Extraction of Industry Coagglomeration Patterns from Small Area Statistics: An Approach by the FDR-Based Cluster Detection and the Frequent Pattern Mining of Industry Clusters

Ryo Inoue

Abstract

Identifying the way in which the enterprises of related businesses locates in proximity occur – the coagglomeration of industries – is a key solution to understand the geographic concentration mechanisms of industries. Many research in geography and spatial economics have been addressing this issue to explain and theorize such mechanism. This paper proposes a new statistical approach that allows extracting coagglomeration patterns of industries from available national datasets. The approach takes two steps; the first step finds spatial clusters of each industry using the false discovery rate-controlling statistical test, and the second step searches for colocation relationships among industries through the frequent pattern mining of detected cluster location and the Monte Carlo simulation. This approach identifies coagglomeration patterns of industries. The proposed new method is applied to the 500-meter grid data of the 2012 Economic Census for Business Frame of Japan, to check its applicability and validity.

R. Inoue (Corresponding author)
Graduate School of Information Sciences, Tohoku University, 6-6-06, Aramaki Aoba, Aoba, Sendai, Miyagi 980-8579 Japan
Email: rinoue@plan.civil.tohoku.ac.jp

1. Introduction

The coagglomeration of industries, namely the phenomenon wherein offices and factories of related businesses are located near one another, is one of the key issues affecting the mechanisms of geographic concentration of industries and formation of urban areas. Many researchers in geography and spatial economics have developed various theories to explain these mechanisms, and conducted empirical studies to understand the actual condition of this phenomenon (e.g., Ellison and Glaeser, 1999; Duranton and Overman, 2005; Ellison et al., 2010).

Previous empirical studies proposed several indices that quantify the levels of industry coagglomeration; these indices help researchers analyze whether a group of industries accumulates in the same regions. However, they do not provide locational information where industry coagglomeration occurs, and they do not shed light on the combinations of industries which comprise industry coagglomeration.

Using small area statistics on economic activities, this study aims to discover coagglomeration patterns of industries and obtain information that will help pinpoint specific types of neighboring industries at particular locations. This study proposes an approach composed of two steps: spatial cluster detection of each industry and pattern mining of colocated industrial clusters.

The first procedure finds spatial clusters of each industry based on the false discovery rate (FDR)-controlling statistical test, which can avoid multiple testing problems. Caldas de Castro et al. (2006) first introduced the FDR-controlling statistical test to the geographic context, and Brunsdon and Charlton (2011) applied it to the spatial cluster detection problem. FDR-based spatial cluster detection is a powerful statistical method, since it can detect multiple clusters without multiple testing problems.

The second procedure searches for the colocation relationship between industries from cluster locations of each industry to discover of coagglomeration patterns of industries. The frequent pattern (FP)-growth algorithm (Han et al., 2000), one of the fastest frequent pattern mining algorithms, is applied to find the possible patterns of agglomerated industries in the same region, and the significance of found patterns are tested by the Monte Carlo simulation.

The applicability of proposed approach is tested on 500-meter grid square data of the 2012 Economic Census for Business Frame of Japan.

2. Previous Empirical Metrics of Agglomeration and Coagglomeration of Industries

Ellison et al. (2010) introduced two indices quantifying the degrees of co-agglomeration between industries: the Ellison and Glaser (1997) indices (hereinafter, referred to as the EG indices) and the Duranton and Overman (2005) index (hereinafter, referred to as the DO indices).

The EG indices use employment statistics aggregated by geographic regions such as states and counties. The EG index that tests the agglomeration of industry *i* is

$$\gamma_i \equiv \frac{\sum_{m=1}^{M}\left(s_{mi}-x_m\right)^2 \Big/ \left(1-\sum_{m=1}^{M} x_m^2\right) - H_i}{1-H_i} \tag{2.1}$$

where *m* indexes geographic regions, s_{mi} is the share of industry *i*'s employment in region *m* over industry *i*'s employment in all regions, x_m is the share of all employment in region *m* over all employment in all regions, and H_i is the enterprise-level Herfindahl index of industry *i*. The index value is large when industry *i* clusters. The agglomeration index is expanded to a coagglomeration index of a group of *I* industries,

$$\gamma_I^c \equiv \frac{\sum_{m=1}^{M}\left(s_m-x_m\right)^2 \Big/ \left(1-\sum_{m=1}^{M} x_m^2\right) - H - \sum_{i\in I}\hat{\gamma}_i w_i^2\left(1-H_i\right)}{1-\sum_{i\in I} w_i^2} \tag{2.2}$$

where w_i is the industry *i*'s share of the total employment in the group of *I* industries, s_m is the share of employment in the group of *I* industries in each region: $s_m = \sum_{i\in I} s_{mi}$, H is the enterprise-level Herfindahl index of the group of *I* industries: $H = \sum_{i\in I} w_i^2 H_i$, and $\hat{\gamma}_i$ is the estimated EG index of industry *i*'s agglomeration.

A simpler formula of EG index for the industry *i*'s agglomeration is also defined as

$$\gamma_i = \frac{\sum_{m=1}^{M}\left(s_m-x_m\right)^2}{1-\sum_{m=1}^{M} x_m^2} \tag{2.3}$$

and a simpler formula of EG index for the coagglomeration of industries *i* and *j* as

$$\gamma_{ij}^{c} = \frac{\sum_{m=1}^{M}\left(s_{mi} - x_m\right)\left(s_{mj} - x_m\right)}{1 - \sum_{m=1}^{M} x_m^2} \tag{2.4}$$

The EG indices have advantages in that they can utilize the statistics aggregated by regions, which are easy to obtain, to check the industry agglomeration and coagglomeration. However, Duranton and Overman (2005) criticized that the EG indices disregard the agglomeration and coagglomeration of industries composed of pairs of enterprises located near each other but not in the same region; the EG indices have the modifiable area unit problem (MAUP). Then, they proposed the distance-based DO indices. The DO index of agglomeration of industry *i* is

$$\hat{K}_A^{Ent}(d) = \frac{1}{n(n-1)h}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} f\left(\frac{d - d_{ij}}{h}\right) \tag{2.5}$$

where n is the number of enterprises in the industry, d_{ij} is the Euclidean distance between enterprises i and j, and $f(\,)$ is a Gaussian kernel density function with bandwidth h. This index takes into account all distances between pairs of enterprises. If the number of employees at each enterprise is known, the agglomeration index can be represented as

$$\hat{K}_A^{Emp}(d) = \frac{1}{h\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} e(i)e(j) f\left(\frac{d - d_{ij}}{h}\right) \tag{2.6}$$

where $e(i)$ is the number of employees of enterprise i. These DO indices are expanded to analyze the coagglomeration of industries, such that

$$\hat{K}_{ij}^{c\ Ent}(d) = \frac{1}{n_i n_j h}\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} f\left(\frac{d - d_{rs}}{h}\right) \tag{2.7}$$

$$\hat{K}_{ij}^{c\ Emp}(d) = \frac{1}{h\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} e(r)e(s)}\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} e(r)e(s) f\left(\frac{d - d_{rs}}{h}\right) \tag{2.8}$$

where n_i and n_j are the number of enterprises in industries i and j, respectively.

It is worth mentioning that the DO-indices share a similarity with Ripley's *K* function (Ripley, 1976), namely the analysis of spatial point distribution. The definition of the *K* function at distance *d*, *K*(*d*), is the expected number of extra points within distance *d* of a randomly chosen point over the point density:

$$\hat{K}(d) = \frac{1}{\hat{\lambda}\pi d^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I\left(d_{ij} \le d\right) \tag{2.9}$$

where $\hat{\lambda}$ is the estimated density, which is the total number of points over the total size of the study area, and *I*( ) is the indicator function. The cross-*K* function $K_{ij}(d)$, namely the expansion of the *K* function to the two point processes, is defined as the expected number of type *j* points within distance *d* of a randomly chosen type *i* point. The estimator of the cross-*K* function is

$$\hat{K}_{ij}(d) = \frac{1}{\hat{\lambda}_i \hat{\lambda}_j \pi d^2} \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} I\left(d_{rs} \le d\right) \tag{2.10}$$

where $\hat{\lambda}_i$ is the estimated density of point process *i*, which is the total number of type *i* points over the total size of the study area (e.g. Cressie, 1993).

The DO indices and *K* functions utilize the detail location information of enterprises, which is hard to obtain; they are free from the MAUP. However, they have difficulties in analyzing the coagglomeration which consists of more than two industries; it is hard to define the proximity of more than two points, to search the combinations of more than two points which are located in their neighborhood, and to understand the statistical properties of indices which is composed of more than three different point processes.

These EG and DO indices have been applied to many datasets in several countries to analyze the agglomeration and coagglomeration of industries. However, they suffer two major limitations of analysis, as described below.

The first limitation is that they cannot indicate the regions where industry coagglomeration occurs. The previous indices can judge whether industries cluster in their neighborhoods, but do not provide locational information on industry coagglomeration.

The second limitation is that it is difficult to identify the patterns of industries which have coagglomeration. The EG indices needs to preset a combination of industries to calculate an index value. Since the possible patterns of industries, composed of more than two industries, are numerous, discovering coagglomeration patterns of industries is almost impossible without an effective search algorithm.

This study takes the following two-step approach to overcome limitations by the previous methods: i) spatial cluster detection of each industry using the false discovery rate (FDR)-controlling method, and ii) pattern mining of colocated industrial clusters using the frequent pattern mining algorithm. The significance of discovered patterns of industries are tested by the Monte Carlo simulation. This approach therefore can discover patterns of industries that have similarities in their spatial distributions as well as the regions where such industrial patterns are located.

3. Spatial Cluster Detection of Each Industry Using the FDR-Controlling Method

3.1 Two Major Methods for Multiple Spatial Cluster Detection of Point Events

This study considers the locations of enterprises as point event data and searches regions wherein enterprises accumulate using the statistics aggregated by regions. When detecting regions of industry agglomeration, it is general to assume that multiple cluster regions exist. Two approaches are proposed recently for detection of multiple clusters of point events; Mori and Smith (2010) offered a method as an extension of spatial scan statistic (e.g., Kulldorff, 1997), and Brunsdon and Charlton (2011) proposed the application of the FDR-controlling procedure (Benjamini and Hochberg, 1995).

The former method bases the spatial scan statistic. The spatial scan statistic is a famous method to detect regions where point events cluster; it evaluates the degree of point accumulation in the given area by the likelihood ratio, namely the likelihood of the alternative hypothesis that assumes the given area is a cluster region in which the large number of point events are located over the likelihood of the null hypothesis that assumes the area is not a cluster region. After searching the area with the maximum likelihood ratio, its statistical significance is tested through comparison to the distribution of the maximum likelihood ratio from the random point distribution obtained from the Monte Carlo simulation. The spatial scan statistic and its derivations are widely used for cluster detection; however, there is a limitation for detecting multiple clusters, since the alternative hypothesis presumes that a single cluster exists. Although the secondary and other clusters can be detected under the condition that the former detected clusters exist at the detected locations, this limitation spoils the availability of spatial scan statistic-based cluster detection.

Mori and Smith (2010) proposed the evaluation of multiple cluster models as an expansion of the spatial scan statistic using the Bayesian Information Criterion (BIC). This method forms "cluster schemes" that set the multiple cluster candidates, estimates the density parameters for all candidates in each cluster scheme based on the point distribution assumption, and calculates the BICs. After the cluster scheme with the maximum BIC is selected, its significance is tested through the Monte Carlo simulation. Since model selection by the BIC can consider the number of clusters as well as their locations, it is a promising multiple cluster detection method. However, since the numbers of possible cluster schemes are huge and the efficient search procedure is not proposed, this method might take excessive time for detecting clusters, especially in small area analysis.

The latter method is based on the FDR-controlling procedure; it can not only avoid the multiple testing problem but also achieves greater statistical power than family-wise error rate controlling methods, that is, another approach for multiple testing (e.g. Holm, 1979).

The multiple testing increases the occurrence of false discoveries (type I errors) by chance. Benjamini and Hochberg (1995) defined the FDR as the expected value of the proportion of false discoveries to the rejected null hypotheses, and proposed a testing procedure that keeps the FDR less than the given significance level α. Brunsdon and Charlton (2011) utilized it for cluster detection; the method configures the set of alternative hypotheses that each region is a cluster, and tests the null hypotheses according to the FDR-controlling procedure.

The latter method is advantageous in that it requires far less calculation amount compared to the former method to find cluster regions. This study employs the FDR-controlling method to detect regions with each industry's agglomeration.

3.2 Detection of Industrial Clusters Using the FDR-Controlling Method

Let G denote the area of interest, and suppose G is segmented into subregions. Let Z denote one of the subregions in G, Z^C denote a complement region of Z in G, n_Z denote the count of point events in Z, n_{Z^C} denote the count of point events in Z^C, a_Z denote the size of Z, and a_{Z^C} denote the size of Z^C. The sizes of regions could be defined by their respective areas, or the number of enterprises of all industries in the regions. Here, assume that the

spatial distributions of points in Z and Z^C conform to the Poisson distributions in which the counts of points in each region are proportional to the sizes of the regions. Then,

$$n_Z \sim Poisson\left(\lambda_Z a_Z\right),\ n_{Z^C} \sim Poisson\left(\lambda_{Z^C} a_{Z^C}\right) \tag{3.1}$$

where λ_Z and λ_{Z^C} are the density parameters in Z and Z^C, respectively. The alternative hypothesis, which considers that points are clustered in Z, is

$$H_1 : \lambda_Z > \lambda_{Z^C} \tag{3.2}$$

and its null hypothesis is

$$H_0 : \lambda_Z = \lambda_{Z^C} . \tag{3.3}$$

Now suppose that the observed number of points in G is N. Under the condition that N points are located in G, n_Z conforms to the following binomial distribution

$$n_Z \sim Bi\left(N, \frac{\lambda_Z a_Z}{\lambda_Z a_Z + \lambda_{Z^C} a_{Z^C}} \right) \tag{3.4}$$

If the null hypothesis is true,

$$n_Z \sim Bi\left(N, \frac{a_Z}{a_Z + a_{Z^C}} \right). \tag{3.5}$$

Then, when the observed point count in Z is n, the p-value of the null hypothesis for Z, p_z, is

$$p_Z = \sum_{i=n}^{N} \binom{N}{i} \left(\frac{a_Z}{a_Z + a_{Z^C}} \right)^i \left(\frac{a_{Z^C}}{a_Z + a_{Z^C}} \right)^{N-i} . \tag{3.6}$$

After the p-values of the null hypotheses for all subregions are calculated, these hypotheses are tested by the FDR-controlling procedure proposed by Benjamini and Hochberg (1995).

4. Discovery of Colocated Patterns of Industrial Clusters

A frequent pattern mining algorithm first extracts possible industrial patterns which might constitute coagglomeration. It was first proposed by Agrawal and Srikant (1994), and is commonly applied in the analysis of consumer buying behavior, to understand which combinations of items are bought together. It distinguishes a frequent combination by *support*, that is, the proportion of the number of regions where a combination of industries is colocated to the total number of regions in the study area. The s*upport* can be interpreted as an estimate of occurrence probability of a combination. The combinations whose frequencies are more than or equal to the given threshold are extracted as the frequent patterns.

Suppose that Table 1 represents the cluster detection results. Clusters of industries A to D, are located in Regions I to V. The *support* of industry A is 60%, as clusters of industry A are located in three regions, namely regions I, III, and V, whereas the total number of regions is five. Similarly, the *support* of pattern {A, B, C} is 40%, as it is found in regions I and V. When the threshold is set to 40%, eight patterns, {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, {A, D}, {A, B, C}, are extracted from the results of cluster detection of each industry.

Table 1. Example of spatial distribution of clusters

Regions	Clustered industries
I	A, D
II	B
III	A, B, C
IV	ϕ
V	A, B, C, D

Usually a common threshold value is set in a pattern mining analysis; however, the common threshold might be too large to find patterns that include industries with few clustered regions. In the coagglomeration analysis, it is meaningful to find the patterns of industries which are located in the limited regions but colocated in most of those regions. Thus this study sets not only the common threshold but also the threshold on the basis of the industry whose clustered regions are the fewest in the pattern.

When the numbers of industries and regions are large, the frequent pattern mining becomes a time consuming process. Several algorithms have been proposed to overcome this problem; this study utilizes the FP-growth algorithm, one of the fastest pattern mining algorithms (Han et al. (2000)).

After the possible industrial patterns are extracted, their significance is tested by the Monte Carlo simulation. If all regions have a uniform cluster occurrence probability of each industry, the numbers of cluster-colocated regions conform to the Poisson distribution under the null hypotheses that assume clusters of each industry are independently distributed. However, as most industries, especially service industries, tend to be located in urban areas where many people live, it is inappropriate to assume the same cluster occurrence probability to regions; this study sets the cluster occurrence probability of an industry in each region to be proportional to the number of clustered industries in each region. This setting makes difficult to estimate the distributions of numbers of cluster-colocated regions under the null hypotheses without simulation; this study employs the Monte Carlo simulation. A cluster occurrence of each industry at each region is simulated under the given probability, regions where simulated clusters are colocated are counted for each possible extracted industrial pattern, and their counts by simulation are compared with those by the data to test the significance of industrial patterns.

5. Application

5.1 Dataset

The 2009 Economic Census for Business Frame is a dataset covering all establishments and enterprises in Japan as of July 1, 2009. This study uses the 500-meter grid square statistics which record the number of enterprises in each grid square. The industries are classified into 86 major groups as per the Japan Standard Industrial Classification, revised on November 12, 2007 (Ministry of Internal Affairs and Communications, 2007). The total number of 500-meter grid squares is 1,515,129 in the entire nation, and the statistics contain the records only on 336,646 grid squares wherein at least one enterprise is located; it is important to note that this is a zero-truncated dataset. The total establishment numbers are 6,009,389.

5.2 Detection of Clusters of Each Industry

The cluster detection of each industry was carried out under the condition that the upper limit of FDR is 0.01. The cluster detections are conducted based on the area density of enterprises. The density estimators are corrected considering that the statistics are zero-truncated data (e.g. Cohen, 1960).

There are 26,184 grid squares containing at least one industrial cluster, amounting to 1.7% of the total grid squares in Japan and 7.8% of industry-located grid squares. Table 2 shows the top ten industries with the widest clustered areas, and Fig. 1 and 2 show examples of cluster detection results near Tokyo. The figures clearly show that the retail industries are located in central business districts and along railway lines (black thick lines), while the manufacturing industries show different spatial distributions.

Most of the results look adequate; however, the clusters of industry "49. Postal activities, including mail delivery," which consists of delivery stations, have an issue. The enterprises of postal activities are usually located at intervals to provide nationwide services; only two enterprises at maximum are located in one grid square. The estimated density parameter is very small, 0.022 enterprises per square kilometer; that is, 0.0056 enterprises per grid square. As a result, all square grids where at least one enterprise is located are selected as clustered areas. The Poisson distribution is not an appropriate model for the distribution of this industry and the 500-meter grid squares are too small compared to the number of enterprises. This is related to the MAUP; it is difficult to decide the appropriate size of spatial units used in the analysis.

Table 2. Top ten industries with broad clustered areas

Industrial categories		Number of grid squares detected as clusters
60	Miscellaneous retail trade	7,569
76	Eating and drinking places	7,563
78	Laundry, beauty, and bath services	7,431
58	Retail trade (food and beverage)	7,116
69	Real estate lessors and managers	5,452
83	Medical and other health services	4,423
49	Postal activities, including mail delivery	3,682
7	Construction work by specialist contractor, except equipment installation work	2,545
57	Retail trade (dry goods, apparel, and apparel accessories)	2,545
6	Construction work (general), including public and private construction work	2,441

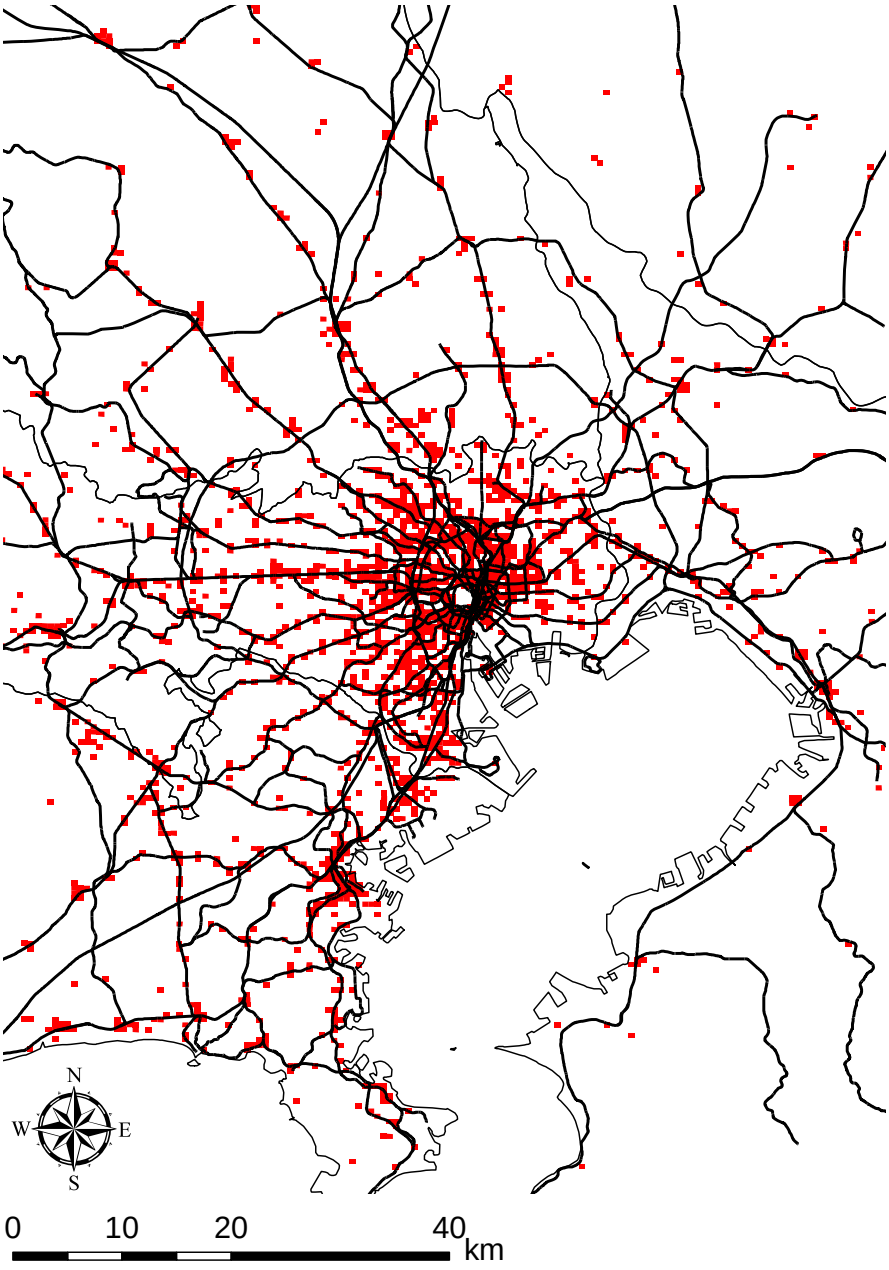

Fig. 1. Clusters of “60. Miscellaneous retail trade”

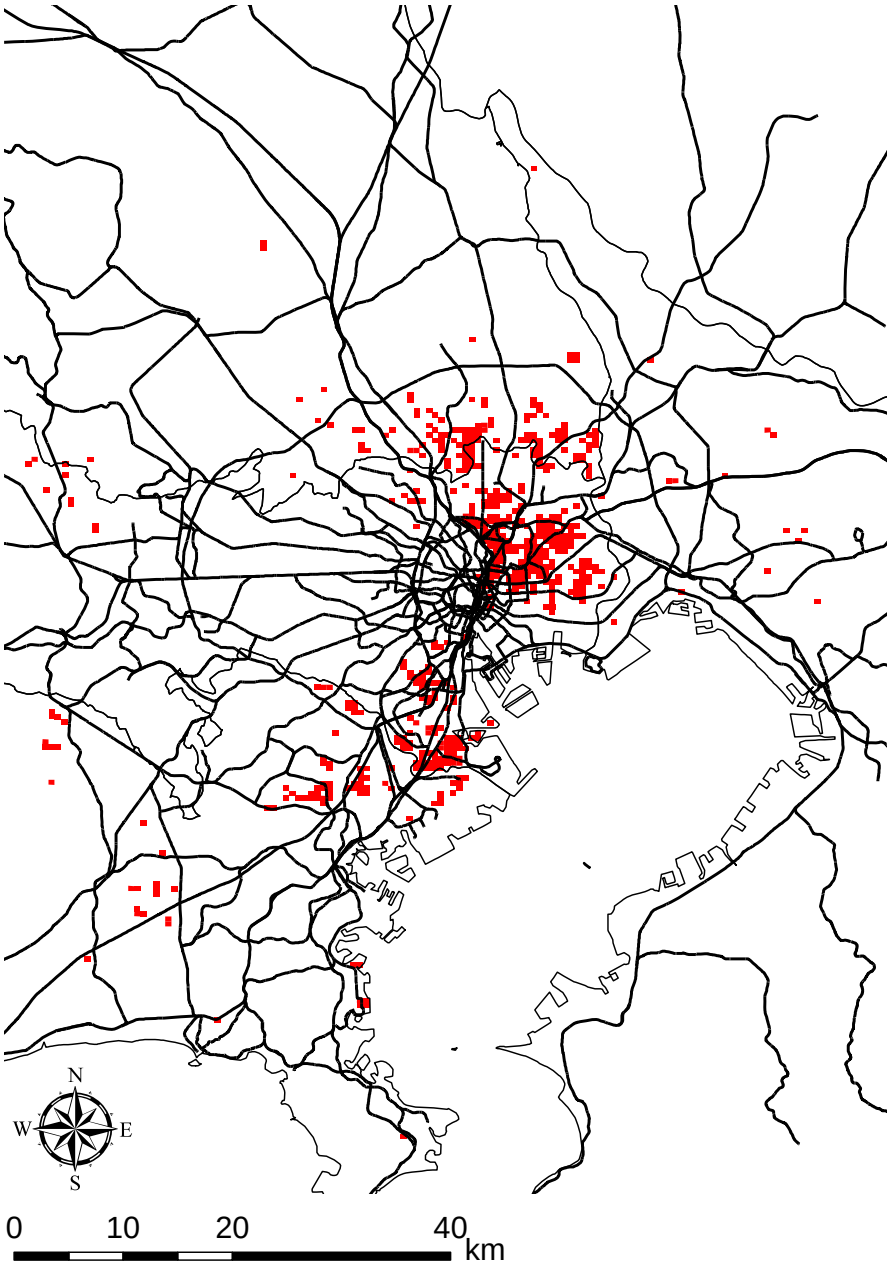

Fig. 2. Clusters of "24. Manufacture of fabricated metal products"

5.3 Extraction of Coagglomeration Patterns of Industries

The thresholds for possible pattern extraction are set to the 0.25% of grid squares where at least one industry accumulates as the common threshold and the 1% of clusters of industry whose clusters are the fewest in the pattern. The frequent pattern mining algorithm extracted 137,799 patterns, and 3,338 patterns that are not a subset of others.

Then the Monte Carlo simulation tests the significance of extracted patterns; the simulation is repeated 999 times, and the significance of patterns is decided under the condition that the upper limit of FDR is 0.01, as this process causes also multiple testing problem. The simulation rejects the patterns whose subsets are non-significant patterns; 22,787 patterns are judged as significant, 1,342 patterns not a subset of others.

The patterns with the largest number of industries are shown in Table 3. Checked industry categories indicate industries included in the patterns. Most of the industries in the larger patterns are service industries, and the grid squares show that these industries are colocated in the central business district of Japan. Fig 3 shows the location of pattern #1.

The analysis reveals that 45 industries are not colocated with any other industries. Most of them have less than five hundred cluster grid squares; however, “49. Postal activities, including mail delivery”, “75. Accommodations”, and “11. Manufacture of textile mill products” are have more than a thousand clustered grid squares but does not have any coagglomeration.

Table 3. Industries identified in the largest patterns

Patterns		#1	#2	#3
39	Information services			✓
57	Retail trade (dry goods, apparel, and apparel accessories)	✓	✓	
58	Retail trade (food and beverage)	✓	✓	✓
60	Miscellaneous retail trade	✓	✓	✓
68	Real estate agencies	✓	✓	✓
72	Professional services, n.e.c.	✓	✓	✓
74	Technical services, n.e.c.		✓	✓
76	Eating and drinking places	✓	✓	
78	Laundry, beauty, and bath services	✓	✓	✓
79	Miscellaneous living-related and personal services	✓	✓	✓
82	Miscellaneous education, learning support	✓	✓	✓
83	Medical and other health services	✓	✓	✓
92	Miscellaneous business services	✓		✓
Number of observed grid squares		247	225	220

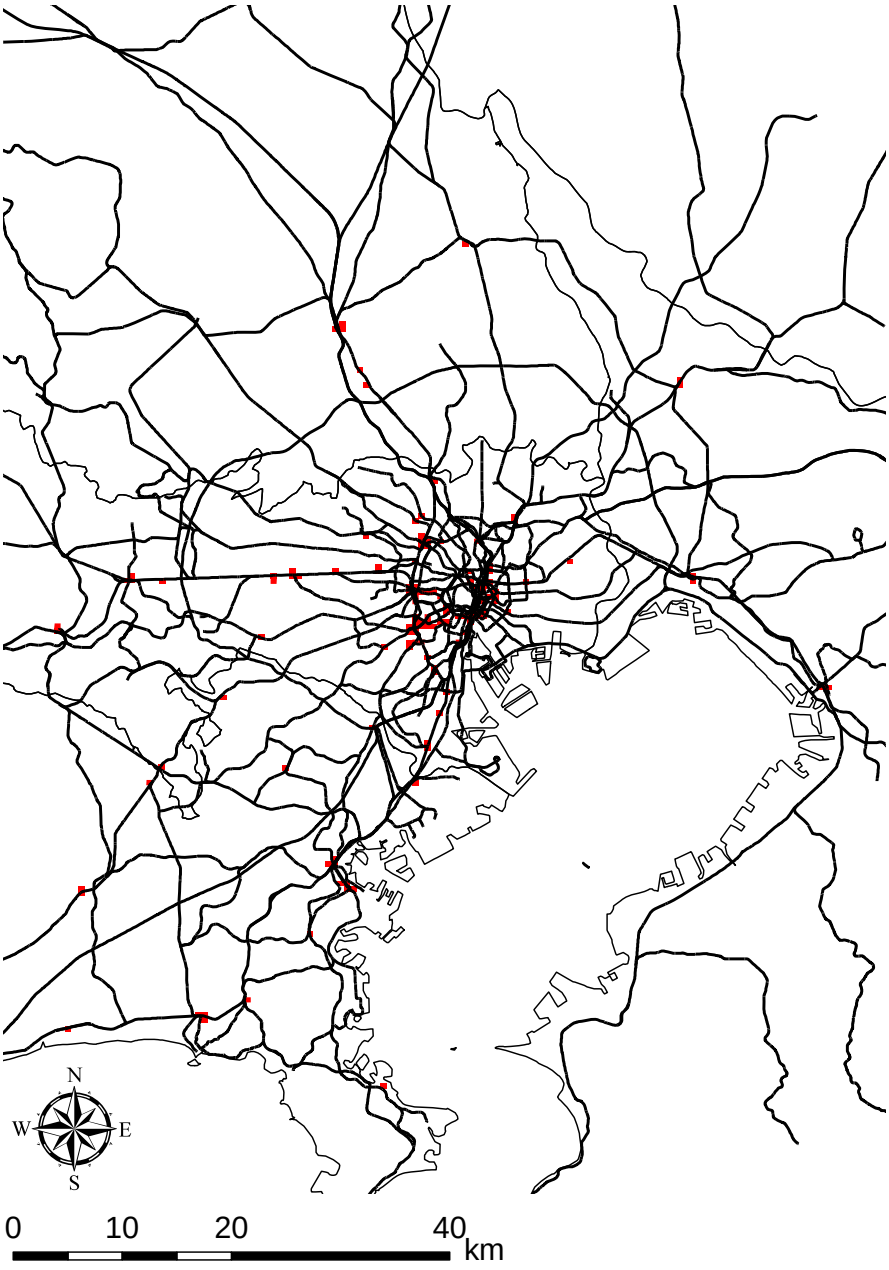

Fig. 3. Location of coagglomeration pattern #1

5.4 Discussion

The advantages of the proposed approach are as follows:

1. It can indicate regions where industry coagglomeration is observed.
2. Each industrial clusters are detected with the FDR-controlling statistical test, which can detect multiple clusters by avoiding multiple testing problems.
3. The possible coagglomeration patterns of industries are discovered by the data mining algorithm; since the algorithm has scalability, it is applicable to the analysis even when the numbers of regions and industry classifications are large.
4. The significance of coagglomeration patterns of industries is tested by the Monte Carlo simulation. The simulation considers that the industries tend to be located in urban areas, and finds the significant patterns which are not occur by chance.

The proposed approach has some disadvantages and possible aspects to improve:

1. The results of the proposed approach reveal a significantly large number of patterns of industries; it is hard to interpret the results, as 1,342 industrial patterns are extracted. It is, thus, necessary to summarize and visualize the relationships of industries extracted from the analysis of the locations of each industry.
2. The proposed approach remains true to the principles of the agglomeration and coagglomeration indices proposed/used by previous studies. However, the output of proposed approach does not give any information about the manner in which the coagglomeration of industries occurs. It only reveals that certain industrial patterns are colocated. Further analyses are needed to better understand the mechanism of coagglomeration of industries.
3. As is similar to the Ellison and Glaser (1997) indices, the proposed approach uses statistics aggregated by geographic regions. It has the MAUP; the results will change if the regions used for aggregation are different.

6. Concluding Remarks

This study proposed an approach to discover coagglomeration patterns of industries through point event cluster detection and pattern mining of colocated industrial clusters. Compared to previous indices, the proposed approach is advantageous in that it is able to discover the patterns of industry coagglomeration and identify locations of industry coagglomeration. The

extracted industry coagglomeration patterns reveal the relatedness of industries, and will also help classify regions based on patterns of located industries; it might be a useful information that helps policymakers understand the characteristics of regions and develop effective policies to attract businesses to their regions.

However, as stated in Section 5.4, the proposed approach has some drawbacks, which we intend to explore on in the future research.

Acknowledgement

This study was supported by JSPS KAKENHI Grant Number 26289169. The 500-meter grid square statistics of the 2009 Economic Census for Business Frame provided by Sinfonica is used as the CSIS Joint Research (No. 456) using spatial data provided by Center for Spatial Information Science, The University of Tokyo. The author thanks Kohei Shiga for his research support.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*, 487–499.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57(1), 289–300.

Brunsdon, C., & Charlton, M. (2011). An assessment of the effectiveness of multiple hypothesis testing for geographical anomaly detection. *Environment and Planning B: Planning and Design*, 38, 216–230.

Caldas de Castro, M., & Singer, B. H. (2006). Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis*, 38, 180–208.

Cohen, A. C., Jr. (1960). Estimating the parameter in a conditional Poisson distribution. *Biometrics*, 16(2), 203–211.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. New York: Wiley.

Duczmal, L., Kulldorff, M., & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, 15(2), 1–15.

Duranton, G., & Overman, H. G. (2005). Testing for localization using micro-geographic data. *The Review of Economic Studies*, 72(4), 1077–1106.

Ellison, G., & Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy*, 105(5), 889–927.

Ellison, G., & Glaeser, E. L. (1999). The geographic concentration of industry: Does natural advantage explain agglomeration? *The American Economic Review*, 89(2), 311–316.

Ellison, G., Glaeser, E. L., & Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *The American Economic Review*, 100, 1195–1213.

Han, J., Pei, H., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. ACM Press, New York, NY, USA.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

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

Ministry of Internal Affairs and Communications of Japan. (2007). *Japan Standard Industrial Classification* (Rev. 12, November 2007).

Mori, T., & Smith, T. (2010). A probabilistic modeling approach to the detection of industrial agglomeration. *KIER Discussion Paper*, 777, 1–54.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2), 255–266.