abstract: Growing concerns to land use planning in fine zone level, the necessity to understand the structure of urbanized area is increased. This study purposes to clarify the applicability of topic model to geographical characteristics, named geographic topic model. Using the proposed model, a couple of metropolitan area in Kyusyuu area is analyzed to compare the geographical characteristics of both cities and its change between two cross sections. As a previous work of geographic topic model, Tsukai and Tsukano (2016) tried to apply the topic model for 5 metropolitan areas and successfully extracted the geographical topics from the dataset, but the following two issues were remained as unsolved tasks. One is to establish the procedure to fix the number of geographical topics, and the other is to utilize topic load matrix, effectively. This study purposes to overcome the above problems, and to confirm the applicability and validity of the proposed method. The information obtained from factor load matrix is aggregated with several different ways such as dominant plot mapping, weighted aggregation of topics with other attributes and the composite topic aggregation along with the accessibility to public transportation. he The empirical analysis in Kyusyu and Fukuoka city showed the usefulness of geographical topic model to evaluate the land use characteristics.

Key words: TOD, policy assessment, comparative analysis

1. Introduction

In Japan, growth of population in CBD area accompanied with the expansion of suburban areas has been observed for past 50 years. The expanded suburban area has often lower density, suffering from poor road access and lack in public transport access. Considering the decrease in total population, further expansion of lower density area should be stopped. In order to make proper land use incentives and regulations, a novel land use scheme called “optimizing built environment”. This scheme aims to

make compact city, along with the public transport network. In order to achieve effective policy making, the necessity to understand the structure of metropolitan area is increased.

There are several approaches to observe the geographical characteristics of metropolitan area. Conventionally, the images observed from satellites has been used for land covering monitoring. Grigorașa and Urițescu (2019) estimated the land covering class to find the relationship with land surface temperature. The empirical study at Bucharast in Romania showed the significant decrease in vegetation areas, and a negative correlation between vegetation and land surface temperature, resulting in heat island phenomena. Renne *et al.* (2016) estimated a model to regress network accessibility and built environment on transit commuting share of the cities in the Unites States. The estimated model clarified the effectiveness of public transport access. Guan *et al.* (2019) analyzed the area development around public transport station. Xu and Yang (2019) clarified the correlation between public transport access including transferring cost and land use characteristics, by estimating geographically weighted regression model (GWR). Dadashpoor *et al.* (2019) classified the land scape patterns to find the longitudinal change of the relationship between land space and land use, using GWR. In this study, the drastic changes in suburban area was clarified. Zeng, Yang and Dong (2017) integrated geographical big data such as Point-Of-Interst or Open –Street-Map with SNS big data such as "Weibo" to evaluate land use efficiency based on GWR. Applying their model to 40 megalopolises in China, the potential and problems of the target cities are quantitatively clarified. Flores *et. al* (2019) proposed a novel approach to classify land covering into surface characteristics groups using the pictures with very fine resolutions. This study proposed an efficient algorithm to make the dictionaries for the classification with convolutional neural network was compared with other deep learning methods. Mohamed and Worku (2019) tried to integrate land covering with "land use" in order to clarify the growth and sprawls around the metropolitan area. Hereafter, land use indicates the land attributes such as demographic, social or economic activity data, other than land covering mainly observed by satellite image or aerial photograph on land surface. Their study in Addis Ababa in Ethiopia explored to get the land covering classification and the land use map was overlaid with land use map to find the (in)correspondence of land development with the master plan of the city. Zhang *et al.* (2019) proposed an integrated deep learning approach enabling the simultaneous estimation of land covering with land use. In this study, convolutional neural network with multilayer perceptron was applied.

Owing to geographical information system with accessible geographic database, several aspects of land characteristics for each municipality are available at fine spatial scale. In Japan, "National Land Numerical Information" facilitated by ministry of land, infrastructure and transportation provides the geographic/demographic/economic characteristics of whole national area covering with 1 km squared mesh scale. Such the detailed information is useful for fine scale analysis on land use structure, however, the data handling with many attributes and with large samples is not so easy. Understanding the structure of metropolitan area with its dynamics requires a comprehensive

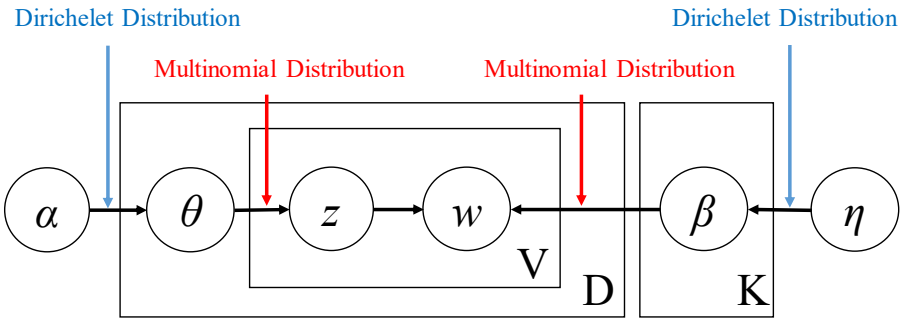

Fig.1 Graphical Model of Topic Model

analysis not only in built environment, but also demographic, industrial or commercial characteristics.

This study purposes to clarify the applicability of topic model to geographical characteristics, named geographic topic model. Using the proposed model, a couple of metropolitan area in Kyusyuu area is analyzed to compare the geographical characteristics of both cities and its change between two cross sections. Tsukai and Tsukano (2016) tried to apply the topic model for 5 metropolitan areas and successfully extracted the geographical topics from the dataset, but the following two issues were remained as unsolved tasks. One is to establish the procedure to fix the number of geographical topics, and the other is to utilize topic load matrix, effectively. In this study, these problems are tackled with proposing feasible data processing procedure.

Topic model is used for a topic estimation problem from the documents with huge number of words. The calculation procedure used for such purpose, principal component analysis or factor analysis are well-known algorithm to make reduction of dataset. While the principal component or factor analysis is based on eigen value / eigen vector decomposition, the topic model is based on singular value decomposition with stochastic and hierarchical mathematical structure. The advantage of topic model is positive estimation of parameters, while negative estimates often appears in conventional approaches. Positive estimates of parameters make much easier interpretation for topic, and the mesh-topic load matrix is conveniently used for further analysis to find the relationship with other attributes.

2. Topic model

Topic estimation problem is to find several “significant” or “interpretable” topics from large amount of document (composed of huge words), in other words, to find algorithms to reduce the information of original data. Although several variation of topic model are developed, the basic idea of topic model is shown in Latent Dirichlet Allocation (LDA; Blei *et al*., 2003). LDA assumes that topic is not directly observed but existing as latent variable. Each document has several different topics, not occupied by single topic. LDA has two level of stochastic structure. The upper level is hyper parameter level for document to latent topic index parameters in each document, and for latent topic

to vocabulary parameters in each topic. The hyper parameters are assumed to follow Dirichlet distribution. The lower level is latent topic index for each word, and for observed word to latent topic index and topic-wise vocabulary parameters. The latent topic index and observed words are assumed to follow multiple distribution. By taking the hierarchical structure with stochastic assumption, the huge number of parameters in lower level are loosely restricted by upper level structure with hyper parameters. In topic model, hyper parameters are estimable. In order to keep the consistency with conventional studies in topic model, following specification of topic model is explained by using the terms in document analysis.

Suppose a set of document as D. Each document : d is composed of several words : $N^{(d)}$, and a set of words including all the words appearing in d at n-th order is $\{w^{(n,d)}\}_{n=1}^{N^{(d)}}$. Each of word $w^{(n,d)}$ is indicated with 1-of-V expression as $w_v^{(n,d)} \in \{e_v\}_{v=1}^{V}$. 1-of-V expression gives an unique number for all vocabularies in D from 1 to V , then n-th word in a vocabulary v is indicated the vector with 1 for column v and with 0s for all the other v(s). $w^{(n,d)}$ is an element of the matrix W with N rows and V columns. Here, N is the total number of words, by summed up $N^{(d)}$ for all the documents.

LDA assumes that each vocabulary has an unobserved potential topic for each word in document d : $z^{(n,d)} \in \{e_k\}_{k=1}^{K}$, and that each document d has unique topic distribution : $\widetilde{\theta}_d$, and each topic k has different vocabulary distribution: β_k. Focusing on a set of vocabulary in d without considering the order of word, probability to observe a word and potential topic $w^{(n,d)}$ and $z^{(n,d)}$ can be formulated in multinomial distribution (1) and (2), respectively.

$$p(z^{(n,d)}|\widetilde{\theta}_d) = Multi_K(z^{(n,d)};\widetilde{\theta}_d) \tag{1}$$

$$p(w_v^{(n,d)}|z^{(n,d)},\beta_1,...,\beta_K) = \prod_{k=1}^{K}\{Multi_V(w^{(n,d)};\beta_k)\}^{z_k^{(n,d)}} \tag{2}$$

In order to estimate $\widetilde{\theta}_d$ and β_k in multinomial distribution, Dirichlet distribution is assumed as conjugate prior distribution of multinomial distribution in (3) and (4), respectively.

$$p(\widetilde{\theta}_d|\alpha) = Dir_K(\widetilde{\theta}_d;\alpha) \tag{3}$$

$$p(\beta_k|\eta) = Dir_V(\beta_k;\eta) \tag{4}$$

Where, α and η are the hyper parameters in each Dirichlet distribution.

Topic distribution for each document is indicated in $\Theta = (\widetilde{\theta}_1,...,\widetilde{\theta}_D)^t$ with D rows and K columns, and vocabulary distribution for each topic is indicated in $B = (\beta_1,...,\beta_K)^t$ with K rows and V columns, where superscript t indicates transpose. The simultaneous distribution of observed data : $W = [\{w^{(n,d)}\}_{n=1}^{N^{(d)}}]_{d=1}^{D}$ and unobserved potential topic : $Z = [\{z^{(n,d)}\}_{n=1}^{N^{(d)}}]_{d=1}^{D}$ is formulated in (5).

$$
\begin{aligned}
&p(W, Z|\Theta, B) \\
&= \prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}} p(w^{(n,d)}|z^{(n,d)}, \beta_k)\, p(z^{(n,d)}|\widetilde{\theta}_d) \\
&= \prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}}\prod_{k=1}^{K}(\Theta_{d,k}\prod_{v=1}^{V} B_{k,v}^{w_v^{(n,d)}})^{z_k^{(n,d)}}
\end{aligned}
\qquad (5)
$$

In order to simplify the model structure, marginal probability of observed data W with N rows and V columns is calculated by summed up of potential topic $z^{(n,d)}$, shown in (6). Further, W in 1-of-V expression for each word is replaced by M in bag of words expression for each document in (7).

$$
\begin{aligned}
&p(W|\Theta, B) \\
&= \sum_{Z} p(W, Z|\Theta, B) \\
&= \prod_{d=1}^{D}\prod_{n=1}^{N^{(d)}}\Big(\sum_{z^{(n,d)}\in\{e_k\}_{k=1}^{K}}\prod_{k=1}^{K}(\Theta_{d,k}\prod_{v=1}^{V} B_{k,v}^{w_v^{(n,d)}})^{z_k^{(n,d)}}\Big) \\
&= \prod_{d=1}^{D}\prod_{v=1}^{V}\Big((\Theta B)_{d,v}\Big)^{\sum_{n=1}^{N^{(d)}} w_v^{(n,d)}}
\end{aligned}
\qquad (6)
$$

$$
M = (m_1, ..., m_D)^t, \quad M_{d,v} = \sum_{n=1}^{N^{(d)}} w_v^{(n,d)} \qquad (7)
$$

M in (8) has D rows and V columns. If u_d indicating a d-th vector in $U = (u_1, ..., u_D)^t = \Theta B$ shown in (8) is used, (7) is updated with (9).

$$
\begin{aligned}
&p(M|\Theta, B) \\
&= \prod_{d=1}^{D} N^{(d)}! \prod_{v=1}^{V} \frac{\Big((\Theta B)_{d,v}\Big)^{M_{d,v}}}{M_{d,v}!} \\
&= \prod_{d=1}^{D} Multi_{V, N^{(d)}}(m_d; u_d)
\end{aligned}
\qquad (8)
$$

$$
M \approx \Theta B \qquad (9)
$$

(9) is obtained when the number of words in d, $N^{(d)}$, is multiplied for each row in (8), since (8) is a multinomial distribution for the probability of words in each document. As shown in (9), LDA will give the approximated decomposition of matrix M by ΘB with rank K, which is smaller than the number of columns : V in M. ΘB, The parameter in LDA indicates the probability density, so then all the elements in the matrix have positive values.

Since Dirichlet distribution is conjugate prior distribution of multiple distribution, the posterior distribution of lower level also follows multiple distribution. Therefore, the estimation procedure of LDA or topic model can rely on the Bayesian techniques. MCMC, Gibbs sampler and Variational Bayes are proposed as the parameter estimation procedure. In this study, Collapsed Variational Bayes

method to find the fixed point of parameters proposed by Sato and Nakagawa (2015) is used because of its efficiency of calculation. As discussed above, hyper parameters are also estimated by fixed point algorithm with several iterations. In terms of likelihood space characteristics, unfortunately, topic model does not have single summit but multiple (steep) summit. Just reported in the previous studies, most of topic model behaves not so instable. In our study, several trials of initial values in estimation was made.

The model choice is made to refer log-likelihood (LL) at the conversion of calculation. As a non-informative initial LL, the probability for each topic K and vocabulary V are given by uniform probability over all the alternatives, the initial LL can be given as follows.

$$\begin{aligned} L\left(w_{di}|\Theta B_0\right) &= \sum_d \sum_i \log p\left(w_{di}|\Theta B_0\right) \\ &= \sum_d \sum_i \sum_k \log \frac{1}{K} \cdot \frac{1}{V} = -N \log V \end{aligned} \tag{10}$$

As shown in (10), initial LL of topic model depends on N and V, but is irrelevant to K. In this study, LL ratio given by initial LL and the last (converged) LL is used to compare the modes.

In order to check the similarity of estimated topics or topic distribution of a mesh, cosine similarity between vector p and q in (11) is used.

$$S_{pq} = \cos\left(p,q\right) = \frac{p \cdot q}{|p||q|} \tag{11}$$

An appropriate number of topics K* is found by following steps. First, upper limit of K (K_max) is found as the first local peak of K by increasing K as 1,2,3…, because our previous trials reported that the estimated LL over K_max becomes decrease and the similarity of topics over K_max becomes higher (Tsukai and Tsukano, 2018) . Between 1 to K_max, K* is determined as the maximum number of K which do not include similar topics found in (11). In the following analysis, threshold of S_{pq} is tried from 0.7, 0.8 and 0.9, checking the interpretability of topics.

In an application of topic model to geographical data, D and V in the above model (number of documents and number of vocabulary) are replaced with number of area / meshes and number of discretized attributes summed up for all the classes, respectively. The total number of words : N is obtained with the sum of $N^{(d)}$, which is same for all the areas as the number of attributes used in the model. The details of these constant is shown at data description, in Chapter 3.

3. Data set

The input information to topic model is called Bag-of-Words (BOW), which counts the number

of vocabulary for each document. In case of geographical data, attributes are count data such as population, size of area and so on. The difference in data characteristics between ordinal document-vocabulary counts and our mesh-attribute data is the range of counting; the former ranges in lower in positive number, but the latter ranges much wider. In order to convert the mesh-attribute data into BOW style, this study proposes to discretize the attribute distribution by using multiple dummy variables. Such the data processing is called binning. By applying binning to geographical data, distributional characteristics of geographical data mimics ordinal BOW for document data.

At the estimation of topic model, number of topics K has to be given. Therefore, to fix an appropriate number of topics requires iterative estimations of the model. In this study, the following two step procedures are proposed. Firstly, (local) summit of log-likelihood is searched. Then the similarity of topics in terms of topic-vocabulary matrix is calculated by using cosine similarity. Referring to the determined threshold of the similarity, some of similar topic is merged. Topic load matrix is used to check the transition of topics between two cross-sections.

The dataset used in this study is Kyusyuu and Fukuoka metropolitan area, recorded in 2000 and 2010. The population of both cities have more than million, and the sum of urbanized area is about 4000 squared kilometers. 34 attributes are collected form national census, economic census and office and company statistics. The number of meshes in the target area is totally 2113 for each cross section, by dropping unused area. In order to common topics in two cross sections, the datasets in 2000 and 2010 are stacked, so then totally 4226 samples are inputted. Looking on the attribute distribution, most of attributes are skewed at no observation. Therefore, zero is treated as an independent class for all attributes. Each of distribution was binned by natural classification. In this study, 7 classes are set by several trials. Natural classification gives a set of threshold based on the second derivative of distribution (of each attributes). Other characteristics of the dataset is “NA: No Answer”, which is labeled at the mesh with one or a few number of observations, in order to mask the characteristics of a very small number of targets in a mesh. Including zero and NA, totally 9 classes are set for each attribute. Since all the attribute observation in each mesh is classified into one of 9 classes through the binning of attributes, all the mesh has 34 words being equal to the number of attributes. As a result, V is 34 * 9 = 306, D is 4226, $N^{(d)}$ is 34 and N is 4226 * 34 = 143484.

4. Estimation of geographical topics

1) Comparison of topic distribution between Kitakyusyu and Fukuoka

Fig. 2 shows the result of searching in optimum number of topics. The local peak of LL appears at K=36, and the maximum number of topics being different for any couples in the set is obtained at K=16 under 0.8 of the threshold in S_{pq}. The topics obtained with other threshold was tried but the interpretability of them was not good. Table 1 shows the topics we estimated. The title of topics is named by referring to KV matrix: B in (9), which indicate the contribution of each discretized

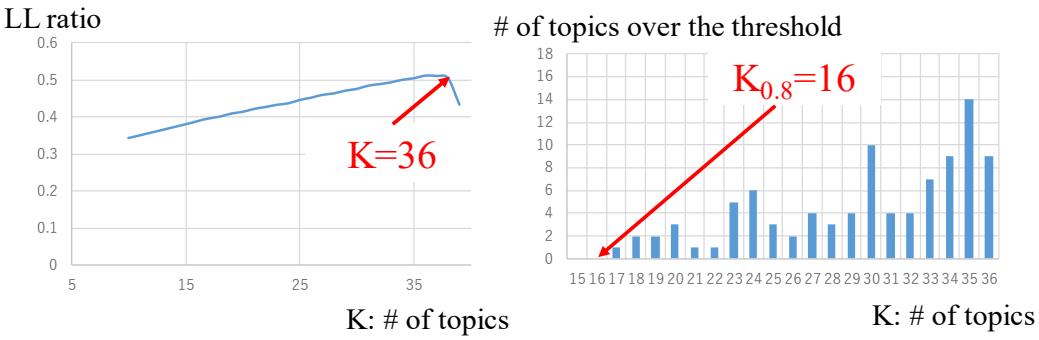

Fig.2 Finding an appropriate number of topics

Table 1 Obtained topics in estimated model

Higher Agglomeration
Middle Agglomeration
Inhabitants : higher density
Inhabitants : middle density
Inhabitants : middle to low density
Low density inhabitants
Low density inhabitants with agriculture
Low density inhabitants and low density commercial
Low density commercial
Low density industory
Low density industory with agriculture
Not used - 1
Not used - 2
Not opend - 1
Not opend - 2
Not opend - 3

attributes. The estimated topics are as follows; two kinds of CBD agglomerations (higher and middle), three kinds of inhabitants (higher, middle and middle to low density) two kinds of low density inhabitants (with no other and with agriculture), two kinds of commercial (with no other and with lower density of inhabitants), two kinds of industries (with no other and with agriculture), two kinds of vacant land and three kinds of no answers. Since the target meshes include wild land (at mountainous side) or water surface along the coastal line, 5 of 16 topics gives no substantial information in land use.

A map plotting for the dominant topic, i.e. the topic with the highest load in the mesh in each cross section, obtained from DK matrix θ in (9) is shown in Fig.3 in 2000 and Fig. 4 in 2010, respectively. Kitakyuusyuu city is located at the Eastern side of target area, and it is an industrial city in coal mining from the beginning of 20th century. Owing to the proximity to coal mine, the first national steel work in Japan was established at Kitakyusyu in 1901. Even after the closure of coal

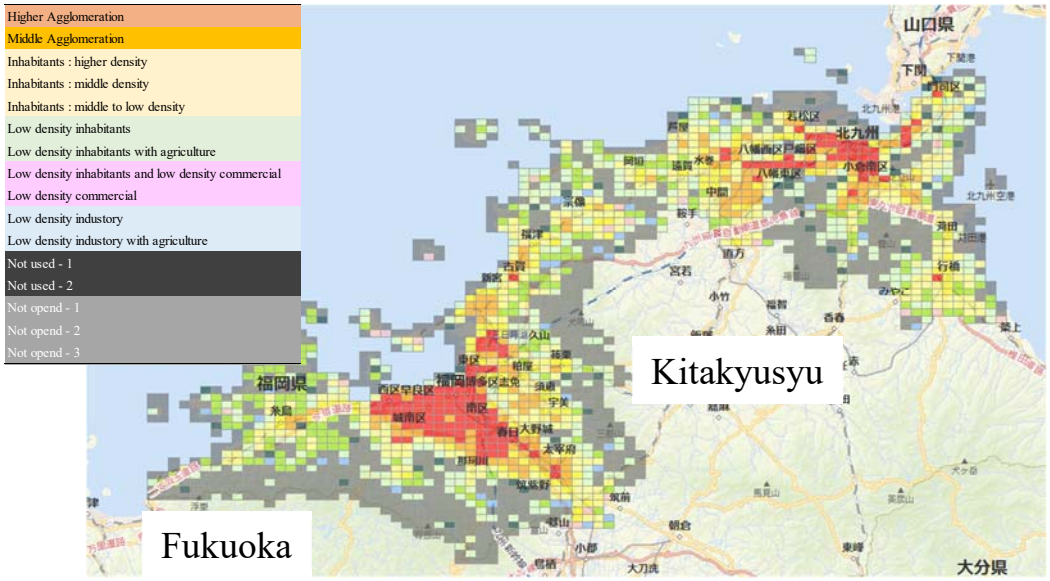

Fig.3 Distribution of dominant topics in 2000

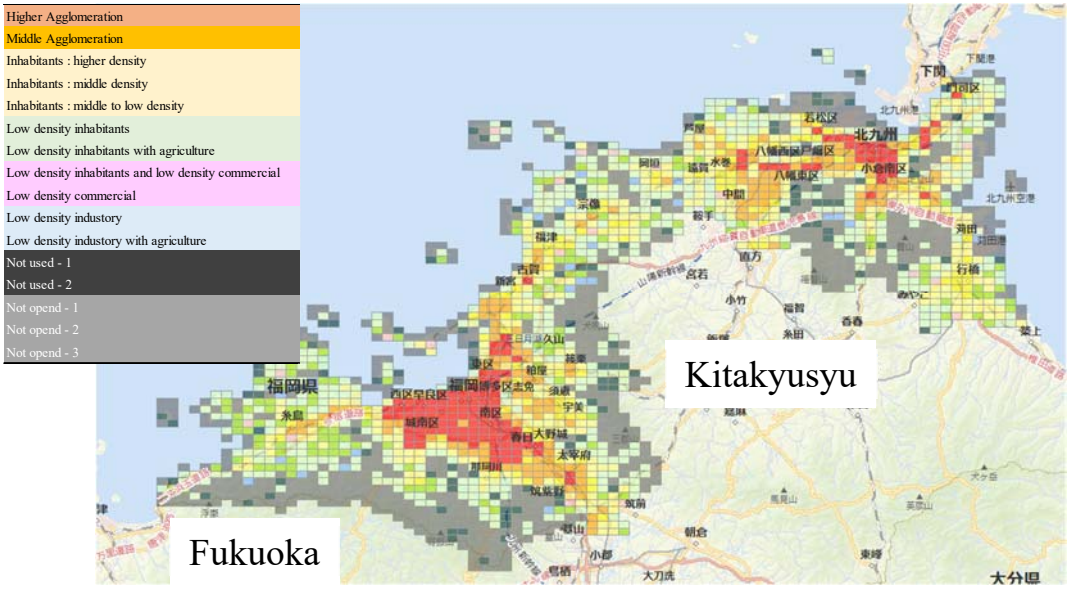

Fig.4 Distribution of dominant topics in 2010

mining in Kitakyusyu area in 1976, the main industry in this area is steel work and related productions. On the other hands, Fukuoka city is located at the Western side and it is very famous commercial area from 8 to 9 century. Fukuoka is well known as the spot of the Mongolian Invasions in 12th century, and the base of Korean troops in 16th century, and then now it is the biggest city in Kyusyu island. The higher agglomeration area colored with red is including the Shinkansen (High Speed Rail) station matching with central business district. As going outside from the CBD, less density area and more low density area are surrounding. The spatial distributions of estimated topics in both cities are naturally distributing from central business district (CBD) to suburban and to peripheral area.

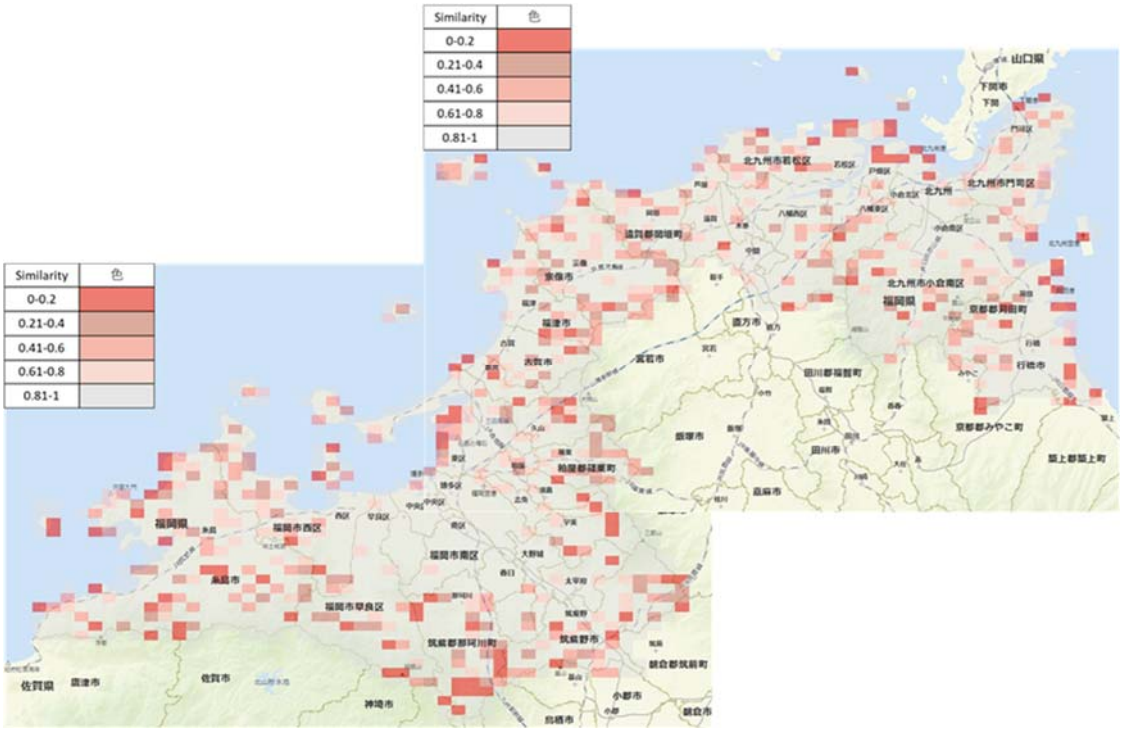

Fig.5 Change in land use by cosine similarity between 200 to 2010

Therefore, the proposed model could successfully give a quantitative characteristics of spatial topic distribution.

Comparing with fig.3 to 4, the expansion of CBD area is clarified. In Fukuoka, CBD area is expanded, and the suburban areas in 2000 are dynamically developed in 2010. Suburban area in Fukuoka is also expanded to the outside of CBD, so then the fringe area of Fukuoka city has turned to be denser land use. On the other hand, in Kitakyusyu, CBD area becomes small and low-density suburban area is increased.

In order to clarify the change in land use from 2000 to 2010, cosine similarity for each mesh is calculated. Fig.5 shows the spatial plot of cosine similarity in Kitakyusyu and Fukuoka, respectively. On this figure, the land use in both cities has been mainly changed at the surrounding area of the most agglomerated area, suburban area and the fringe area of suburban area. Looking on the railway line, the sites with some land use change are seen within the higher agglomeration or middle agglomeration dominant meshes in Kitakyusyu, while such the change are not observed in Fukuoka. F Fig.6 shows the topic share of both cities in two cross-sections. In Kitakyusyu, higher agglomeration has decreased with 0.7 points (from 5.4 to 4.7), while middle agglomeration has increased with 0.4 points (7.5 to 7.9). Sum of three topics of inhabitants has increased with 0.9 points (from 19.1 to 20.2). Sum of two topics of low density inhabitants has increased with 0.4 points (from 15.5 to 15.9). In Fukuoka, higher agglomeration has increased with 0.6 points (from 6.6 to 7.2), and also middle agglomeration has increased with 0.3 points (6.6 to 6.9). Sum of three topics of inhabitants has increased with 0.3 points

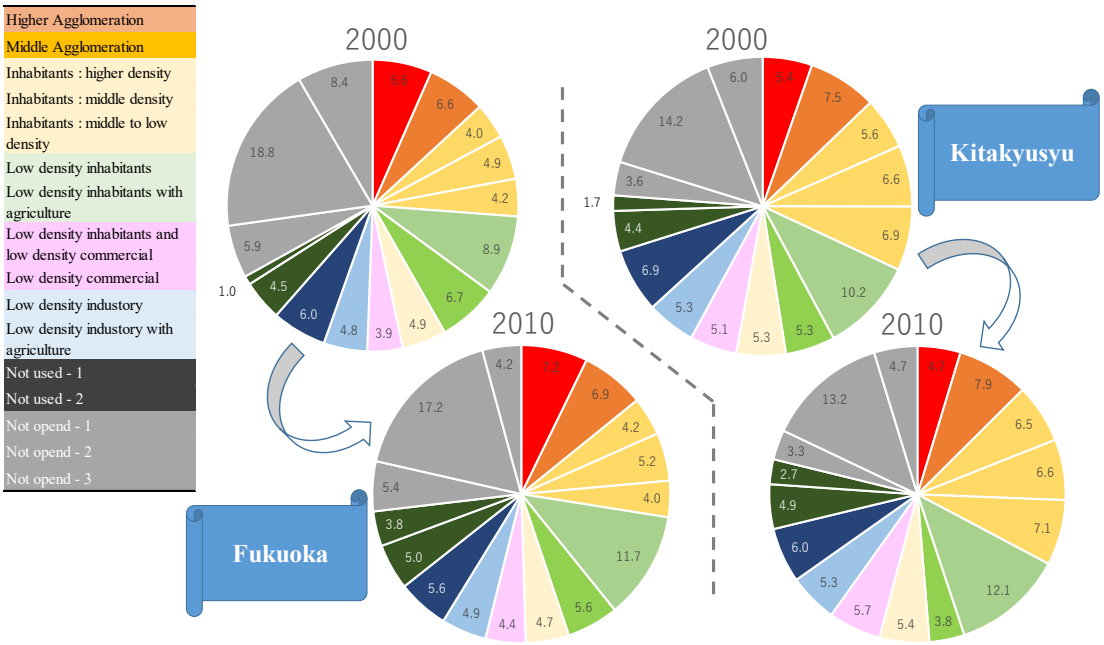

Fig.6 Topic share in Kitakyusyu and Fukuoka in 2000, 2010

(from 13.1 to 13.4). Sum of two topics of low density inhabitants has increased with 0.7 points (from 15.6 to 16.3). All the above results indicated that Kitakyusyu has been declined in CBD (higher agglomeration) but the middle density agglomeration has been increased. Observing the map of Fig.3 and 4, some of higher agglomeration has been declined to middle agglomeration. On the other hand, Fukuoka has been grown in the higher agglomerated area and in middle agglomeration. The Fig. 3 and 4 shows that some of suburban area becomes higher agglomeration area.

The development of built area in each topic are compared in both cities, shown in Table 2. Overall increase of built area from 2000 to 2010 in Kitakyusyu is 1.31, and that in Fukuoka is 1.34, so then the increase ratio is not so different. Some of notable difference of built area development except not used and not opened class are found in higher agglomeration topic (0.99 in Kitakyusyu: 1.25 in Fukuoka), low density inhabitants (1.58 in Kitakyusyu: 1.92 in Fukuoka and low density commercial (1.33 in Kitakyusyu: 1.56 in Fukuoka). The development in built area indicates the decline of CBD area in Kitakyusyu. In Fukuoka, higher agglomeration area and middle agglomeration surrounding CBD can attract buildings. Further in Fukuoka, low density inhabitants and low density commercial are also attracting buildings.

In order to clarify the access to public transportation, the topic with larger inhabitants as higher agglomeration, middle agglomeration, inhabitants: higher density and inhabitants: middle density are focused to check the public transport access. The public transport in this aggregation include the stations of Shin-kansen, other railway, subway and bus lines in 2014. Table 3 shows the change in topic share by distance band. Focusing on higher agglomeration, Kitakyusyu succeed to attract it

Table 2　Change in Built area in each topic (km^2)

Topics	Kitatyusyu			Fukuoka		
	2000	2010	2010 / 2000	2000	2010	2010 / 2000
Higher Agglomeration	31.7	31.5	0.99	69.8	87.2	1.25
Middle Agglomeration	38.9	48.5	1.25	54.5	72.1	1.32
Inhabitants : higher density	22.6	33.9	1.50	24.1	36.2	1.50
Inhabitants : middle density	19.8	25.5	1.29	22.9	33.7	1.47
Inhabitants : middle to low density	14.3	20.2	1.41	11.4	15.6	1.37
Low density inhabitants	7.1	11.2	1.58	7.1	13.6	1.92
Low density inhabitants with agriculture	5.9	6.2	1.05	9.4	11.6	1.23
Low density inhabitants and low density commercial	9.8	12.1	1.23	14.4	16.7	1.16
Low density commercial	12.9	17.2	1.33	11.4	17.8	1.56
Low density industory	6.6	10.2	1.55	6.7	9.3	1.39
Low density industory with agriculture	5.9	6.6	1.12	7	8.1	1.16
Not used - 1	2.7	3.3	1.22	2.5	3.2	1.28
Not used - 2	1.3	2.4	1.85	0.4	1.2	3.00
Not opend - 1	1.1	0.8	0.73	1.2	1.2	1.00
Not opend - 2	1.1	3.8	3.45	0.8	0.6	0.75
Not opend - 3	1.1	5.4	4.91	2	0.6	0.30
total	182.8	238.8	1.31	245.6	328.7	1.34

Table 3　Changes of topic share in distance band (pts)

Higher Agglomeration

Distaince to the nearest station	Kitakyusyu	Fukuoka
below 1 km	7.56	1.24
1 to 2 km	-6.07	-0.48
2 to 3 km	-1.33	-0.42
3 to 5 km	-0.26	-0.43
5 to 10 km	0.1	0.06
above 10 km	0	0.03

Middle Agglomeration

Distaince to the nearest station	Kitakyusyu	Fukuoka
below 1 km	-2.87	-0.9
1 to 2 km	4.04	0.13
2 to 3 km	-0.58	0.14
3 to 5 km	-0.79	0.61
5 to 10 km	0.19	0.02
above 10 km	0	0

Inhabitants : Higher density

Distaince to the nearest station	Kitakyusyu	Fukuoka
below 1 km	-2.18	-3.49
1 to 2 km	-0.9	2.97
2 to 3 km	-0.11	0.37
3 to 5 km	2.55	0.07
5 to 10 km	0.64	0.08
above 10 km	0	0.01

Inhabitants : Middle density

Distaince to the nearest station	Kitakyusyu	Fukuoka
below 1 km	-0.54	-2.46
1 to 2 km	0.28	0.57
2 to 3 km	1.35	2.21
3 to 5 km	-1.54	0.43
5 to 10 km	0.45	-0.75
above 10 km	0	0

around the public transportation (below 1 km band), while it is decreased in 1 to 2 km band. Fukuoka also shows the same tendency with Kitakyusyu, but the increase / decrease of the topic is weak. In terms of middle agglomeration, it is significantly increased in 1 to 2 km band with significant decrease in below 1 km band in Kitakyusyu, while the increase is observed in 3 to 5 km band in Fukuoka. Same as higher agglomeration, the increase / decrease of the topic is weak in Fukuoka. In case of inhabitants: higher density, significant decrease of the topic is commonly observed in below 1 km band for both cities, but the increase of the topic occurs in 3 to 5 km in Kitakyusyu, while that occurs in 1 to 2 km

band in Fukuoka. About inhabitants: Middle density, significant increase is commonly observed in 2 to 3 km band, while the decrease of the topic occurs in 3 to 5 km in Kitakyusyu, while that occurs in below 1 km in Fukuoka.

2) Discussion

Through the application of topic model to Kitakyusyu and Fukuoka city, the performance of proposed model was confirmed good, since the estimated topics were rational and its spatial distribution meets the natural expectation of land use (e.g. agglomerated topic appears at CBD area, other topics are surrounding the topics with slightly decreasing the population density or land use intensity). Due to the difference in characteristics of economic and commercial activities in both city, Kitakyusyu generally seems to be declining, while Fukuoka seems to be growing. The above acceptance can be supported with ordinal statics aggregated for municipality unit, but the detailed land use analysis in our study can give more detailed change in land use with spatial characteristics.

Kitakyusyu has many meshes to be changing its land use along the railway line on Fig.5. Such the change in Kitakyusyu cannot necessarily cause positive renovation or revitalization of the area with some changes, however, the changes indicates decline of higher agglomeration instead of increases in inhabitants related topics, as shown in fig.3 and 4. Built area analysis in table.2 also showed that decline of built area in higher agglomeration topic in Kitakyusyu. On the other hands, change of topic share in table 3 showed that higher agglomeration topic becomes closer to public transportation station, and also middle agglomeration topic attracted around the station (except the nearest distance band). Inhabitant related topics in Kitakyusyu have different dynamism with agglomeration related topics, from the viewpoint of public transport access. Table 3 tells that the topics in higher density of inhabitants and middle density inhabitants are attracted at rather far areas from public transport station. To sum up the trend in land use change in Kitakyusyu, higher agglomeration topic becomes decline, but its location has been re-organizing to be attracted around the public transportation stations. In terms of inhabitants related land use, development with lower accessibility to public transportation are observed, which would be occurred around the suburban road side area.

Fukuoka's land use change seems to be better than Kitakyusyu, in terms of higher agglomeration topic. The share of higher agglomeration and middle agglomeration has been increased, and the built area in these topics are also increased. The analysis on built area and public transport access in table 2 and 3, respectively, however, clarified the problems in Fukuoka. Built area in Fukuoka is significantly increased within low density inhabitants topics and low density commercial topics. In fig.2 and 3, these topics are appearing at the fringe of metropolitan area. These observations suggest that Fukuoka faces weak sprawl phenomena in these areas. Note that sprawl phenomena is not avoidable at a growing phase of the city. In case of Fukuoka, inner city developments such as higher agglomeration topic, middle agglomeration topic, inhabitants: higher density topic and inhabitants:

lower density topic seems to be good. Therefore, continuous watching for the whole metropolitan area is necessary to make appropriate land use regulations.

5. Conclusion

This study developed the detailed land use analysis in Tsukai and Tsukano (2018), by applying the model to two metropolitan areas in Japan. The empirical analysis in Kyusyu and Fukuoka city showed that the applicability and validity of the proposed method. The information obtained from factor load matrix is aggregated with several different ways in dominant plot mapping, weighted aggregation of topics with other related attributes and the composite topic aggregation along with the accessibility to public transportation. The advantage of proposed model is to give the detailed information in land use type, with its locational information. The combined use of the output of our model with other geographical information is useful to understand the detailed characteristics of the target area. Based on the outputs, the characteristics in land use change in both cities can be discussed.

As further analysis on the estimated topic, combination with other geographical characteristics should be discussed. For example, compactness of land use not only around the public transportation, but also road side is required to clarify the fringe land use problems. In terms of model development, more detailed classification in agglomerated area is desired for the empirical usefulness. For this purpose, not only the effective algorithm to give fine classification of input data, but also the way of pre-processing of input data should be developed. For example, about the latter issue, binning of continuous attributes (how to discretize such the attributes) should be improved. Along this issue, techniques in informatics to extract the geographical characteristics should be discussed, and then the performance of proposed process will be checked through the application of topic model.

References

Blei, A., Ng, M. and Jordan, M.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, 993-1022, 2003.

Dadashpoor, H., Azizi P. and Moghadasi, M.: Land use change, urbanization, and change in landscape pattern in a metropolitan area, Science of the Total Environment, 655, 707–719, 2019.

Flores, E., Zortea, M. and Scharcanski, J.: Dictionaries of deep features for land-use scene classification of very high spatial resolution images, Pattern Recognition, 89, 32–44, 2019.

Grigorașa, G. and Urițescu, B. : Land Use/Land Cover changes dynamics and their effects on Surface Urban Heat Island in Bucharest, Romania, International Journal of Applied Earth Observations Geoinformation, 80, 115–126, 2019.

Guan, C. : Spatial distribution of high-rise buildings and its relationship to public transit development in Shanghai, , Transport Policy, *Article in Press*, 2019.

Mohameda, A. and Workua, H. : Quantification of the land use/land cover dynamics and the degree of urban growth goodness for sustainable urban land use planning in Addis Ababa and the surrounding Oromia special zone, Journal of Urban Management, 8, 145–158, 2019.

Renne, J., Hamidi, S. and Ewing, R. : Transit commuting, the network accessibility effect, and the built environment in station areas across the United States, Research in Transportation Economics, 60, 35-43, 2016.

Sato, I. and Nakagawa, H.: Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation, KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1035-1044, 2015.

Shena, X., Wanga, X., Zhou, Z., Luc, Z. and Lv, T. : Evaluating the effectiveness of land use plans in containing urban expansion: An integrated view, Land Use Policy, 80, 205–213, 2019.

Tsukai, M. and Tsukano. M.: An Analysis on fine-scale geographical data by using topic model, Journal of Japan society of civil engineering D3, 74, 2, 111-124, 2018 (in Japanese).

Xu, W. and Yang, L. : Evaluating the urban land use plan with transit accessibility, Sustainable Cities and Society, 45, 474–485, 2019.

Zeng, C., Yang, L. and Dong, C. : Management of urban land expansion in China through intensity assessment: A big data perspective, Journal of Cleaner Production, 153, 637-647, 2017.

Zhang, C., Sargent, I., Panc, X., Li, H., Gardiner, A., Hare, J. and Atkinson, P. : Joint Deep Learning for land cover and land use classification, Remote Sensing of Environment, 221, 173–187, 2019.