Spatial Cluster Detection on Detailed Data without Constraint of Continuousness

Akihito Ujiie and Junya Fukumoto

Abstract

Spatial cluster is the set of geographical units where concentration of events is observed. Spatial clusters provide useful information for understanding mechanism and characteristic of socioeconomic activities. Several methods have been proposed for cluster detection. However, there is no existing method relaxes a constraint on adjacency of geographical units that compose clusters. Constraint that requires exact adjacency may have significant impact on detected clusters, especially in the case of detailed data. In this study, we propose a new cluster detection method relaxes constraints on shape and adjacency. Along the lines of model-based clustering, we assume spatial data arise through a probabilistic model. Employing Potts model on the probabilistic model, we can embed constraints on shape in the probabilistic model and relax constraints on geometric shape. The applicability of the proposed method is tested on case studies using mesh data of Japanese economic census.

A. Ujiie (Corresponding author) • J. Fukumoto
Graduate School of Information Sciences, Tohoku University, Aoba 6-6-06, Aramaki, Aobaku, Sendai, JP
Email: khujiie@plan.civil.tohoku.ac.jp

J. Fukumoto
Email: fukumoto@plan.civil.tohoku.ac.jp

1 Introduction

All socioeconomic activities are held on a geographical space. The phenomenon on the geographical space is more or less dependent on its location. It is useful for understanding mechanism and characteristic of socioeconomic activities to focus on relation between socioeconomic activities and their location. In fact, researchers on new economic geography (NEG) focus on spatial property of economic activities and produce impressive results.

Recently, it has become much easier to get detailed spatial data such as mesh data or movement-locus data obtained by GPS. There exist a lot of methods that help us get exploratory findings from spatial data. Cluster detection analysis is one of the exploratory analyses for spatial data. Detected clusters will be useful for understanding socioeconomic activities. For example, detecting industrial agglomerations will provide empirical examination for results of NEG and basic information for decision-making of government on industry policies and urban planning. Several methods have been proposed for cluster detection in epidemiology and criminology for the purpose of investigating cause of infection disease or effective prevention of crimes. In this paper, we focus on cluster detection on discrete geographical space consists of geographical units such as municipalities or meshes.

In cluster detection, analysts face the problem of deciding constraints on geometric shape of spatial clusters. Absence of objective criterion makes spatial clustering analysis difficult. This problem has been pointed out by researches that proposed cluster detection methods such as Besag & Newell (1991), Kulldorff & Nagarwalla (1995), and Kulldorff (1997). There have been various ways to assume geometric shape of spatial clusters: circle (Kulldorff & Nagarwalla (1995)), ellipse (Kulldorf et al. (2006)), convex hull (Mori & Smith (2013)) and set of adjacent geographical units (Duczmal & Assunção (2004), Tango & Takahashi (2005) and Inoue et al. (2013)). However, there is no existing method relaxes a constraint on adjacency of geographical units. Constraint that requires exact adjacency may have significant impact on detected clusters, especially in the case of detailed data. There exist two main reasons to assume adjacency of geographical units in existing methods. First, analysts have to set arbitrary criterion that control the maximum distance between non-adjacent geographical units that belong to the same cluster. Second, assumption of adjacency simplifies the cluster detection problem by reducing the number of combination of geographical units to be considered. However, there is no guarantee real spatial clusters are composed only of adjacent geographical units. In the case of detailed spatial data, we argue that assumption of adjacency is not valid.

The aim of this study is to propose a new cluster detection method that relaxes constraints on shape and adjacency of geographical units that compose spatial clusters. Along the lines of model-based clustering, we assume spatial data arise through a probabilistic model. Employing Potts model on the probabilistic model, we can detect clusters through relaxed constraints on geometric shape. The applicability of the proposed method is tested on case studies using mesh data of Japanese economic census.

2 Existing methods

A large number of statistical methods have been developed for spatial data. These methods can be grouped into two categories. Methods in first group describe quantitative relationship among spatial data by means of regression models. First group includes a lot of methods developed in geostatistics and spatial econometrics. Another group consists of methods that provide exploratory findings. Techniques referred to as exploratory spatial data analysis (ESDA), which was first defined by Anselin (1994), are categorized as second group. Second group includes cluster detection methods and spatial clustering methods. Cluster detection and spatial clustering have a certain thing in common that both of the methods detect set of geographical units. However, we can draw a difference between the two analyses. Cluster detection is intended to find accumulation of events. On the other hand, spatial clustering is intended to find classification of geographical space.

Methods for cluster detection and spatial clustering can be categorized into two approaches on the way to deal with a probabilistic model that describes generation process of spatial data. Methods in first approach get detection results by estimating the probabilistic model based on observed data. Methods in second approach get detection results without estimating the probabilistic model. First group includes cluster detection method proposed by Mori & Smith (2013) and spatial clustering methods based on model-based clustering. Second group includes cluster detection method named spatial scan statistic and spatial clustering methods based on density-based clustering.

As far as we know, all existing cluster detection methods assume exact adjacency of geographical units that compose a cluster. On the other hand, spatial clustering methods such as model-based clustering and density-based clustering allow us to relax the constraint. One of the advantages of model-based clustering is that it has potential to allow us to extend cluster detection method to a method that combines multiple spatial data.

In what follows, we review representative cluster detection methods and model-based clustering approach.

2.1 Cluster detection method

2.1.1 Spatial scan statistic

Kulldorff & Nagarwalla (1995) (KN) proposed spatial scan statistic finds a statistically-warranted cluster. Spatial scan statistic is the one of the most widely used cluster detection methods. KN referred a candidate of spatial cluster as window, which is the set of geographical units that meets the constraints on geometric shape of a cluster. In spatial scan statistic, whole study area is searched by windows; one of the windows is judged as spatial cluster in cases where the window has the most significant accumulation. More precisely, KN assumed observed data is realization of a probabilistic model. Poisson distribution is commonly used. Under this assumption, KN made a null hypothesis and an alternate hypothesis. Null hypothesis argues there is no difference in degree of accumulation between inside and outside of the window. On the other hand, alternate hypothesis argues the degree of accumulation inside the window is higher than that of outside of the window. The likelihood ratio of these two hypotheses is calculated to evaluate each window. Spatial scan statistic has a problem with detecting multiple clusters because recursive application of it confronts multiple testing problem. This limitation comes from the fact that the null and alternate hypotheses assume that there exists only one cluster in the study area.

KN assumed the shape of a spatial cluster was circle. But, there is a good chance that real spatial clusters have various geometric shapes. For example, cluster along a railroad or river ought to be slim-line. Duczmal & Assunção (2004), Tango & Takahashi (2005) and Inoue et al. (2013) proposed spatial scan statistic assume that a cluster to be combination of adjacent geographical units.

2.1.2 Mori & Smith (2013)

Mori & Smith (2013) (MS) proposed a cluster detection method detects more than one cluster simultaneously for the purpose of detecting industrial agglomerations. Evaluating all candidates of cluster (i.e. windows defined by KN) together, MS evaded multiple testing problem. MS referred set of windows as cluster scheme. MS assumes observed spatial data arise through a probabilistic model that describes the relationship between spatial data and a cluster scheme. Spatial clusters are detected through a selection of one cluster scheme on the basis of Bayesian information criterion (BIC). MS employs dartboard model that regards each establishment as a dart. Under a certain cluster scheme, an establishment selects one window with a probability proportional to value of a parameter estimated for each window and subsequently selects one geographical unit in the window with a probability proportional to its square measure. Additionally, analysts need to constrain geometric shape of spatial clusters because MS's model is independent from shape of clusters. MS assumes a cluster has to meet following two constraints. First constraint is that shape of a cluster is convex hull on a network that declares time-distance among geographical units. Another constraint is a cluster must not have any geometric hole.

2.2 Model-based clustering

In image understanding and pattern recognition, a lot of approaches for image segmentation have been developed; Cheng et al. (2001) provides one of reviews. Image segmentation methods allocate each picture element to one of groups where the same object is caught in original image. One of the major approaches is model-based clustering, which is employed by Celeux et al. (2003), Chen et al. (2005) and Cucala & Marin (2013) etc.

Model-based clustering assumes observed image data arise through a probabilistic model that describes the relationship between image data and a configuration, which denotes certain image segmentation. Image segmentation is carried out by searching configuration that has largest posterior probability. The probabilistic model has two phases. First, a configuration rises up stochastically under a prior distribution. Second, image data is generated based on the configuration and likelihood. In the framework of Bayesian statistics, likelihood is a probabilistic model that describes the relationship between realized configuration and image data.

A number of researchers employ q-state Potts model on prior distribution of configurations. Q-state Potts model is the probabilistic model proposed in statistical physics for the purpose of explaining the behavior of magnetic materials. Let's suppose that every lattice point of two-dimensional square lattice has a spin which takes one of multiple states; the number of states each spin can take is denoted by q. Q-state Potts model gives relatively high probability to a configuration where close-set spins take the same state. Regarding each spin and spin's state as picture element and group of the element, respectively, this model can be employed as a prior distribution in image segmentation. By doing so, analysts can embed spatial autocorrelation of image data into the probabilistic model that describes the generation process of observed image.

3 Cluster detection with relaxed constraints

We propose a new cluster detection method based on the framework of model-based clustering. We assume spatial data arise through a probabilistic model. Regarding each spin and spin's state as geographical unit and region that the unit belongs to, respectively, we employ q-state Potts model on prior distribution of the probabilistic model that describes generation process of observed spatial data. This approach enables us to relax constraints on adjacency of units.

3.1 Framework

In this section, we explain the framework. We suppose the study area is discrete geographical space that consists of N geographical units. The observed data in geographical unit i is denoted by f_i ; $\mathbf{f} \equiv (f_1,\cdots f_N)^T$. We assume there exists K spatial clusters and a non-cluster region in the study area; every geographical unit belongs to one of those $(K+1)$ regions. The region that geographical unit i belongs to is denoted by a_i ($\in \{0,1,\cdots,K\}$); configuration of spatial cluster is denoted by $\mathbf{a}$ ($\mathbf{a} \equiv (a_1,\ldots,a_N)^T$), which is equivalent to MS's cluster scheme in terms of holding information. We assume observed spatial data is generated through following probabilistic process.

1. One configuration rises up under a prior distribution $p(\mathbf{a} \mid \beta)$. β is set of parameters included in the prior distribution.
2. Spatial data $\mathbf{f}$ is generated under a conditional probability distribution on $\mathbf{a}$ denoted by $p(\mathbf{f} \mid \mathbf{a},\gamma)$, which is called as likelihood in Bayesian statistics. γ is set of parameters included in likelihood.

Under those assumptions, we detect clusters by selecting a configuration on the basis of ICL, which is the one of most commonly-used model selection criteria in model-based clustering. Birnacki et al. (2000) proposed ICL defined by Eq. 3.1.

$$ICL(K) = \int p(\mathbf{f},\mathbf{a} \mid \theta_K)\pi(\theta_K)d\theta_K \tag{3.1}$$

$\theta_K \equiv \beta \cup \gamma$. Birnacki et al. (2000) approximate Eq. 3.1 by Eq. 3.2.

$$ICL(K) \approx \log p(\mathbf{f}, \hat{\mathbf{a}} \mid \hat{\theta}_K) - \frac{v_K}{2} \log N \tag{3.2}$$

$\hat{\theta}_K$ is maximum likelihood estimate of parameters. v_K is the number of parameters included in the model. $\hat{\mathbf{a}}$ denotes configuration that has largest posterior probability under $\hat{\theta}_K$; $\hat{\mathbf{a}}$ defined by Eq. 3.3.

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} p(\mathbf{a} \mid \mathbf{f}, \hat{\theta}_K) \propto p(\mathbf{f}, \mathbf{a} \mid \hat{\theta}_K)\pi(\hat{\theta}_K) \tag{3.3}$$

BIC employed by MS could lead to overfitting in terms of clustering because it selects the model that best fits observed data. On the other hand, the purpose of ICL is to select the most valid model for clustering and ICL tends to select a model that has smaller number of clusters than BIC. Baudry et al. (2014) said that ICL is known to select stable and reliable number of clusters in practice. In this study, we employ ICL as model selection criteria and select a configuration based on ICL. The spatial cluster detection problem is defined by Eq. 3.4.

$$\begin{aligned}(\hat{\mathbf{a}}, \hat{\beta}, \hat{\gamma}) &= \arg\max_{\mathbf{a},\beta,\gamma} ICL(\mathbf{a}, \beta, \gamma) \\ &= \arg\max_{\mathbf{a},\beta,\gamma} \ln p(\mathbf{f} \mid \mathbf{a}, \gamma) p(\mathbf{a} \mid \beta) - \frac{v_K}{2} \ln N\end{aligned} \tag{3.4}$$

All parameters are supposed to have non-informative prior distribution.

3.2 Formulation of likelihood and prior distribution

We argue there are three requirements for the probabilistic model describing the generation process of data. First requirement is that the probabilistic model takes relatively high probability in cases where geographical units which have high observed value belong to cluster regions apart from low-value geographical units. Second requirement is that the probabilistic model takes relatively high probability in cases where close-set geographical units belong to the same regions. Third requirement is that the probabilistic model takes nearly zero in cases where the configuration has a cluster that violates constraints on geometric shape. In our probabilistic model, multivariate normal distribution which is employed as likelihood expresses the first requirement; q-state Potts model which is employed as prior distribution expresses the second and third requirements.

3.2.1 Likelihood

We employ multivariate normal distribution defined by Eq. 3.5 on likelihood.

$$p(\mathbf{f} \mid \mathbf{a}, \boldsymbol{\mu}, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(f_i - \mu_{a_i}\right)^2}{2\sigma^2} \right\} \tag{3.5}$$

$\boldsymbol{\mu}(\equiv \{\mu_0, \ldots, \mu_K\})$ and σ are parameters; $\gamma = \boldsymbol{\mu} \cup \sigma$. This likelihood takes high probability in cases where differences in data among geographical units that belong to the same region are small. This means likelihood of Eq. 3.5 satisfies the first requirement explained in the opening of section 3.2. The estimate of parameters derived from Eqs. 3.4-3.5 is as follows.

$$\begin{cases} \hat{\mu}_k = \sum_{i \in C_k} f_i \Big/ N_k \\ \hat{\sigma} = \sqrt{\sum_i \left(f_i - \mu_{a_i}\right)^2 \Big/ N} \end{cases} \tag{3.6}$$

N_k denotes the number of geographical units belonging to region *k*. C_k denotes the set of geographical units belonging to region *k*.

3.2.2 Prior distribution

We utilize q-state Potts model defined by Eq. 3.7 to define prior distribution.

$$p(a \mid \alpha) = \frac{1}{Z(\alpha)} \exp\left(-H(a, \alpha)\right) \tag{3.7}$$

α is a parameter ($\alpha \geq 0$); Z is normalization constant. H denotes energy function of Potts model defined by Eq. 3.8.

$$H(a, \alpha) = -\sum_{k \in \{0, \cdots, K\}} \sum_{(i,j) \in N_{C_k}} \frac{1}{2} \alpha w_{ij} \tag{3.8}$$

The spatial weight coefficient between geographical units i and j is denoted by w_{ij}. N_{C_k} denotes the set of all pairs of neighboring units in region *k*. Configurations with low energy get high probability in the prior distribution. This prior distribution gives prior probability which is significantly bigger than 0 to every configuration. Therefore, original Potts model has the potential to permit configurations that have peculiar spatial clusters. We make a proposal to add new penalties which express constraints on geometric shape of spatial clusters into energy function of Potts model in or-

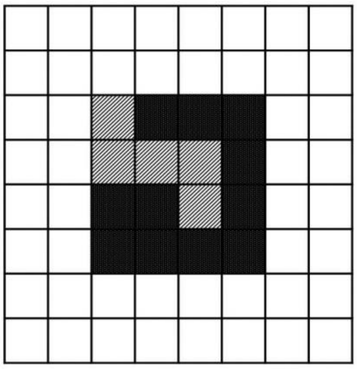


Fig. 1. Example of configuration that complication penalty is intended to avoid

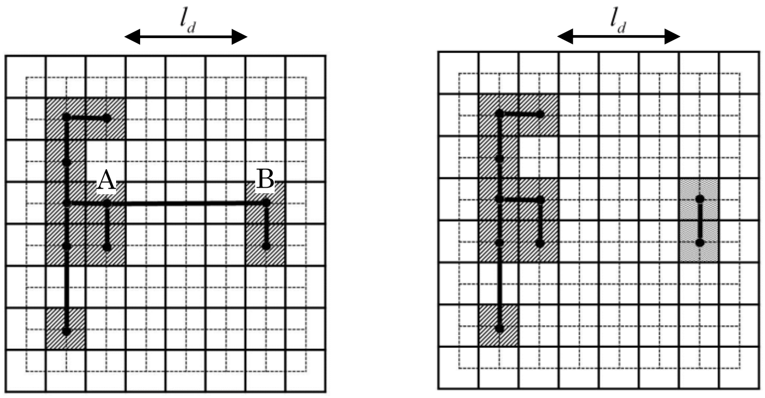
der to deal with this problem. Heavily penalized configurations get nearly zero prior probabilities.

We propose two penalties. First penalty named *distance penalty* is intended to avoid configurations that have a cluster consists of critically separated geographical units. Original Potts model permits configurations have a cluster that consists of distant geographical units. We need to pose a penalty on the energy function to evade such configurations. Second penalty named *complication penalty* is intended to avoid complex configurations where a lot of clusters are in a jumble like Fig. 1. In Fig. 1, each grid square expresses geographical unit; a pattern of each square shows the region where the unit belongs. Plain units belong to non-cluster region. Likelihood defined by Eq. 3.5 has the potential to give high probability to such configurations because it is independent from geometric shape of spatial clusters. We need to pose another penalty to evade such configurations.

To embody the idea of penalties, we utilize imaginary time-distance network. We think it is best to employ the real transportation network for detecting clusters of socioeconomic activities. But, it is difficult to embed real transportation network in the probabilistic model. For this reason, we define the distance between two geographical units on imaginary time-distance network denoted by *G*. *G* is undirected and weighted graph. Nodes of *G* are geographical units and links of *G* are put between adjacent geographical units. The weight of link is time-distance between the geographical units.

Utilizing graph *G*, we formulate the *distance penalty* as Eq. 3.9.

$$P_d = \sum_{k\in\{1,\ldots,K\}} \sum_{(i,j)\in L_{C_k}} \left(d_{ij}\delta(d_{ij} > l_d) \right) \tag{3.9}$$



a Configuration that has distance penalty

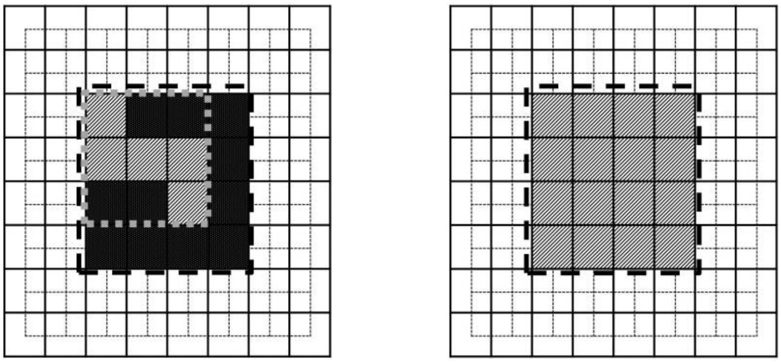
b Configuration that doesn't has distance penalty

Fig. 2. Distance penalty

P_d denotes the value of distance penalty. l_d is a parameter that define the maximal distance that is not penalized. d_{ij} is the shortest distance between geographical unit *i* and *j* on network *G*. L_{C_k} is the set of links included in minimum spanning tree denoted by T_k , which is defined on complete graph G_k . G_k is undirected and weighted graph. Nodes of G_k are geographical units that belong to region *k* and the weight of links is the shortest distance on *G*. $\delta(d_{ij} > l_d)$ denotes delta defined in Eq. 3.10.

$$\delta(d_{ij} > l_d) = \begin{cases} 1 & \mathit{if} \ d_{ij} > l_d \\ 0 & \mathit{otherwise} \end{cases} \tag{3.10}$$

Fig. 2 shows the idea of distance penalty defined by Eq. 3.9. In Fig. 2, each grid square drawn by solid line expresses geographical unit. The dashed grid expresses the time-distance network *G*. We suppose that adjacent geographical units share a side of squares and time-distances between all adjacent units are the same. The length of two-headed arrow express the l_d . Pattern of square shows the region where the geographical unit belongs. Plain units belong to non-cluster region. Bold lines express minimum spanning tree *T*. The configuration of Fig. 2a is penalized because minimum spanning tree *T* has the link whose the time-distance is longer than l_d between units A and B. On the other hand, the configuration of Fig. 2b is not penalized because the penalized link at Fig. 2a was removed. Formulating distance penalty by Eq. 3.9, we can avoid clusters consists of critically separated units.



a Configuration that has complication penalty

b Configuration that doesn't has complication penalty

Fig. 3. Complication penalty

We formulate the *complication penalty* as Eq. 3.10, utilizing convex hull on time-distance network *G*.

$$P_c = \sum_{k\in\{1,\ldots,K\}} \sum_{k'\in\{1,\ldots,K\}\backslash k} s_{k,k'} \tag{3.11}$$

$s_{k,k'}$ denotes an square measure of overlapping area between convex hulls of region k and . The convex hulls are defined on graph *G*. Fig. 3 shows the idea of complication penalty. Fig. 3 is drawn in the same way with Fig. 2. Squares drawn in heavy dashed line express convex hulls on time-distance network *G*. The configuration of Fig. 3a is penalized by complication penalty because convex hulls are overlapping. On the other hand, the configuration of Fig. 3b is not penalized because overlap of two convex hulls observed in Fig. 3a was resolved. Formulating complication penalty by Eq. 3.11, we can avoid complex configurations where a lot of clusters are in a jumble.

Adding these two penalties into energy function of Potts model, geometric shape of clusters is constrained. The prior distribution is defined by Eq. 3.12.

$$p(\mathbf{a}\,|\,\alpha,b_d,b_c,l_d) = \frac{1}{Z(\alpha,b_d,b_c,l_d)}\exp\left(-H(\mathbf{a},\alpha,b_d,b_c,l_d)\right) \tag{3.12}$$

b_d and b_c are parameters. The energy function is defined by Eq. 3.13.

$$H(\mathbf{a},\alpha,b_d,b_c,l_d) = -\sum_{k\in\{0,1,\cdots,K\}} \sum_{(i,j)\in L_{C_k}} \frac{1}{2}\alpha w_{ij} + b_d P_d + b_c P_c \tag{3.13}$$

In this paper, analysts set value of parameters included in the prior distribution. By doing so, we can evade calculation of intractable normalization constant. Estimation of these parameters is one of challenges for the future.

3.3 Solving method employing genetic algorithm

The spatial cluster detection problem defined by Eq. 3.4 is combinatorial optimization problem. We suppose it is nearly impossible to get globally optimal solution. We use genetic algorithm (GA) as a heuristic search technique to the problem. We regard configuration and ICL as gene and fitness function, respectively. Termination condition of GA is that absolute value of difference between best ICL value and mean ICL value is smaller than threshold level for 5 successive generations.

In this study, GA has three operators: local search, mutation and duplication of competent individuals. Based on following observation, we add local search as operator of GA so that GA reaches good solution in a short time. Though it is nearly impossible to get globally optimal solution, we can roughly expect the characteristic of competent individuals. By the formulation of likelihood and prior distribution, we can say that competent individuals have penalties-zero spatial clusters consist of geographical units where observed data are high. However, local search is more likely to go into locally optimal solution. Mutation plays the primary role in avoiding locally optimal solution. Generally, locally optimal solutions are searched through crossover, which generates new individuals. In our algorithm, locally optimal solutions are searched through the local search. We suppose we don't need to search locally optimal solutions again by crossover. We just duplicate competent individuals as an alternative to crossover. Duplication of individuals makes our algorithm faster because it evades calculation of ICL. As a matter of fact, the most time-consuming part is calculation of minimum spanning tree T and convex hull on time-distance network G for every cluster. We can make the algorithm faster by decreasing the number of calculation of ICL.

4 Application

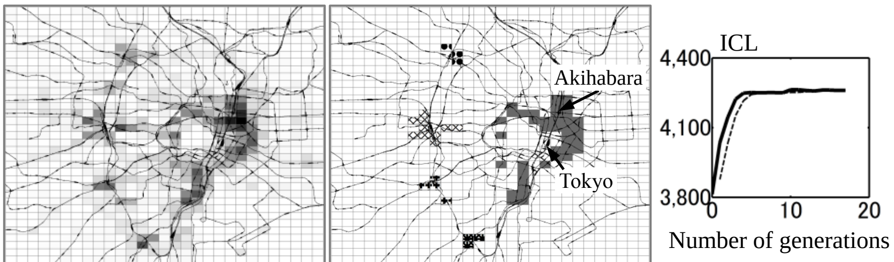
In this chapter, the applicability of the proposed method is tested through case studies using aggregated data of 2009 Japanese economic census that shows the number of establishments for each *nibunnoichi chiiki* mesh, which is approximately square, 500 m on a side. The study area is approximately square area 16km on a side that consists of 1073 meshes of heart of Tokyo. The proposed method is applied on data that shows the density of establishments for each mesh. Our method is also intended to be applied to data aggregated on municipalities, which have great difference in square measure. So, observed data is standardized by square measure. The square measure is calculated on polygon data which can be downloaded from the web page managed by Japanese National Statistic Center (http://www.e-stat.go.jp/SG1/estat/eStatTopPortalE.do). In this paper, time-distance between adjacent meshes is set by straight-line distance between median points of meshes. We assume that adjacent meshes share a side of mesh. As to spatial weight matrix, we suppose $w_{ij} = 1/d_{ij}^2$.

Parameters of the prior distribution are set up as $\alpha = 4$, $b_d = b_c = 10^6$ and $l_d = 2$ km. α is non-negative parameter. As α takes larger value, close meshes are more likely to belong to a same region. If α is too big, we cannot detect clusters because all meshes belong to a single region. We chose an adequate value for parameter α so that we would get meaningful results. b_d and b_c are set to large value enough to dismiss penalized configurations. As to l_d , we chose adequate value so that we could detect multiple clusters that consist of non-adjacent meshes. In this paper, we show results of two industries both are classified in information and communications industry.

4.1 Spatial clusters of information services industry

Fig. 4 shows the result of spatial cluster detection for information services industry. Fig. 4a shows the density of establishments; dark meshes have high density. We also show the railroad network in Fig. 4a and b. Fig. 4b shows the result of cluster detection; pattern of each mesh shows the region where the mesh belongs. Plain meshes belong to non-cluster region. Fig. 4c shows the convergence process of GA. Solid line and dashed line expresses the best ICL value and mean value of ICL of each generation, respectively.

We can point out following three points as to Fig. 4. First, fig. 4c indicates GA at least converged with one of the locally optimal solutions. Second, multiple clusters are detected. There are five detected clusters in Fig. 4b. Third, proposed method detects clusters that consist of non-adjacent meshes. For example, the detected cluster around Tokyo and Akihabara station has several non-adjacent meshes.



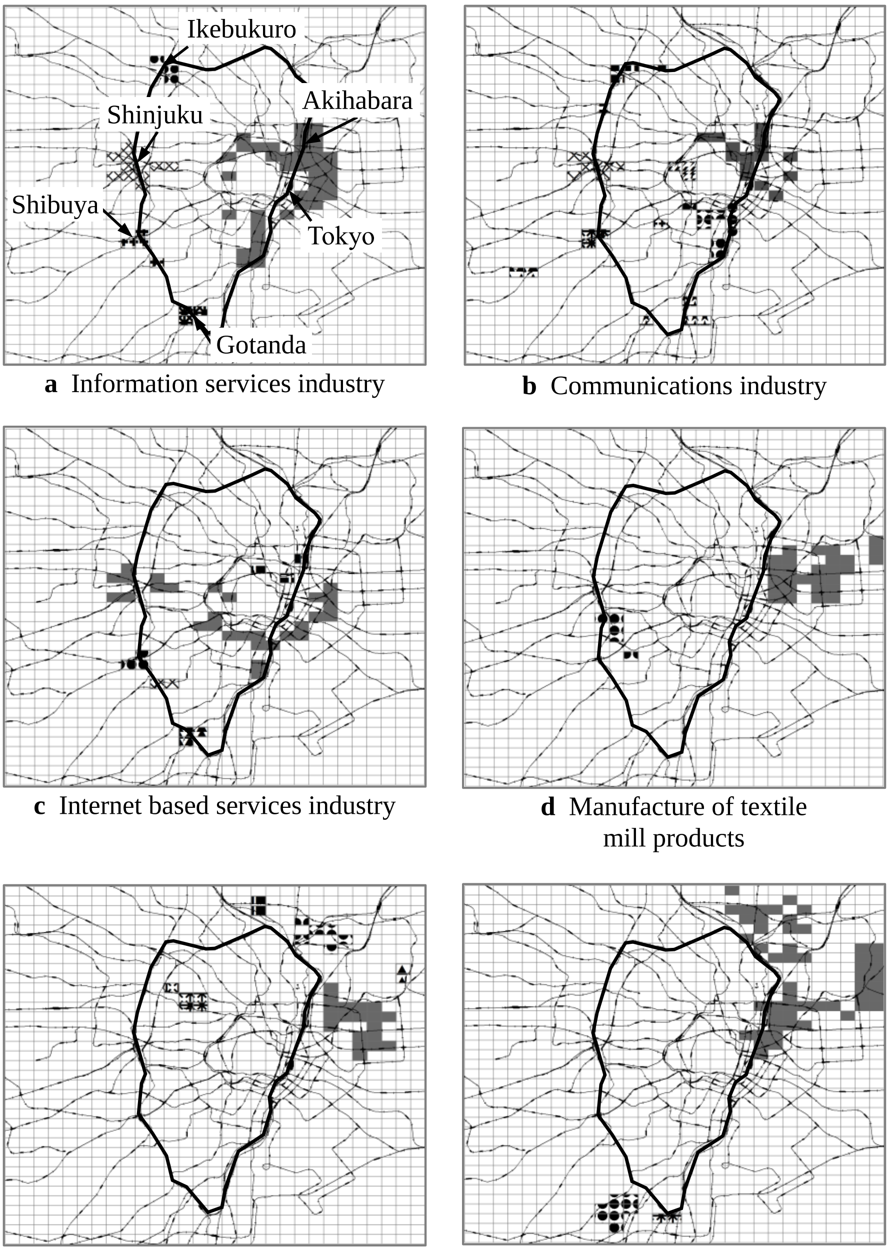
a Density of establishments **b** Detected clusters **c** Convergence process of GA

Fig. 4. Clusters of information services industry

4.2 Industrial agglomeration of six industries

Fig. 5 shows that results of cluster detection for six industries including information services industry. Figs. 5a-c and Figs 5d-f show results of industries classified into tertiary industry and secondary manufacturing, respectively. The heavy line indicates Yamanote Line, which is a belt line railway that connects most of Tokyo's major stations.

We point out following three points as to Fig. 5. First, we can reaffirm the fact that our method detects clusters that consist of non-adjacent meshes. Such clusters are detected in all six industries. Second, as in Figs. 5a-c, three industries classified into tertiary industry have in common that most clusters locate at a nearby site of major stations. Shinjuku and Shibuya stations clearly lie at the heart of detected clusters in all three tertiary industries. Clusters are also detected around Ikebukuro, Akihabara, Tokyo and Gotanda stations in one or two industries. Third, as in Fig. 5d-f, most detected clusters for industries classified into secondary manufacturing lie outside of the ring of the Yamanote Line; all three industries have a cluster located at east side of the ring of the Yamanote Line. However, there are several differences among industries of secondary manufacturing. For example, Manufacture of plastic products except otherwise classified (Fig. 5f) has clusters located at south side of the ring of the Yamanote Line, on the other hand other two industries don’t have such clusters. These results indicate that industries of secondary manufacturing have wider variety of location pattern than tertiary industries. More detailed comprehensive analysis is an issue in the future.



a Information services industry

b Communications industry

c Internet based services industry

d Manufacture of textile mill products

e Manufacture of pulp, paper and paper products

f Manufacture of plastic products, except otherwise classified

Fig. 5. Clusters of six industries

5 Concluding remarks

In this study, employing model-based clustering approach, we proposed a new cluster detection method that relaxes the constraints on geometric shape of spatial clusters. Relaxed constraints on geometric shape were expressed as two penalties added into energy function of Potts model which is employed on the prior distribution. We applied proposed method on mesh data of 2009 Japanese economic census. The result of case study shows that proposed method can detect clusters that consist of non-adjacent geographical units; our method can make clear the location of industrial agglomeration using detailed data.

Proposed method is characterized by Potts model and the two penalties, or distance penalty and complication penalty. The framework of our method can be used to detect clusters constraining geometric shape in the same way as existing cluster detection methods. For example, if we set up the maximal distance that is not penalized by distance penalty (denoted by l_d) as distance between adjacent geographical units, our method detect clusters assuming exact adjacency of geographical units. If we add one more penalty requiring circular clusters, our method detect such clusters. However, it should be noted that the fact remains that analysts face the problem of absence of objective criterion for setting constraints on geometric shape of spatial clusters.

Proposed method has two problems to be solved. First, we need to speed up calculation and analyze larger study area including an entire economic zone. Though we have already employed several devisals, GA still takes approximately 30 min. on average to detect clusters from data whose size is the same as case studies of this paper. Main cause of this problem lies in repeating calculation of minimum spanning tree and convex hull for every cluster. We are going to take in a devisal that reduces the number of calculating of minimum spanning trees and convex hulls.

Second, the parameter α of original Potts model should be estimated from data for the purpose of reducing the impact of arbitrary setting made by analysts.

Acknowledgment

This work is the production of collaborative research with Center for Spatial Information Science, The University of Tokyo, which provided mesh data of 2009 Japanese economic census.

References

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. In M. Painho (Ed.), *New tools for spatial analysis*. Luxembourg: Eurostat.

Baudry, J.-P., Cardoso, M., Celeux, G., Amorim, M. J. and Ferreira, A. S. (2014). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*. doi:10.1007/s11634-014-0177-3.

Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154(1), 143-155.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719-725.

Celeux, G., Forbes, F. & Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, 36, 131-144.

Chen, F., Tanaka, K. & Horiguchi, T. (2005). Image segmenta-tion based on bathe approximation for Gaussian mixture model. *Interdisciplinary Information Sciences*, 11(1), 17-29.

Cheng, H. D., Jiang, X. H., Sun, Y. & Wang, J. (2001). Color image segmentation: advances and prospects. *Pattern Recognition*, 34, 2259-2281.

Cucala, L. & Marin, J.-M. (2013). Bayesian inference on a mix-ture model with spatial dependence. *Journal of Computa-tional and Graphical Statistics*, 22(3), 584-597.

Duczmal, L. & Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45, 269-286.

Inoue, R., Kasuya, S. & Watanabe ,T. (2013). Spatio-temporal cluster detection of point events by hierarchical search of adjacent area unit combinations. *Proceedings of 13th International Conference on Computers in Urban Planning and Urban Management*, Paper 51, USB memory.

Kulldorff, M. (1997). A spatial scan statistic. *Communication Statistic Theory and Method*, 26(6), 1481-1496.

Kulldorff, M., Huang, L., Pickle, L. & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25, 3929-3943.

Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters – detection and influence. *Statistic in Medicine*, 14, 799-810.

Mori, T. & Smith, T. E. (2013). A probabilistic modeling ap-proach to the detection of industrial agglomerations. *Journal of Economic Geography*, 1-42.

Tango, T. & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(11).