

Article

Transition of Socio-Demographic Characteristics in Urban Areas by Applying a Topic Model to Small Area Units

Makoto Tsukai ^{1,*}, Satoko Ohno ² and Yuta Tsukano ³

¹ Department of Civil and Environmental Engineering, Graduate School of Applied Science and Engineering, Hiroshima University, Higashi-Hiroshima 739-8527, Japan

² Independent Researcher, Kochi 780-0850, Japan; sa06.ohstk@gmail.com

³ Kansai Airports, 1, Sensyu-Kuko-Kita, Izumisano 549-8501, Japan; yuta.tsukano@kansai-airports.co.jp

* Correspondence: mtukai@hiroshima-u.ac.jp; Tel.: +81-824-24-7827

Abstract: Under the depopulation society in Japan, the hollowing out and suburbanization of urban areas have become very serious problems, but an appropriate analytical tool for land use transition has not yet been proposed. This study analyzes the transitions in socio-demographic characteristics of small area units in the Fukuoka and Kitakyushu metropolitan areas by applying a topic model to geographical data. Plotting the topic shares on a map clarified the spatial distribution of topics, and the transitions between two cross-sections were analyzed along with other geographical characteristics. Our empirical study showed that the topic model could clearly and quantitatively describe the transitions between two cross-sections of these urban areas. The topic model revealed that the urban center of the Fukuoka metropolitan area was expanding, while the urban center of the Kitakyushu metropolitan area was shrinking. In suburban areas, both metropolitan areas had increasing low-density residential and commercial land use. In the Kitakyushu metropolitan area, this transition could seriously threaten the sustainability of land use, since the total population had significantly decreased.

Keywords: suburbanization; land use characteristics; transition; topic model; local city in Japan



Citation: Tsukai, M.; Ohno, S.; Tsukano, Y. Transition of Socio-Demographic Characteristics in Urban Areas by Applying a Topic Model to Small Area Units. *Sustainability* **2022**, *14*, 1010. <https://doi.org/10.3390/su14021010>

Academic Editors: Tatsuhito Kono and Nao Sugiki

Received: 12 November 2021

Accepted: 24 December 2021

Published: 17 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In Japan, the rapid influx of population and motorization has led to the spread of urbanized areas into low-density suburbs since the 1960s [1]. Uncontrolled suburban development, known as sprawl, leads to inefficient infrastructure provision and maintenance costs for people living in the area [2]. Suburban sprawl is associated with a decline in the residential population and commercial functions of urban centers; the expansion of residential areas has continued in the era of depopulation that began in 2008, threatening the sustainability of the regions. Ohashi and Phelps (2021) analyzed the suburban area of the Tokyo metropolitan area and argued for in-depth policy for this area, since the shrinking of the area is more complex and heterogeneous than its expansion [3].

In 2014, a land use plan named the “Grand Design of Japan in 2050” was announced by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) in Japan, which proposed a “compact city” concept composed of spatially agglomerated bases for daily life and a network of frequent public transport connecting the bases. Miyauchi et al. (2021) analyzed suburban area size and urban population to find a stable relationship among them [4]. In order to mitigate the uncontrolled expansion of urban areas or reshape them into a sustainable form, many local governments have adopted the “compact city” as a goal of urban structure in their master plan for urban design, and policies were developed to encourage people to return to the city center [5]. Such policy making requires a detailed method of understanding the current state of land use.

The development of geo-statistical data and advances in GIS technology have made it possible to analyze land use and population dynamics in fine spatial units smaller than the

municipality. In Japan, national land use data supplied by MLIT provide a land use dataset at the tertiary mesh level for multiple years across the entire land. In addition, the National Census and Economic Census provide data on population, households, and number of offices and employees, also at the tertiary mesh level. These standardized geographic information databases are useful for analyzing land use change and population dynamics, and have greatly contributed to the development of research on urban structure and urban policy [6]. However, when analyzing an entire city or region, the data handling for the number of meshes and their many attributes becomes a nuisance due to the enormous amount of information. While a map for each attribute, for example, is convenient for grasping that attribute's characteristics at a glance, a large number of maps for individual attributes cannot give insights for the comprehensive assessment of the target area.

In order to understand changes in urban structure, it is necessary to observe the changes in land use, population dynamics, commerce, and industry in a comprehensive manner, and to extract an outline of changes and notable areas [7]. If we could develop an analytical tool that can accomplish this task, it would be of great help for grasping the changes in urban structure corresponding to the real functions of the target area.

There are several approaches for grasping the land use characteristics of urbanized areas. Conventionally, images observed from satellites have been used for land cover monitoring. Grigoraşa and Urişescu (2019) estimated land cover classes to find their relationships with land surface temperature [8]. This empirical study in Bucharest, Romania showed a significant decrease in vegetation areas, and a negative correlation between vegetation and land surface temperature, resulting in heat island phenomena. By using detailed mesh information from a city, Rahman et al. (2021) proposed an integrated index to quantify the compactness of neighborhood districts in order to assess the need of a neighborhood policy for each district [9]. Renne et al. (2016) estimated a model to regress network accessibility and built environment on transit commuting share in cities in the United States [10]. Pan et al. (2018) analyzed urban structure focusing on the distance to the transportation network by using an accessibility index on small land cell units [11]. The estimated model clarified the effectiveness of public transport access. Guan (2019) analyzed the area development around public transport stations [12]. Xu and Yang (2019) clarified the relationship between public transport access, including transferring cost and land use characteristics, by using a geographically weighted regression model (GWR) [13]. Dadashpoor et al. (2019) classified landscape patterns to find the longitudinal change in the relationship between land space and land use, using GWR [14]. In this study, the drastic changes in suburban areas were clarified. Zeng et al. (2017) integrated geographical big data such as point of interest databases and OpenStreetMap with SNS big data such as Weibo to evaluate land use efficiency based on GWR [15]. Applying their model to 40 megalopolises in China, the potential and problems of the target cities were quantitatively clarified. Flores et al. (2019) proposed a novel approach for classifying land cover into groups of surface characteristics using images with very fine resolutions [16]. This study proposed an efficient algorithm for defining the dictionaries for the classification using convolutional neural networks. A comparison of the proposed algorithm with other deep learning methods showed improved performance. Mohamed and Worku (2019) tried to integrate land cover with land use in order to clarify growth and sprawl around a metropolitan area [17]. Hereafter, land use refers to land attributes such as demographic, social or economic activity data, as opposed to land cover, which is mainly observed by satellite images or aerial photographs of the land surface. Their study in Addis Ababa, Ethiopia, explored land cover classification and the resulting land cover map was overlaid with a land use map to find the correspondence or lack thereof 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 [18]. In this study, a convolutional neural network with a multi-layer perceptron was applied.

The topic model is used for a topic estimation problem from the documents with a huge number of words. Among the statistical approaches, principal component analysis or factor

analysis are well-known algorithms to make reductions of the original dataset. While the principal component or factor analysis is based on eigenvalue/eigenvector 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. Tsukai and Tsukano applied the topic model to geographic data of 23 attributes, such as the area of various types of land, the number of population and households, and the number of business offices, and extracted eight topics as land use types [19]. However, the transition of land use topic was not yet analyzed.

This study purposes to clarify the applicability of topic model to geographic data, by analyzing urban structural changes such as urban hollowing out and suburbanization from the geographically extracted “topics” (hereafter referred to as geographic topics) of the target urban area. A topic model originally used for text mining from documents is developed to apply for geographical data which gives geographic topics, in this study. The proposed model enables us to assess the transition of geographical topics. The model outputs will show how the target area has sprawled (or not sprawled) and enables to give the quantitative indices about the transition of land use. The target area of our analysis is Fukuoka and Kitakyusyu metropolitan areas because these metropolitan areas are adjacent, but their land use characteristics and transition is contrastive. Fukuoka metropolitan area has been grown in population, while Kitakyusyu metropolitan area has been suffered from depopulation. In this study, we set the empirical hypothesis as follows; the proposed model can successfully capture the characteristics of the couple of metropolitan areas. If the above hypothesis is accepted through the estimated geographical topics, we can confirm the model capacity.

2. Methods

2.1. Topic Model

The topic estimation problem is to make a summary from documents consisting of a large number of words, and the summarized information is called a topic. There are many topic models that have been proposed, and the basic idea behind them is shown in the probabilistic topic generation model using Dirichlet distribution, called Latent Dirichlet Allocation (LDA) [20]. The graphical model of topic model is shown in Figure 1. In LDA, topics are assumed to be latent variables and a document is composed of multiple topics. The following is an overview of the model discussed in our previous study [19]. In the following model specification, we use the identical terms with those in text analysis for which the topic model was originally developed, for ease of the readers to refer to the papers.

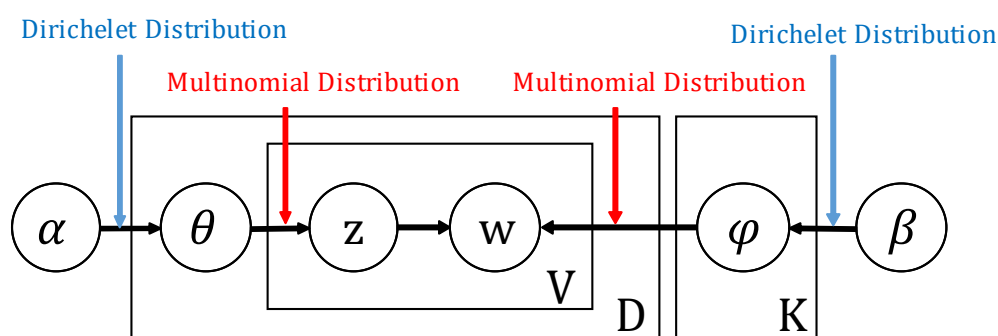


Figure 1. Graphical model of topic model.

Consider a set of D documents. Each document d consists of N^d words, and the n -th word of document d is $\{w^{n-d}\}_{n=1}^{N^d}$. The 1-of- V representation of a lexicon in a document

is a representation in which a set of vocabularies are assigned a unique number in order, and when the lexical number of the n -th occurrence of a word in each document is v , the v -th element of the vector $w^{n,d}$ is set to 1 and all the others are set to 0. $w^{n,d}$ has $N \times V$ elements in the whole document, where N is the total number of words or is the sum of the number of words in each document N^d .

In LDA, we assume that each lexicon belongs to a latent topic $z_k^{n,d} \in \{e_k\}_{k=1}^K$. We assume that each document has a different topic distribution θ_d , and each topic k has a different vocabulary distribution φ_k . If we ignore the order in which words appear in a document, and consider only the cohesion of words that co-occur in a document, these can be expressed by Equations (1) and (2) using the multinomial distribution.

$$p(z^{n,d}|\theta_d) = \text{Multi}_{K,1}(z^{n,d}; \theta_d) \quad (1)$$

$$p(w_v^{n,d}|z^{n,d}, \varphi_1, \dots, \varphi_K) = \prod_{k=1}^K \text{Multi}_{V,1}(w_v^{n,d}; \theta_d)^{z_k^{n,d}} \quad (2)$$

To estimate the parameters of the multinomial distribution θ_d and φ_k , we assume its conjugate prior distributions by the Dirichlet distribution. These are expressed in Equations (3) and (4).

$$p(\theta_d|\alpha) = \text{Dir}_K(\theta_d; \alpha) \quad (3)$$

$$p(\varphi_k|\beta) = \text{Dir}_K(\varphi_k; \beta) \quad (4)$$

We define the topic distribution of each document as the document parameters $\Theta = (\theta_1, \dots, \theta_D)^t$ with D rows and K columns. The lexical distribution of each topic is defined as the topic parameters $\Phi = (\varphi_1, \dots, \varphi_K)^t$ with K rows and V columns. The superscript t in the right shoulder indicates transposition. The simultaneous distribution of the observed data $W = \left[\left\{ w^{n,d} \right\}_{n=1}^{N^d} \right]_{d=1}^D$ and the latent variables $Z = \left[\left\{ z^{n,d} \right\}_{n=1}^{N^d} \right]_{d=1}^D$ can be expressed in Equation (5).

$$\begin{aligned} p(W, Z|\Theta, \Phi) &= \prod_{d=1}^D \prod_{n=1}^{N^d} p(w^{n,d}|z^{n,d}, \varphi_k) p(z^{n,d}|\theta_d) \\ &= \prod_{d=1}^D \prod_{n=1}^{N^d} \prod_{k=1}^K (\Theta_{d,k})^{\sum_{v=1}^V \Phi_{k,v}^{w_v^{n,d}}} \end{aligned} \quad (5)$$

For the sake of model comparison, let us eliminate the latent topics $z^{n,d}$ and obtain the marginal probabilities for W (N rows, V columns). This is expressed in Equation (6). In addition, the 1-of- V representation of W is rewritten with the data M in the bag-of-words representation of the document unit defined by Equation (7).

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

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

As shown in Equation (7), M obtained by this operation will have D rows and V columns, then Equation (8) is obtained.

$$\prod_{d=1}^D \prod_{v=1}^V ((\Theta\Phi)_{d,v})^{\sum_{n=1}^{N^d} w_v^{n,d}} = \prod_{d=1}^D \prod_{v=1}^V ((\Theta\Phi)_{d,v})^{M_{d,v}} = p(M|\Theta, \Phi) \quad (8)$$

Here, u_d is the d -th row vector of matrix $U = (u_1, \dots, u_D)^t$. Equation (6) can then be rewritten as a probability distribution over the data M , as shown in Equation (9).

$$p(M|\Theta, \Phi) = \prod_{d=1}^D N^{d!} \prod_{v=1}^V \frac{((\Theta\Phi)_{d,v})^{M_{d,v}}}{M_{d,v}!} = \prod_{d=1}^D \text{Multi}_{V,N^d}(m_d; u_d) \quad (9)$$

$$M \approx \Theta\Phi \quad (10)$$

Equation (9) is a probabilistic model of the bag-of-words data M in units of documents, which is decomposed in the products of latent parameter matrices of $\Theta\Phi$. The simplified notation of this structure yields Equation (10). Equation (10) also shows that LDA is a matrix decomposition model that approximates the observed data M with a low-rank matrix $\Theta\Phi$ whose rank is the number of topics K .

2.2. Application of Topic Model into Geographical Data

In this study, we apply the topic model to geographic information data to extract geographic topics. The advantage of the topic model is its flexibility to extract topics from a set of co-occurring vocabularies. On the other hand, geographic feature in an aggregated spatial unit (i.e., a squared-kilometers unit) is characterized by a vector of attributes in terms of land use, socio-demographic information and economic activities. As described in Section 2.1, topic model requires many documents recording the set of co-occurrence vocabularies that is expressed in 1 of V form. If the geographic data is successfully converted or processed into the form suitable for the input to topic model, we would utilize the higher ability of the topic model to extract the geographic topic giving an adequate feature of each location. For this purpose, we consider the correspondence of data format between a set of documents and the geographic data, documents and vocabularies in text data correspond to meshes and the attributes in geographic characteristics. In the following, we consider the procedure of processing geographic information data as input to a topic model. The input data for the topic model is an $N \times V$ matrix that counts the number of words V in each document N . The geographic information data is an $N \times V$ matrix with mesh N in the row direction and attributes V in the column direction. In the case of geographic information data, it is an $N \times V$ matrix with N meshes in the row direction and V attributes in the column direction.

The topic model is a sparse learning algorithm, but if we simply take the geographic attributes in the columns of input data, the number of attribute V is very small compared to the number of mesh N , and the data is not sparse. Therefore, we create an $N \times V'$ matrix data by processing the attributes with positive continuous values into a discretized classes V' . First, consider a distribution of an attribute value over all the meshes. Note that such the graph of distribution is drawn by setting certain intervals of aggregation, called classes. If we introduce a dummy variable (1/0) to indicate the correspondence of the record in an attribute of a mesh to a certain class of an attribute and the dummy variables are defined for all classes and for all attributes, we will obtain the 1-of- V representation for each mesh about the distributional information of the set of attributes. By stacking over the vectors of 1-of- V representation for each mesh, we can obtain the Bag-of-Words form with rows for meshes and columns for (discretized) attributes. The above preprocessing increases the number of spatial attributes and achieve a sparse structure of the data.

The output of the topic model is an $N \times K$ matrix Θ , where the rows represent the mesh and the columns represent the distribution of geographic topics, and a $K \times V'$ matrix Φ , where the rows represent the geographic topics and the columns represent the distribution of class-specific attributes, as shown in Figure 2. The former reveals the relationship between meshes and geographic topics, and the latter reveals the dominant class-attributes composes of a geographic topic. In Θ , the higher/lower values for columns of a row show the dominant/minor discretized attribute of the geographic topic, thus a set of them gives the characteristics of a mesh. In other words, Θ can be considered as a set of factor loading vectors of each mesh. In Φ , the higher/lower values for columns of a row show

the dominant/minor class-attributes that are likely to co-occur in a geographical topic. In other words, Φ can be considered as a set of class-attributes loading vectors of each topic, so it gives the interpretations of geographic topics.

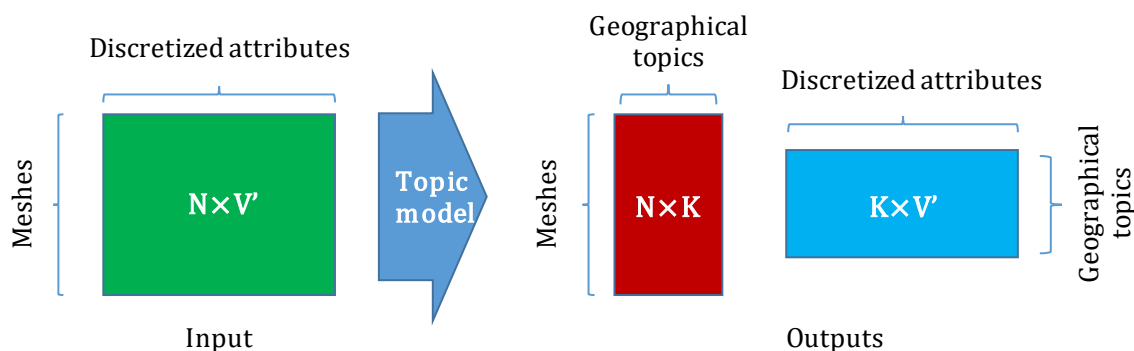


Figure 2. Input and outputs of topic model.

2.3. Model Selection

Since Dirichlet distribution is a conjugate prior distribution of multinomial distribution, the posterior distribution of the lower level also follows multiple distribution. Therefore, the estimation of topic model can rely on the Bayesian techniques. For example, 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 [21]. In their algorithm, hyper parameters are also estimated by fixed point algorithm with dozens of iterations. Unfortunately, the likelihood space of topic model does not have a single summit but multiple (steep) summits. Even though its outlook is difficult, most of the previous studies reported that the estimation of topic model is acceptably stable. In our study, several trials of initial values in estimation was made, in order to check the model stability.

The model choice is made to refer log-likelihood (LL) at the conversion of calculation in Equation (11). 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 in Equation (12).

$$L(w^{n,d}|\Theta, \Phi) = \sum_d \sum_n \log p(w^{n,d}|\Theta, \Phi) \quad (11)$$

$$\sum_d \sum_n \log p(w^{n,d}|\Theta, \Phi_0) = \sum_d \sum_n \sum_v \log \frac{1}{K} \cdot \frac{1}{V} = -N \log V \quad (12)$$

As shown in Equation (12), 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 LL at the last or the converged is used to compare the models.

In LDA estimation, the number of geographical topics to be extracted, K , has to be determined. In other words, the procedure to determine K referring to the output is left to be answered. For the sake of higher ability to describe the input data, larger K is preferred, while smaller K is desired for easier handling of the data. In this study, the likelihood ratio is to give the upper limit of the range of K . The likelihood ratio indicates the goodness of fit of the model in a statistical sense, and the number of topics with the local maximum value is the best fit as increasing of K . However, even if the statistical goodness of fit is high, the outputs contain several similar topics, and as such, the set of topics may not be easy to interpret. Therefore, K_{max} is the upper limit of the range of consideration, and if similar topics are found, they should be reduced from the estimated topic set. As an index of topic

similarity, this study uses cosine similarity between topic vectors as a measure of similarity. The similarity between topic k and topic k' , S is defined in Equation (13),

$$s_{kk'} = \frac{U_k \cdot U_{k'}}{|U_k| |U_{k'}|} \quad (13)$$

where the larger the value of $s_{kk'}$, the higher the similarity between topic k and topic k' . However, there is no clear basis for setting a threshold for the similarity criterion that would give an appropriate K . Therefore, we set multiple thresholds and check for the topics for which the number of similar topic pairs exceeding the threshold becomes zero for the largest k when the number of topics is sequentially reduced from K_{max} . In this study, the values of 0.7, 0.8 and 0.9 above $\cos 45^\circ$ are used as the thresholds of similarity. Then, we extract the geographical topics corresponding with 0.7, 0.8 and 0.9 as the thresholds of similarity, respectively. Then, the extracted results are compared, and the most suitable number of topics for analysis is adopted in terms of topic interpretation.

2.4. Methodological Feature of Geographical Topic to Measure the Sprawl Phenomena

The proposed model outputs geographic topics as a set of discretized land use characteristics and the loading vectors of geographic topics for each mesh. In order to clarify the characteristics of the proposed method, the 8 indices of sprawl phenomenon proposed by Galster et al. (2001) are reviewed [22]. The indices are density, continuity, concentration, clustering, centrality, nuclearity, mixed use and proximity. These indices are supposed to be used for mesh data, because meshes are regularly arranged in space, which simplifies their measurement. Each index, if its value is low, indicates that the mesh or a subset consisting of multiple neighboring meshes is sprawling. In our approach, density is naturally considered by making discretized attributes with its rank, so then the estimated geographic topics will show the density of land use. Mixed use would be directly measured by the loading vector of geographic topic. Proximity can be measured by mixed use for neighboring meshes. Centrality is an average distance to CBD for each attributes. Continuity is geographically gradual change of a specific attribute. In the geographical topic model, the changing of the geographical topic to the neighbor's topic corresponds to it. Concentration, clustering and nuclearity will be measured by an extent of spatially uneven distribution of geographical topics. An advantage of our approach is to extract the geographical feature of the target area is obtained by data-driven analysis, which is not possible in an ad hoc approach to use multiple indices as in [5], [6] or [9].

The above discussion indicates that we can set the following three conditions for acceptance of the hypothesis in this study such that a geographical topic is named for the density characteristics (density), that geographical topic can indicate CBD area (centrality) and that the spatial distribution of topics is gradually changing from CBD to suburban areas (Continuity). For simplicity, other features are not considered in this study.

3. Data

The input information to the topic model is 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, etc. The difference in data characteristics between the original document-vocabulary counts and our mesh-attribute records is the range of counting. The former ranges are lower in positive number, but the latter ranges are much wider. In order to convert the mesh-attribute records into BOW format, this study proposes to discretize a distribution by using multiple dummy variables for each attribute. Such data processing is called binning. By applying binning to geographical data, continuous distribution of an attribute of geographic information can be converted to BOW format.

At the estimation of the topic model, number of topics K has to be given. 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 the similar topics are merged. Topic load matrix is used to check the transition of topics between two cross-sections.

The dataset used in this study is the Kyusyu and Fukuoka metropolitan areas, recorded in 2000 and 2010. We can set a much later period for study, but any period is possible for testing the applicability of the geographical model. Another reason for setting the period 2000–2010 is that the demographic trends in Fukuoka and Kitakyushu are more contrasting in this period than in any other. The population of Fukuoka metropolitan area in 2000 and in 2010 are about 2.337 million and 2.513 million, respectively, while the population of Kitakyushu metropolitan area in 2000 and in 2010 are about 1.481 million and 1.425, respectively (Census in Japan in 2000/2010). In the Kitakyushu metropolitan area, a natural decrease of population was started in 2000 and social decrease of the population continued after 1985. In the Fukuoka metropolitan area, a natural increase and social increase of population was observed between 2000 to 2010. The sum of urbanized area is about 2000 squared kilometers. Thirty-four attributes are collected from the national census, economic census and office and company statistics, the attributes of which are shown in Figure 3. The number of meshes in the target area is totally 2113 for each cross section, by dropping the unused.

Category	Item	Category	Item
Age group population	0 to 14 age group		no. of offices in wholesale and retail
	15 to 64 age group		no. of offices in finance and insurance
	65 and above		no. of offices in real estate
Population by residence period	with less than 10 years of residence		no. of offices in accomodation and
	continuous residence since birth		no. of offices in education and learning
Workers by industry	primary industry		no. of offices in medial and welfare
	second industry	Third industry	no. of employees in wholesale and retail
	thied industry		no. of employees in finance and
	one person		no. of employees in real estate
Number or type of households by number of persons in household	two persons		no. of employees in accomodation and food services
	Three persons		no. of employees in education and learning support
	four persons		no. of employees in medial and welfare
	more than five persons	Public sector	number of offices
	nuclear family		number of employee
Housing type	housing in single-family house	land use area	paddy field
	housing in apartment		other agriculture
Second industry	number of offices		
	number of employee		

Figure 3. Thirty-four attributes inputted to geographical topic model.

In order to estimate the common geographical 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 the attributes are skewed due to the missing observation. In order to consider the information brought by them, zero or missing observation is set to be an independent class for all attributes. Each distribution was binned by natural classification. In this study, seven classes are set by several trials. Natural classification gives a set of thresholds based on the second derivative of a distribution (of each attribute). Other characteristics of the dataset is “NA: No Answer”, which is given to the mesh in which observations are one or a few, in order to mask the characteristics of a very small number of targets in a mesh. Including zero and NA, a total of nine classes are set for each attribute. Since all the attribute observation in each mesh is classified into one of nine 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, is 34 and N is $4226 * 34 = 143,484$.

4. Geographical Topics and Their Transition

4.1. Naming to Geographical Topics

Figure 4 shows the result of model in the 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 $K = 16$ with given 0.8 of the threshold in similarity. The topics obtained with other thresholds was tried, but the interpretability of them was not good. Figure 5 shows the topics we estimated. The title of topics is named by referring to KV matrix: Φ in (10), which indicates the contribution of each discretized attribute. 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 give no substantial information in land use. The “not disclosed” topics are so named since they were dominated by missing data for which no information has been recorded due to very a low number of observations in the mesh. As shown in these results, density was extracted as the characteristics of geographical topic.

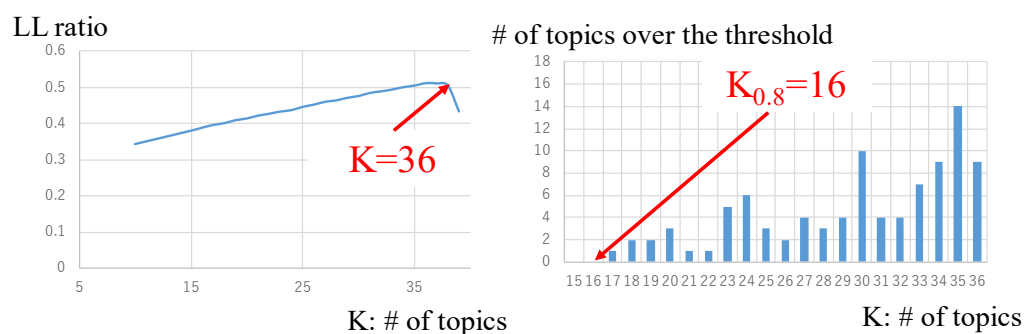


Figure 4. Finding an appropriate number of topics.

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

Figure 5. The name of 16 geographical topics.

4.2. Characteristics of the Estimated Geographical Topics

Kitakyusyu city is located at the Eastern side of the target area, and it is an industrial city in coal mining from the beginning of 20th century. Owing to the proximity to a coal mine, the first national steel work in Japan was established at Kitakyusyu in 1901. Even after the closure of coal 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 the 8th to 9th century. Fukuoka

is well known as the spot of the Mongolian Invasions in 12th century, and the dispatch of troops to Korea in 16th century, and 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.

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 in Equation (10) is shown in Figure 6 in 2000 and in Figure 7 in 2010, respectively. On this map, we can clearly see the CBD meshes for both metropolitan area, so centrality was extracted. 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, indicating the gradual change of geographical topics in space (continuity). Therefore, we can conclude that the proposed model could successfully give a quantitative characteristics of spatial distribution.

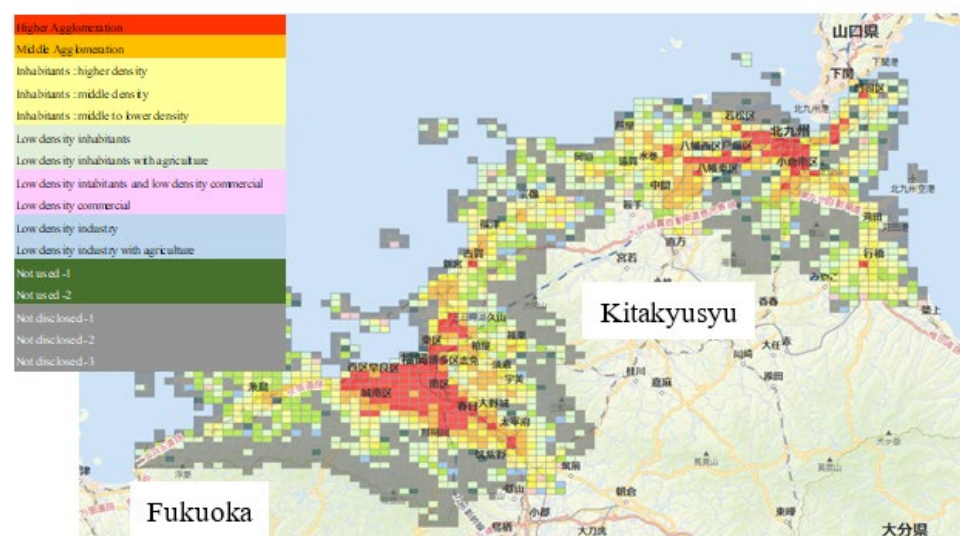


Figure 6. Distribution of dominant topics in 2000 in Fukuoka and Kitakyusyu metropolitan areas.

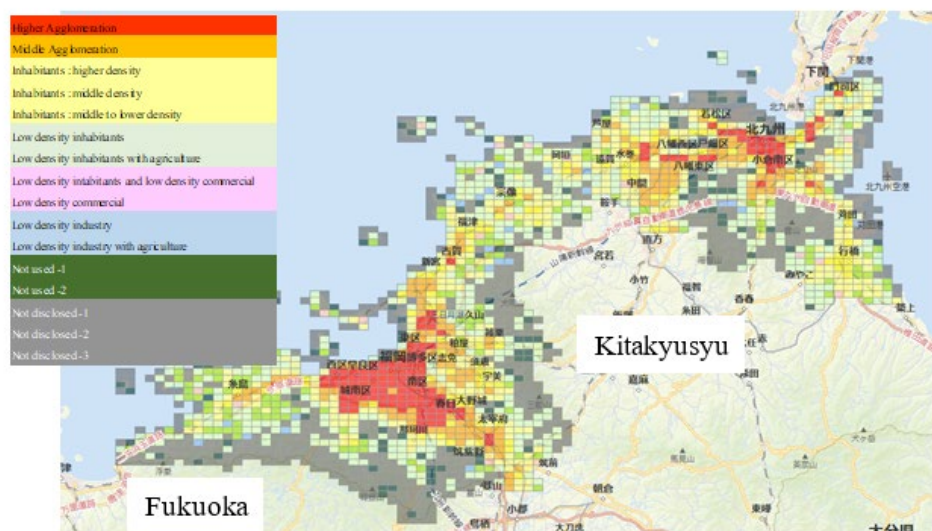


Figure 7. Distribution of dominant topics in 2010 in Fukuoka and Kitakyusyu metropolitan areas.

Comparing Figure 6 with Figure 7, the expansion of the CBD area is clarified for both metropolitan areas. 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 out to be denser land use. On the other hand, in Kitakyusyu, the 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. Figure 8 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 changes are not observed in Fukuoka.

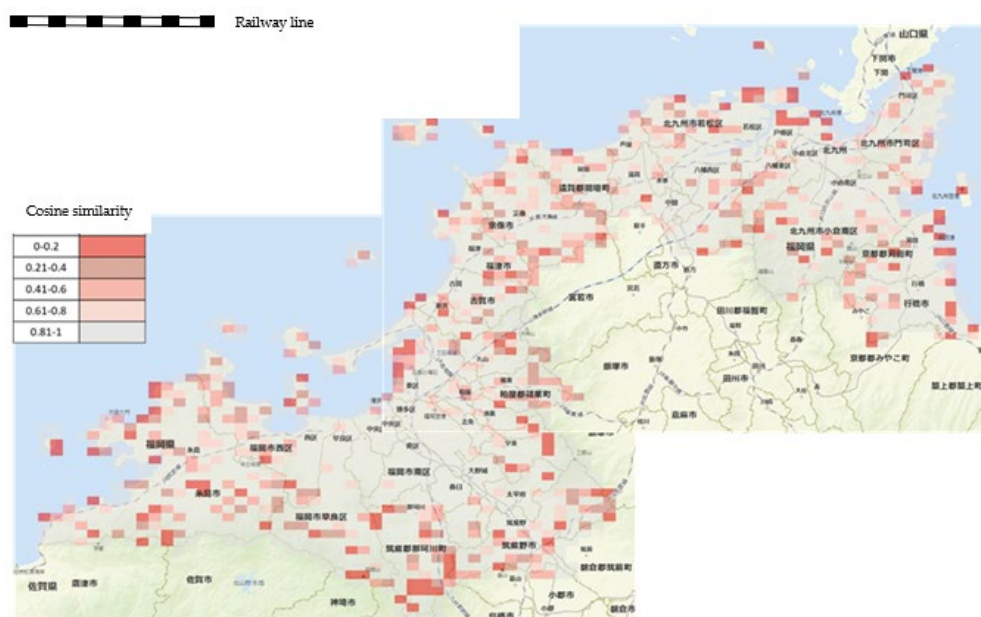


Figure 8. Change in land use by cosine similarity between 2000 to 2010 in Fukuoka and Kitakyusyu metropolitan areas.

Figure 9 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). The sum of three topics of inhabitants has increased with 0.9 points (from 19.1 to 20.2). The 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). The sum of three topics of inhabitants has increased with 0.3 points (from 13.1 to 13.4). The 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 declined in CBD (higher agglomeration), but the middle density agglomeration has been increased. Observing the map of Figures 6 and 7, some of the 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. Figures 6 and 7 show that some of the suburban area becomes higher agglomeration area.

The development of built area in each topic are compared in both cities, shown in Figure 10. 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 that as follows. Fukuoka has higher agglomeration topic (0.99 in Kitakyusyu: 1.25 in Fukuoka), higher density inhabitants (1.58 in Kitakyusyu: 1.92 in Fukuoka) and higher 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 inhabitant area and low density commercial area are also attracting new buildings.

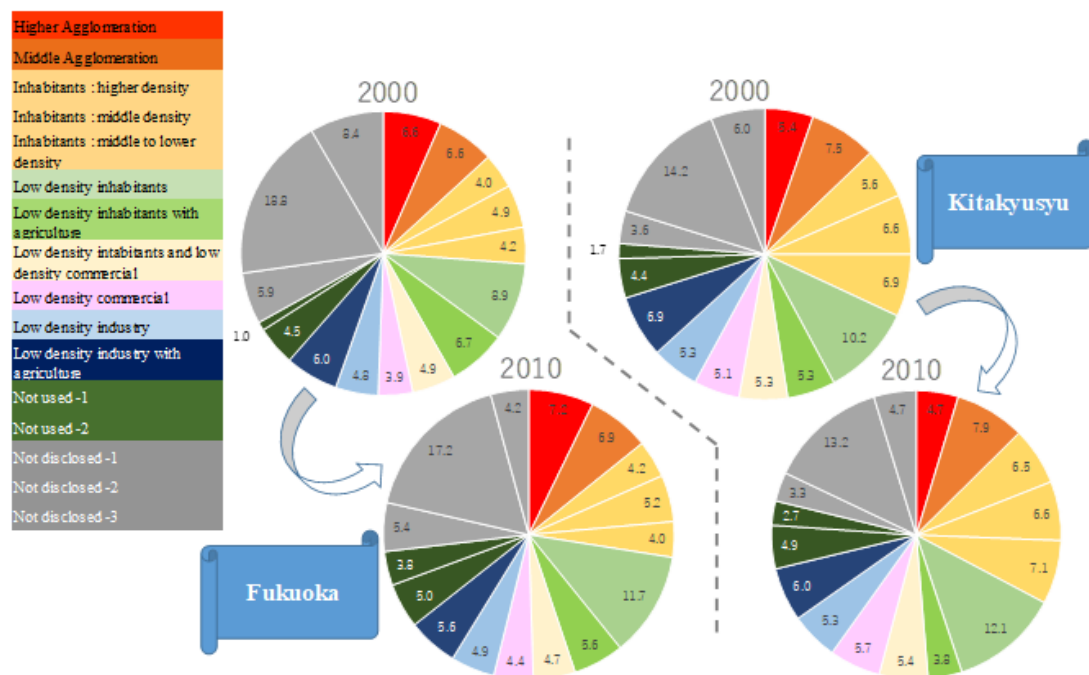


Figure 9. Geographical topic shares (%) in Kitakyusyu and Fukuoka in 2000, 2010.

	Kitakyusyu			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	72.2	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 lower 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 industry	6.6	10.2	1.55	6.7	9.3	1.39
Low density industry with agriculture	5.9	6.6	1.12	7.0	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 disclosed -1	1.1	0.8	0.73	1.2	1.2	1.00
Not disclosed -2	1.1	3.8	3.45	0.8	0.6	0.75
Not disclosed -3	1.1	5.4	4.91	2.0	0.6	0.30
total	182.8	238.8	1.31	245.6	328.7	1.34

Figure 10. Changes in built area in each topic (km²).

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 includes the stations of Shin-kansen, other railway, subway and bus lines in 2014. Figure 11 shows the change in topic share by distance band. Focusing on

higher agglomeration, Kitakyusyu succeeded in attracting it around the public transportation (below 1 km band), while it is decreased in the 1 to 2 km band. Fukuoka also shows the same tendency with Kitakyusyu, but the increase or decrease of the geographic topic was not apparent. 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 areas, the increase or decrease of the geographic topic is not so significant in Fukuoka. In the 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 the 1 to 2 km band in Fukuoka. About inhabitants: Middle density, significant increase is commonly observed in the 2 to 3 km band, while the decrease of the topic occurs 3 to 5 km in Kitakyusyu, while that occurs below 1 km in Fukuoka.

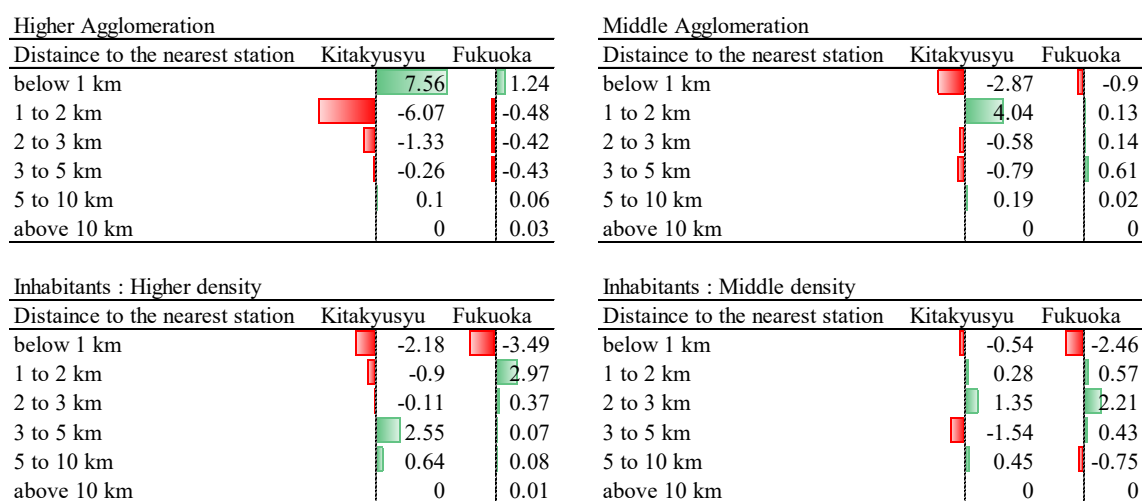


Figure 11. Changes in topic share in distance band (pts).

4.3. Discussion from the Estimated Geographical Topics

Through the application of the topic model to Kitakyusyu and Fukuoka metropolitan areas, the performance of the proposed model was confirmed. For example, agglomerated topic appears at CBD area, other topics are surrounding the topics with slightly decreasing the population density or land use intensity. In the following, the discussion for land use transition and its policy implication enabled by the observation of geographical topics will be demonstrated.

Due to the difference in characteristics of economic and commercial activities in both cities, Kitakyusyu seems to be declining, while Fukuoka seems to be growing. The above findings can be supported with ordinary statistics aggregated for whole the municipality unit, but our study can give the quantitative land use changes with the detailed location. Therefore, the working hypothesis in this study is accepted. Kitakyusyu has many meshes with changing its land use along the railway line on Figure 8. Such a change in Kitakyusyu would not be caused by the renovation or revitalization of those areas, rather by the decline of higher agglomeration areas since the topics related to inhabitants and the built areas in those topics were not increased, as shown in Figures 9 and 10, respectively. Built area analysis in Figure 10 also showed the decline of built area in higher agglomeration topic in Kitakyusyu. Considering the decreasing population with the decline, the expansion of low density land use in suburban area has been progressed in Kitakyusyu. On the other hand, from the change in topic share in Figure 11, we can see that higher agglomeration topics are moving closer to public transportation stations, and medium agglomeration topics are also gathering around stations (except the closest distance band to the stations). Inhabitant related topics in Kitakyusyu have different dynamism with business related topics, from the viewpoint of public transport access. Figure 11 showed that the topics

in higher density of inhabitants and the topics in middle density of inhabitants locate somewhat far from public transport stations. To sum up the trend in land use change in Kitakyusyu, higher agglomeration topic in business declines, but its location has been reorganized around the public transportation stations. In terms of the topics related to inhabitants, those developments with lower accessibility to public transportation were observed, which occurred around the road side in suburban area.

Fukuoka's land use change seems to be desirable rather than Kitakyusyu, in terms of higher and middle agglomeration topics in business. The share of those topics have been increased, and the built area in those topics are also increased. The analysis on built area in Figure 10 and the public transportation access in Figure 11 revealed the problems in Fukuoka. Built area in Fukuoka is significantly increased in low density inhabitant topics and in low density commercial topics. In Figures 6 and 7, these topics appear at the fringe of the metropolitan area. These observations suggest that Fukuoka faces a weak sprawl phenomenon in these areas. Even though the sprawl phenomenon is not avoidable at a growing phase of the city, land use control in the suburbs is necessary to achieve sustainable land use in Japan, where the population is declining. On the other hand, the developments in city center are acceptable, including higher agglomeration topic, middle agglomeration topic, inhabitants: higher density topic and inhabitants: lower density topic. In order to properly control the land use of an entire city, it is important to monitor the land use continuously and quantitatively.

5. Conclusions

This study made the detailed land use analysis by applying the model to two metropolitan areas in Japan, originally developed by Tsukai and Tsukano (2018) [19]. The empirical analysis in Kitakyusyu city and Fukuoka city showed the applicability and validity of the proposed model. The information obtained from the topic loading matrix can be utilized in a variety of ways, including dominant plot mapping, weighted aggregation by topics with other related attributes, and the composite topic aggregation along with the accessibility to public transportation. The estimated geographical topics fulfilled the three conditions for appropriateness of geographical topic model, discussed in 2.4. Therefore, we can conclude that the model is highly capable of capturing land use characteristics. The advantage of proposed model is to give the detailed information in land use types with their locations. Based on the outputs of the model, the characteristics in land use change in the cities were quantitatively clarified. The combined use of the output of the model with other geographical information is useful to find appropriate policies in land use.

Further analysis will make it possible to find a consistent land use policy of a city. For example, if the compactness of land use is clarified in the city center and around the road-side, a comprehensive policy for controlling core, suburban and fringe areas in a city can be discussed. From the viewpoint of usefulness in the observation of sprawl phenomenon, remaining five indices (concentration, clustering, nuclearity, mixed-use and proximity) should be measured and tested of their performance on sprawling, based on the outputs of proposed model. In terms of model development, it is desirable to increase the classification of land use patterns in urban centers in order to increase their empirical usefulness. For this purpose, a novel pre-processing of input data should be developed. For example, binning of continuous attributes (how to discretize the geographic attributes) can be improved. By comparing the performance of the models using different methods of binning attributes as a preprocessing of the input data, the limitations of the current model and the areas for improvement will become apparent.

Author Contributions: Conceptualization, M.T. and Y.T.; methodology, M.T.; software, M.T. and S.O.; validation, M.T., S.O. and Y.T.; formal analysis, S.O. and Y.T.; resources, M.T.; data curation, S.O.; writing—original draft preparation, M.T.; writing—review and editing, M.T.; visualization, S.O.; supervision, Y.T.; project administration, M.T.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://www.e-stat.go.jp/en>.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Martinez-Fernandez, C.; Weyman, T.; Fol, S.; Audirac, I.; Cunningham-Sabot, E.; Wiechmann, T.; Yahagi, H. Shrinking cities in Australia, Japan, Europe and the USA: From a global process to local policy response. *Prog. Plan.* **2016**, *105*, 1–48. [\[CrossRef\]](#)
- Hattori, K.; Kaido, K.; Matsuyuki, M. The Development of Urban Shrinkage Discourse and Policy Response in Japan. *Cities* **2017**, *69*, 124–132. [\[CrossRef\]](#)
- Ohashi, H.; Nicholas, A. Phelps. Diversity in Decline: The Changing Suburban Fortunes of Tokyo Metropolis. *Cities* **2020**, *103*, 102693. [\[CrossRef\]](#)
- Miyauchi, T.; Setoguchi, T.; Ito, T. Quantitative Estimation Method for Urban Areas to Develop Compact Cities in View of Unprecedented Population Decline. *Cities* **2021**, *114*, 103151.
- Sakamoto, K.; Iida, A.; Yokohari, M. Spatial patterns of population turnover in a Japanese Regional City for urban regeneration against population decline: Is Compact City policy effective? *Cities* **2018**, *81*, 230–241. [\[CrossRef\]](#)
- Luan, C.; Liu, R.; Peng, S. Land-Use Suitability Assessment for Urban Development Using a GIS-Based Soft Computing Approach: A Case Study of Ili Valley, China. *Ecol. Indic.* **2021**, *123*, 107333. [\[CrossRef\]](#)
- Shen, X.; Wanga, X.; Zhou, Z.; Luc, Z.; Lv, T. Evaluating the effectiveness of land use plans in containing urban expansion: An integrated view. *Land Use Policy* **2019**, *80*, 205–213. [\[CrossRef\]](#)
- Grigoraşa, G.; Urişescu, B. Land Use/Land Cover changes dynamics and their effects on Surface Urban Heat Island in Bucharest, Romania. *Int. J. Appl. Earth Obs.* **2019**, *80*, 115–126. [\[CrossRef\]](#)
- Rahman, M.H.; Islam, M.H.; Neema, M.N. GIS-Based Compactness Measurement of Urban Form at Neighborhood Scale: The Case of Dhaka, Bangladesh. *J. Urban Manag.* **2021**, *in press*. [\[CrossRef\]](#)
- Renne, J.; Hamidi, S.; Ewing, R. Transit commuting, the network accessibility effect, and the built environment in station areas across the United States. *Res. Transp. Econ.* **2016**, *60*, 35–43. [\[CrossRef\]](#)
- Pan, H.; Deal, B.; Chen, Y.; Hewings, G. A Reassessment of Urban Structure and Land-Use Patterns: Distance to CBD or Network-Based?—Evidence from Chicago. *Reg. Sci. Urban Econ.* **2018**, *70*, 215–228. [\[CrossRef\]](#)
- Guan, C. Spatial distribution of high-rise buildings and its relationship to public transit development in Shanghai. *Transp. Policy* **2019**, *81*, 371–380. [\[CrossRef\]](#)
- Xu, W.; Yang, L. Evaluating the urban land use plan with transit accessibility. *Sustain. Cities Soc.* **2019**, *45*, 474–485. [\[CrossRef\]](#)
- Dadashpoor, H.; Azizi, P.; Moghadasi, M. Land use change, urbanization, and change in landscape pattern in a metropolitan area. *Sci. Total Environ.* **2019**, *655*, 707–719. [\[CrossRef\]](#)
- Zeng, C.; Yang, L.; Dong, C. Management of urban land expansion in China through intensity assessment: A big data perspective. *J. Clean. Prod.* **2017**, *153*, 637–647. [\[CrossRef\]](#)
- Flores, E.; Zortea, M.; Scharcanski, J. Dictionaries of deep features for land-use scene classification of very high spatial resolution images. *Pattern Recogn.* **2019**, *89*, 32–44. [\[CrossRef\]](#)
- Mohamed, A.; 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. *J. Urban Manag.* **2019**, *8*, 145–158. [\[CrossRef\]](#)
- Zhang, C.; Sargent, I.; Panc, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P. Joint Deep Learning for land cover and land use classification. *Remote Sens. Environ.* **2019**, *221*, 173–187. [\[CrossRef\]](#)
- Tsukai, M.; Tsukano, M. An Analysis on fine-scale geographical data by using topic model. *JJSE D3* **2018**, *74*, 111–124. (In Japanese) [\[CrossRef\]](#)
- Blei, A.; Ng, M.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
- Sato, I.; Nakagawa, H. Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation. In Proceedings of the KDD '15 the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 1035–1044.
- Galster, G.; Hanson, R.; Ratcliffe, R.; Wolman, H.; Freihage, J. Wrestling Sprawl to the Ground: Defining and Measuring an Elusive Concept. *Hous. Policy Debate* **2001**, *12*, 681–717. [\[CrossRef\]](#)