Inoue *et al. Computational Urban Science* (2023) 3:35
https://doi.org/10.1007/s43762-023-00107-9

Computational Urban Science

ORIGINAL PAPER

Open Access

Colocations of spatial clusters among different industries

Ryo Inoue[1], Shino Shiode[2] and Narushige Shiode[3*]

Abstract

Spatial colocation has been studied in many contexts including locations of urban facilities, industry entities and businesses. However, identifying colocations among a small number of facilities and establishments holds the risk of introducing false positive in that such a spatial arrangement may have occurred by chance. To account for the association between a group of facilities that frequently colocate with each other, this study proposes a two-step approach consisting of identifying statistically significant clusters of each facility type using the False Discovery Rate (FDR) controlling procedure, and subsequently measuring the colocation of those clusters with the frequent-pattern-growth (FP-growth) algorithm. Empirical analysis of 6 million business and industrial establishments across Japan suggests that 10 out of 86 industry types form clear colocations and their colocations form a multi-layered, cascading structure. The number of layers in the multi-layered structure reflect the city size and the strength of the association between the colocated clusters of industries. These patterns illustrate the utility of detecting colocation of clusters towards understanding the agglomeration of different businesses. The proposed method can be applied to other contexts that would benefit from investigations into how different types of spatial features can be linked with each other and how they form colocations.

Keywords Colocation, False discovery rate, Frequent-pattern growth, Industry agglomeration, Spatial clusters

1 Introduction

Detection of *spatial clusters* usually focuses on identifying concentrations of a single type of spatial feature. As the pursuit of *spatial clusters* is extended to account for more than one type of spatial features in the cluster, its notion becomes close to that of *spatial colocation*. Huang et al. (2004) define spatial colocation as "*subsets of (Boolean) spatial features whose instances are often located in close geographic proximity*" (Huang et al., 2004, p.1472). A typical example of colocation is symbiotic animal or plant species, where specific combination of species live together. A common framework of spatial colocation discovery methods was established in the early 2000s (Morimoto, 2001; Shekhar & Chawla, 2003; Shekhar & Huang, 2001; Yoo & Bow, 2012; Yoo & Shekhar, 2004), mainly in the field of computer science, and a range of algorithms for extracting spatial colocations have been developed since then.

In an urban context, there are many instances where a specific group of businesses and industries are located close to each other and form a spatial agglomeration pattern. A regular occurrence of the same combination of industries is often referred to as colocation and has long been studied in the domains of urban geography, economic geography and data science. However, current definition of colocation allows inclusion of any number of facilities for each industry. This could potentially result in a situation where one or two cases of an industry that is prevalent across the urban space (e.g. a corner news

*Correspondence:
Narushige Shiode
n.shiode@kingston.ac.uk
[1] Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan
[2] Department of Geography, University of London, Birkbeck WC1E 7HX, UK
[3] Department of Geography, Geology and the Environment, Kingston University, Kingston Upon Thames KT1 2EE, UK

stand or a drug store) may have been frequently discovered in close proximity of other types of facilities and, yet, it may be recognised as a case of colocation. Designing a framework to ensure the presence of a sufficient number of facilities from each industry will help us identify the spatial colocation among those industries and, thereby, eliminating instances where only few facilities exist in the same area as a cluster of another type of facility.

Methods for identifying clusters among multiple types of events or facilities have been also limited until recently, partly because of the conceptual and computational complexity involved in designing such a method (Huang et al., 2004; Shekhar & Huang, 2001). To establish a framework for extracting spatial colocations between sufficient number of events and facilities, this study proposes a combined approach between *spatial clustering* (for amassing statistically significant concentration of certain facilities) and *spatial colocation* (that confirms the frequency of the colocation patterns between a group of facilities), in the hope of identifying the colocated clusters of facilities.

2 Literature review

2.1 Spatial cluster detection

Knox (1989) provided a geometrically explicit definition of *spatial clusters* as "*a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance*" (Knox, 1989, p.17). The question of whether spatial features are clustered in space has drawn a strong attention in many fields of research including epidemiology, criminology, geology and ecology. Spatial representation of socio-economic, physical or ecological phenomena can often be reduced to that of simple spatial features, and the interpretation of the way in which these features are dispersed or clustered has been explored widely in those fields (Gatrell, 2002; Lawson, 2006).

Methodologically, a range of cluster detection methods have been developed so far. The existing range of cluster detection techniques can be broadly divided into two categories: (1) those that confirm the presence of clusters and measure the degree of clustered-ness, and (2) those that identify the spatial extent of clusters. In other words, methods in the first category examine the *global tendency to cluster*, whereas those in the second category exclusively search for *the location of local clusters* (Tango, 1999).

2.1.1 Global tendency of clusters

The first category of methods can be further divided into two sub-groups: aggregate, area-based methods and disaggregate, point-based methods. Aggregate methods use the feature counts or rates aggregated to either the same-sized area-based units called quadrats, or irregular spatial units such as administrative districts. The simplest form of areal-based methods is quadrat methods (Boots & Getis, 1988; Ripley, 1981) which apply statistical tests such as χ^2 test to interpret the frequency distribution of crime counts. Another popular strand of aggregate methods is those that use *spatial autocorrelation* indices at a global scale. A range of statistical indices, such as Moran's *I*, Geary's *C* and Getis-Ord *G* (Getis & Ord, 1992; Moran, 1948; Ord & Getis, 1995), have been designed to measure the level of spatial heterogeneity of an attribute value in the form of a correlation coefficient among neighbouring spatial units across the study area, thus providing a global tendency of a locally aggregated structure of the given distribution.

Disaggregate methods have also been well studied, especially in the field of *point pattern analysis* (PPA) (Boots & Getis, 1988; Diggle, 2003). PPA has been developed primarily for measuring the degree of spatial variation in the distribution of point objects, and also for investigating the presence of a statistical anomaly among them. A number of methods have been proposed for analysing such distributions, including distance-based methods that use the point-to-point distances as an index for measuring the degree of point dispersion. They range from the nearest-neighbour distance methods (Clark & Evans, 1954) and k-nearest neighbours (k-NN) method (Cuzick & Edwards, 1990) to Ripley's *K*-function method (Ripley, 1976, 1981). In epidemiology, hypothesis testing for the absence of spatial disease clustering is carried out usually with aggregate data, adjusting for an inhomogeneous background population. Methods that are commonly employed in this process include Besag-Newell's *R* statistic (Besag & Newell, 1991), the maximizing excess events test (Tango, 2000) and the Bonetti-Pagano M statistic (Bonetti & Pagano, 2005).

2.1.2 Local clusters

The second category of methods are designed to identify the location and the extent of clusters themselves at a local scale. These methods maintain the capacity to provide more detailed information on the spatial characteristics of a given distribution that would meet the demands for practical applications in a number of subject fields. A range of methods belong to this category including the following three types of techniques.

The first strand of methods is characterised by the concept of *local spatial autocorrelation*. As stated earlier, spatial autocorrelation statistics have initially been used for measuring the tendency of global clustering. However, these statistics were later extended to serve as an indicator of local clusters, from which a series of local statistical measures were developed. These include local Moran's *I (also called LISA – Local Indicators of*

Spatial Association), local Geary's C and local Getis-Ord G_i^* and G_i (Anselin, 1995; Getis & Ord, 1992; Ord & Getis, 1995), as well as their extension as AMOEBA, a method for finding clusters as flexibly combined neighbouring areas (Aldstadt & Getis, 2006). These statistics help find a set of adjacent areas with similar attribute values identified by a spatial correlation coefficient. It shares the same principle with the global indices discussed earlier, except that the spatial correlation coefficient is calculated for each specific location which allows us to extract local clusters.

The second group of local cluster methods uses the concept of a *search window* which was originally developed in spatial epidemiology (Tango, 1999). A search window usually takes the form of a circle and is used for sweeping exhaustively across the study area to find an area with high concentration of events. It allows us to count the number of individual observations within the search window at each instance and compare them to the expected number of counts under the null model to find any unusual concentration of events (e.g. Besag & Newell, 1991; Diggle & Chetwynd, 1991; Rushton & Lolonis, 1996; Turnbull et al., 1990). The spatial scan statistic (Kulldorff, 1997; Kulldorff & Nagarwalla, 1995) shares a similar concept, but it is much more widely used as it addresses most of the limitations from which the previous methods have suffered; namely it has (1) the capacity to detect cluster size continuously, rather than applying search windows of discrete sizes; and (2) offers a control for the multiple testing problem. More recently, the spatial scan statistic was extended to the space-time dimension (Kulldorff et al., 1998) and to different or more flexible shapes of search windows (e.g. Kulldorff, et al., 2006; Patil & Taillie, 2004; Shiode & Shiode, 2020, 2022; Takahashi et al., 2008; Tango & Takahashi, 2005).

The range of cluster-detection methods discussed so far focus on the concentration of a single, specific type of feature (e.g. patients of a particular disease; or a single species of plant). It is occasionally extended to address clusters of two types of features, but the range of methods for the pair-wise detection of two features are much limited than those for a single type of feature. One of those that can be extended to two features detection is the K-function discussed earlier. It has been extended in the form of cross-K-function, which investigates the proximity and association between two types of features in a cumulative fashion by increasing the search distance (Ripley, 1976). Similarly, Moran's I and its *local* variant, local Moran's I, are also capable of detecting concentrations of two types of features. These extensions are known as Bivariate Moran's I and Bivariate Local Moran's I (Bivariate LISA), and they have been applied in a variety of contexts (Anselin et al., 2002). However, they can be only used for measuring the association of the degree of concentration between two types of features.

2.2 Spatial colocation of clusters

Unlike the cluster-detection methods, spatial colocation methods are designed to extract a group of features that are repeatedly observed as a set across the study area, and there is usually no limit to the number of feature types they can extract as a unit of colocation. The extent of colocation is often represented in the form of an index that serves as a metric for quantifying the frequency of observing the same set of features. They can be considered as the colocation equivalent of the early phase in cluster detection (i.e. measuring the tendency to colocate and confirm the presence of colocated sets). For instance, in the field of economic geography, use of the employment statistics (or the number of employees) aggregated by the geographic regions (e.g. states and counties) could lead to *an industry agglomeration index*. It quantifies whether and to what extent a group of industries locate close to each other (Ellison et al., 2010; Ellison and Glaser, 1997; Duranton & Overman, 2005).

In addition to assessing the overall level of colocation in the form of an index, colocation also refers to the identification of the specific combination of feature types that frequently form a colocation. This process requires a different strand of methods from those that measure the overall intensity of colocation, and they are often categorised into two classes; namely, the *spatial statistics approach* and the *data mining approach* (Huang et al., 2004). Spatial statistics approach uses spatial correlation to characterise the relationship between different types of spatial features. Measures of spatial correlation (between pairwise events) include the cross-K function used in cluster detection and the Pair Correlation Function (PCF) (Illian et al., 2008). In this sense, *spatial correlation* methods can be regarded as dual-purpose methods catering to the needs for *spatial cluster detection* and *spatial colocation*. Similarly, use of spatial modelling approaches for investigating *spatial correlation* is also considered as part of this strand. For instance, using logistic regression, Tonkin et al. (2011) investigated cross-crime association between a pair of crimes using the inter-crime distance and the temporal proximity.

Finding colocation patterns across a large number of features can be computationally expensive, and it was only in recent years that data mining approaches saw much development—specifically, their application to regional-level and larger datasets were only made possible after high computational power became affordable and suitable data-mining methods were developed. These approaches mainly consist of the *area-based (aggregate) approaches* and the *distance-based (disaggregate)*

approaches. Here, we can draw parallels to the categorisation of the spatial clustering methods as discussed above. Majority of the methods belong to the distance-based approaches, where two or more spatial objects are considered as neighbours if the distances between them are no greater than a given distance threshold. The effort to identify colocations using distance-based mining techniques have been extended in several different directions including the space–time analysis of colocated features (Celik, 2015), colocation of fuzzy objects (Ouyang et al., 2017; Wang et al., 2022), and measurement of colocations using the network distance (Morioka et al., 2022a, 2022b, 2022c). Others have also looked into changing the ways in which colocations are measured, and these include generalised methods with some of the parametric constraints in colocation detection being relaxed (Yoo & Bow, 2012), and local demarcation of the colocation regions or the colocation equivalent of the local cluster-detection analysis (Deng et al., 2017).

The series of distance-based colocation methods that process disaggregate individual spatial features have also been met by some challenges: (1) They require a set of threshold distance to be determined a priori and this value often need to be identified in an exploratory fashion; (2) the sheer computational load of colocation calculation means that, usually, only a small sets of features are selected as a target colocation set; and (3) the risks of false positives and management of the random noise, where the algorithm may report colocations even if the features are randomly distributed (Barua & Sander, 2014).

To resolve the issues around scalability and to enable detection of large colocations (e.g. a colocation of 100 different feature types), several effective search algorithms have been proposed for rapidly reducing the number of candidate feature sets, e.g. frequent pattern (FP)-growth algorithm (Han et al., 2000; Yoo & Shekhar, 2006; Xiao et al., 2008) and, more generally, to alleviate the exponential computational load from searching through larger datasets (Celik, 2015; Yoo & Bow, 2012).

These challenges bring us back to the point made earlier in that colocations do not always indicate concentrated features (i.e. spatial clusters), as the focus of colocation analysis is on the *combination* and *variety* represented by the set of colocated features; whilst clustering analysis focuses on the degree of *concentration* of features. Figure 1 shows illustrative examples of colocations and clusters between two types of features. Figure 1(a) illustrates a scenario where two features form separate clusters respectively and these are not colocated, while Fig. 1(b) shows a case where colocations are observed across the study area but no clusters exist, and Fig. 1(c) highlighting a case of colocated clusters. While it is useful to improve on the efficiency of algorithms for colocation detection, in terms of the utilities of colocation detection in the real-world context, we may encounter situations where we would like to find the type of colocations in Fig. 1(c); i.e. colocations of features, each of which has some concentrations.

These could range from the types and the volume of crimes recorded in heavily problematic areas (Shiode et al., 2023), to areas in need of imminent medical attentions (e.g. areas suffering from outbreaks of several different types of epidemics). They require accurate detection of the types of features that are situated closely together and in mass to identify highly-problematic situations. Areas in which a certain level of aggregations (i.e. clusters) are found, or where "*group of occurrences of sufficient size and concentration*" (Knox, 1989) and "*subsets of spatial events whose instances are often located in close geographic proximity*" (Shekhar & Huang, 2001) all refer to situations that demand the spatial colocation of clusters of features. To find such colocations, we will take an *area-based* data mining method, rather than a *distance-based* method, as it helps relieve the computational cost issues by sparing the process of searching across each individual point, and the random noise can be also controlled by aggregating everything into areal units.

Given these backgrounds, this study will propose a method for detecting spatially colocated clusters that could fit many real-world applications. It proposes a method that comprises two steps: (1) spatial cluster detection of each type of features for detecting clusters or hotspots of the respective feature, and (2) colocation extraction for identifying the colocation between

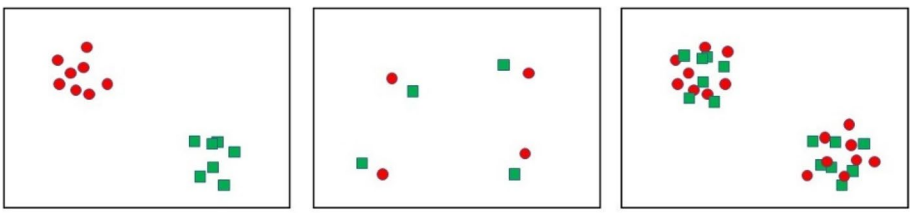

Fig. 1 Illustrative examples of spatial clusters and spatial colocations, (**a**) non-colocated clusters, (**b**) non-clustered colocation, (**c**) colocated clusters

the clusters of different types of features. As detailed below, the first step (cluster detection) uses the False Discovery Rate (FDR)-controlling statistical test, which can alleviate multiple testing problems; and the second step (colocation searches) adopts frequent pattern (FP)-growth algorithm (Han et al., 2000), one of the fastest mining algorithms for frequent patterns. The choice of these methods was based on their overall performance and prevalence in their respective domain. Their performance was not rigorously benchmarked and compared with alternative methods, as a number of existing studies have pursued the topic of efficiency already. They should still offer a sufficient level of efficiency that enables us to focus on the aim of our study, which is to derive colocations of clusters through the two-step approach.

3 Methodology

3.1 Spatial cluster detection using the False Discovery Rate (FDR) controlling method

The first step of the colocated clusters analysis is to detect spatial clusters. Of the two broad categories of cluster detection methods, it belongs to the second group of methods that seeks the *location of local clusters* (i.e. a local method), as opposed to those that examine *the global tendency to cluster* (i.e. a global method). Within the group of local methods, arguably the most widely used strand of methods is the scan statistic-type approach, which was systematised by Kulldorff and Nagarwalla (1995) as a search-window-type method. It identifies statistically significant concentration of events by creating a search window around the centroid of each spatial region and changing the radius of the window continuously to take any value between zero and a predetermined upper limit (Duczmal & Assunção, 2004). Using the likelihood ratio test, the scan statistic detects spatial regions where the underlying event occurrence rates are significantly higher inside the window than those outside.

The spatial scan statistic and its variants are indeed widely used for cluster detection. However, there is a limitation against detecting multiple clusters, since the alternative hypothesis assumes the presence of a single cluster. While it is technically possible to detect the secondary and other clusters by removing the clusters already detected, the limitation imposed on multiple cluster detection makes it difficult to use spatial scan statistic-based cluster detection for the situations of this study when simultaneous detection of a large number of clusters is expected.

To address this issue, two approaches have been proposed: namely, (1) an extension of spatial scan statistic (Mori & Smith, 2010), and (2) a cluster detection method based on the FDR-controlling procedure (Benjamini & Hochberg, 1995; Brunsdon & Charlton, 2011). The former offers multiple cluster variants of the spatial scan statistic using the Bayesian Information Criterion (BIC) (Mori & Smith, 2010). This method formulates *cluster schemes* to identify multiple cluster candidates, estimate the density parameters for all candidates in each cluster scheme based on the point distribution assumption, and calculate the BICs. After the cluster scheme with the maximum BIC is selected, its significance is tested through the Monte-Carlo simulation. While the model selection by the BIC accounts for multiple clusters and their locations, its search procedure is directly affected by the numbers of possible cluster schemes, and it may take long time to detect clusters in small area analysis.

The latter uses the FDR-controlling procedure for detecting clusters (Brunsdon & Charlton, 2011). First introduced to the geographic context by Caldas de Castro and Singer (2006), the FDR-controlling procedure offers a robust statistical method for detecting multiple clusters whilst avoiding multiple testing problems (Brunsdon & Charlton, 2011). It also holds a greater statistical power than the family-wise error rate controlling methods, namely the approaches traditionally used for multiple testing (e.g., Holm, 1979). The FDR-controlling procedure may be summarized as follows. Let us consider the case where m hypotheses are tested, and R null hypotheses are to be rejected (Table 1). The multiple testing increases the type I error occurrence (V) by chance. Benjamini and Hochberg (1995) defined the FDR as an index of false discoveries as follows:

$$FDR = E(V/R), (FDR = 0, \text{if} R = 0) \quad (1)$$

and proposed an FDR-controlling procedure that keeps the FDR less than a predetermined significance level α. The cluster detection method developed by Brunsdon and Charlton (2011) exploits the FDR-controlling procedure for configuring a set of alternative hypotheses that each region is a cluster, and rejecting the null hypotheses. This study will adopt this procedure in the cluster detection stage of the analysis.

Suppose that features within each feature type can be treated as point objects and that the number of these features is aggregated by the regional area unit in which they are contained (e.g. census tracts). Then, the search

Table 1 m number of hypotheses tests

	Rejected null hypothesis	Retained null hypothesis	Total
Null hypothesis is true	V	U	m_0
Alternative hypothesis is true	S	T	$m - m_0$
Total	R	$m - R$	m

for the clusters of each feature type would be conducted by counting these point data. Let G denote the entire study area and suppose that G consist of a finite number of subregions. Let Z denote one of the subregions in G, Z^C a complement region of Z in G, n_Z and n_{Z^C} the count of point features in Z and Z^C respectively; and a_Z and a_{Z^C} the size of Z and Z^C, respectively. The sizes of regions could be defined by their respective areas or the number of features of all feature types in each region. Here, we assume that the spatial distributions of points in Z and Z^C conform to the Poisson distributions in which the point counts within each region are proportional to the sizes of the regions. Then,

$$n_Z/a_Z \sim Poisson(\lambda_Z), n_{Z^C}/a_{Z^C} \sim Poisson(\lambda_{Z^C}) \quad (2)$$

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

$$H_1|\lambda_Z > \lambda_{Z^C} \quad (3)$$

and its null hypothesis is

$$H_0|\lambda_Z = \lambda_{Z^C} \quad (4)$$

Suppose that the number of points observed in G was N. Then, n_Z conforms to the following binomial distribution:

$$n_Z \sim B_i\left(N, \frac{a_Z\lambda_Z}{a_Z\lambda_Z + a_{Z^C}\lambda_{Z^C}}\right) \quad (5)$$

If the null hypothesis was true,

$$n_Z \sim B_i\left(N, \frac{a_Z}{a_Z + a_{Z^C}}\right) \quad (6)$$

Then, the p-value of the null hypothesis of Z, p_z is

$$p_Z = \sum_{i=n_z}^{N} \binom{N}{i} \left(\frac{a_Z}{a_Z + a_{Z^C}}\right)^i \left(\frac{a_{Z^C}}{a_Z + a_{Z^C}}\right)^{N-i} \quad (7)$$

The p-values of the null hypotheses of all subregions in G are calculated, and these hypotheses are tested by the B–H procedure (Benjamini & Hochberg, 1995), a statistical test based on the FDR-controlling procedure.

3.2 Extracting colocation patterns

The second stage of our methodology focuses on the extraction of specific combinations of feature types that are forming spatial colocations across the study area. As described before, the analysis uses the locations of clusters detected for each feature type to extract the combinations of features. If the clusters of feature types A, B, and C are colocated in many regions, this combination is considered as a colocation pattern. To extract such patterns, this study uses a *frequent-pattern-growth algorithm* (Agrawal & Srikant, 1994; Han, et al. 2000). The FP-growth algorithm is run by constructing a frequent pattern tree (FP-tree). The fact that the FP-tree can be built from a single scan of the data and can be also processed in parallel for improving the performance makes it particularly efficient for mining frequent sets in large datasets. It is commonly applied in market research for analysing the consumer purchase behaviour, namely exploring which combinations of items are bought together. Frequent pattern mining distinguishes the frequent pattern by *support*, that is, the count of a combination of features. If the frequency of an observed combination of features reaches or exceed a predetermined threshold (hereafter called the *minimum support*) it will be extracted as one of the frequent patterns; i.e. a colocation.

Suppose that Table 2 represents results of cluster detection. Geographical clusters of four feature types, A, B, C and D, are located across Regions I-V. The *support* of feature type A is 60%, as clusters of feature type 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 III and V. When the *minimum support* is set to 40%, nine patterns, {A}, {B}, {C}, {D}, {A, B}, {A, C}, {B, C}, {A, D}, {A, B, C}, are extracted from the spatial distribution of the clusters.

As the numbers of feature types and regions increase, the search for frequent patterns becomes time consuming. Several algorithms have been proposed to overcome this problem, and this study utilises the FP-growth algorithm (Han et al., 2000), which is one of the most efficient pattern-mining algorithms.

4 Analysis

4.1 Dataset

To test the proposed methodology and to gain insights into any hidden patterns of colocation, we used location data of industries and businesses in Japan as a case study. The study of the geographic concentration of industries has attracted many researchers in geography and spatial

Table 2 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

economics. The colocation of specific combination of industry types, namely the phenomenon wherein offices and factories of related businesses are located near one another, is a key issue that could unravel the mechanisms of geographic concentration of industries. Various theories have been proposed to explain these mechanisms, demonstrated by relevant empirical studies (e.g., Ellison & Glaeser, 1999; Ellison et al., 2010).

The 2012 Economic Census for Business Frame of Japan is a statistical dataset that covers the entire records of establishments and enterprises in Japan. The smallest spatial unit available for this data is a 500-m grid cell, which we will use as a unit of analysis in this study, i.e. the number of establishments of each industry type is aggregated by the 500-m grid cells. We adopt the smallest available spatial unit, because certain types of retails and the service sector in Japan tend to show dense but compact concentrations around train stations and other points of interest, they cannot be detected clearly unless we apply a high resolution, or a sufficiently small spatial granularity. The industries are classified into 16 *main categories*, which are sub-divided into 86 *sub-categories* as per the Japan Standard Industrial Classification (please refer to the Appendix for a complete list of main categories and sub-categories). Although we need 1,515,129 of the 500-m grid cells to cover the entire extent of Japan, the dataset contains records from only 336,646 grid cells wherein at least one enterprise is located. It is important to note that this is a zero-truncated dataset. The numbers of establishments are 6,009,389. To confirm the validity of the method, a simulation test was conducted with this dataset whereby the same number of facilities for each respective industry were randomly reassigned to the 1.5 million grid cells and were tested for colocations. Results showed very few clustered colocations as would be expected from a randomised pattern.

4.2 Detecting clusters of each industry

Cluster detection was carried out under the condition that the FDR value remains below the significance level $\alpha = 0.01$. The analysis was carried out using our original code developed on a C++ compiler platform. Clusters of each industry type were determined by the density of the respective establishments within each grid square. A total of 26,059 grid cells contained at least one industrial cluster, amounting to 1.7% of the total grid cells in Japan and 7.7% of grid cells with at least one industry. Table 3 shows the top ten industries that appeared most frequently as part of colocated clusters.

Looking across the main categories in Table 3, we see 2 types of retail industries (I: Retail trades, and M: Restaurants/diner), which are subdivided further into 4 subcategories: I-57 – Retail Trade (dry goods, apparel); I-58 – Retail trade (food, beverage); I-60 – Retail trade (miscellaneous); and M-76 – Eating and drinking places). In fact, one of the defining characteristics of industrial colocation in Japan is that they tend to form clusters of small retail stores, followed by laundry/dry cleaning and hair salon. In terms of public facilities, post offices and health clinics appeared frequently.

Figure 2 shows an example of detecting clusters of "Miscellaneous retail trade (code 60)" around the greater Tokyo region. We chose to illustrate this example, as they returned the highest number of detected clusters. Figure 2(a) shows the density of the miscellaneous retail trade establishments. Overall, it is very high in the Tokyo region with many areas hosting over 200 of miscellaneous retail trades (including furniture, books, drugs and cosmetics, stationery, tobacco, musical instruments, watches). Figure 2(b) shows the grid cells with clusters detected with the FDR. Clearly, areas with high-density of retail establishments are detected as clusters (around 20% of the grids in the area).

Table 3 Top 10 industries whose clusters appear most frequently in areas of industry colocations

Main Categories	Sub-Categories	Categories of industries	# Clusters in colocations
I	60	Miscellaneous retail trade	7,569
M	76	Eating and drinking places	7,563
N	78	Laundry, beauty, and bath services	7,431
I	58	Retail trade (food and beverage)	7,116
K	69	Real estate lessors and managers	5,452
P	83	Medical and other health services	4,423
H	49	Postal activities, including mail delivery	3,682
D	7	Construction work by specialist contractor, except equipment installation work	2,545
I	57	Retail trade (dry goods, apparel, and apparel accessories)	2,545
D	6	Construction work (general), including public and private construction work	2,441

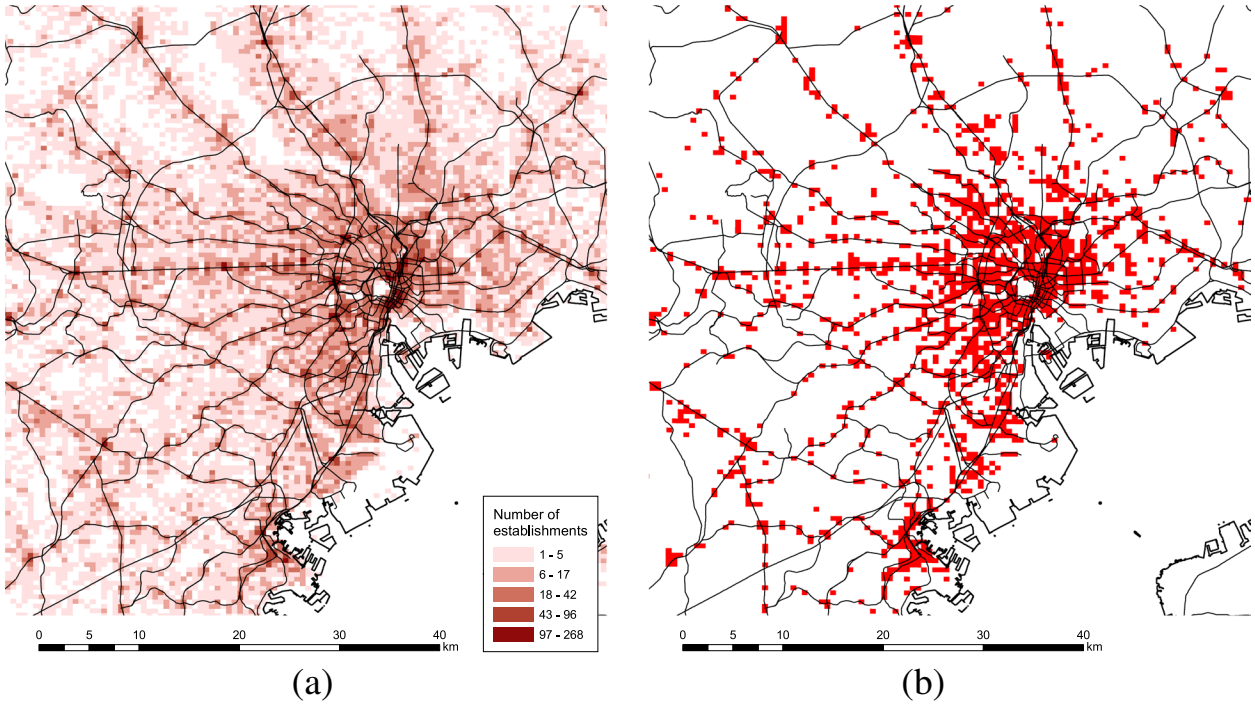

Fig. 2 Clusters of miscellaneous retail trade based on the number of establishments: (**a**) density of the facilities, and (**b**) detected clusters

4.3 Extracting colocation patterns of clusters of industries

Finally, we measured and extracted the colocation patterns among the clusters of various retail and industry establishments using FP-growth algorithm. Colocation method was also developed as a C++programme and was executed in a C++compiler. The minimum support was set as 1% of the number of grid cells having at least one industrial cluster, namely 26,059 grid cells in this case. In other words, a series of combinations of industry clusters were extracted as colocated clusters if more than 260 instances of such combination were observed. The total number of extracted patterns of clustered industry colocation was 24,870 cases. The sheer volume of colocation patterns detected made it difficult to add a meaningful interpretation for every combination of detected colocation of industry clusters. For this reason, we focused on the most frequent combinations that appeared at each colocation size. Table 4 shows the list of most frequently appearing combinations of industry clusters by sub-categories for each respective size of colocation, ranging from the smallest possible combination between a pair of industry types to the largest combination of 10 industry types. Highlighted industry types denote their first appearance in the table. The purpose of producing this table is to identify the most comprehensive set of colocated clusters by eliminating the less frequent duplicates of colocations from the entire set of colocations detected.

Of these colocations, *Retail trade (food and beverage)* (Code 58), *Miscellaneous retail trade* (Code 60), *Eating and drinking places* (Code 76), and *Laundry, beauty and bath services* (Code 78) came up as the industry types that form colocations of clusters most frequently. The first three are typical examples of the food services and retail industry which prevail across the greater Tokyo region. The fourth industry is a *hair and beauty service* that has some strong ties with the food services/retails. Given their frequent and tightly-knit colocation pattern, we will call the set of these four industry types the primary group of industries, in that they are closely related to people's daily life regardless of the size of the city. The next layer of industries are perhaps less frequently colocated but are nonetheless essential to our daily activities, and these include *medical*, *health*, and *education* related services. The tertiary layer industries include other retail, real estate agencies, amusement services, professional, technical and other miscellaneous services. Industries in this group may not be vital to maintaining our daily life but are frequently used and can be found in a medium-sized or a larger town. Figure 3 summarises the multi-layered pattern between the primary, the secondary and the tertiary colocations of clusters of industries.

Comparing Tables 3 and 4 makes us aware that clusters tend to occur in some but not all areas, and they do not necessarily form colocations. These anomalies include construction companies (Codes 6 and 7), and the Postal

Table 4 The most representative colocations of industry clusters for each colocation size

#	Size									
4978	2	58 Retail food	60 Misc retail							
4021	3	58 Retail food	60 Misc retail	78 Beauty						
3452	4	58 Retail food	60 Misc retail	76 Eat/drink	78 Beauty					
2620	5	58 Retail food	60 Misc retail	76 Eat/drink	78 Beauty	83 Med/Health				
1372	6	58 Retail food	60 Misc retail	76 Eat/drink	78 Beauty	82 Misc edu.	83 Med/Health			
901	7	57Retail dry	58 Retail food	60 Misc retail	76 Eat/drink	78 Beauty	82 Misc edu.	83 Med/Health		
569	8	57Retail dry	58 Retail food	60 Misc retail	68 Real estate	76 Eat/drink	78 Beauty	82 Misc edu.	83 Med/Health	
393	9	57Retail dry	58 Retail food	60 Misc retail	68 Real estate	76 Eat/drink	78 Beauty	80 Amusement	82 Misc edu.	83 Med/Health
312	10	58 Retail food	60 Misc retail	68 Real estate	72 Prof svcs.	74 Tech svcs.	76 Eat/drink	78 Beauty	82 Misc edu.	83 Med/Health 92 Misc bus

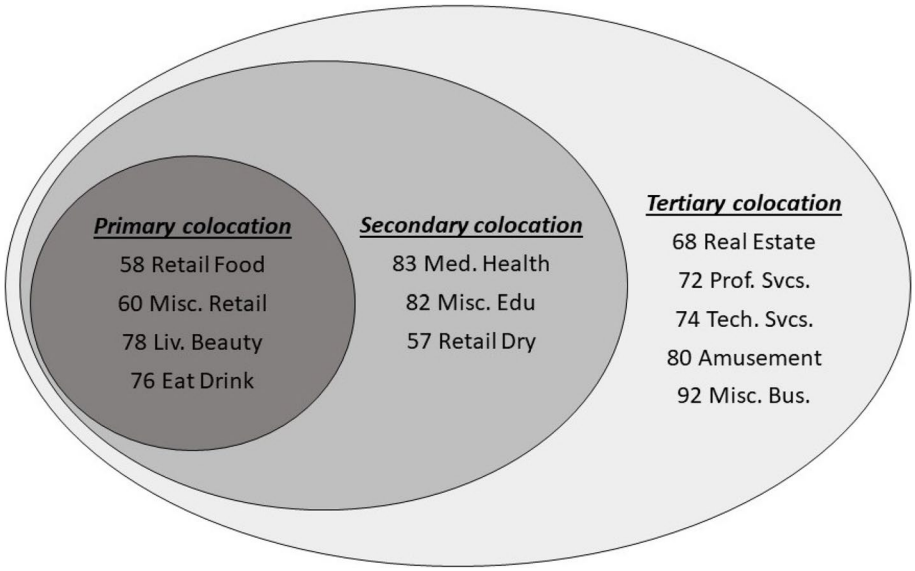

Fig. 3 An illustrative diagram showing the multi-layered structure of industry cluster colocations

Services (Code 49). It shows that clusters are frequently formed in a range of places (Table 3), but they do not always colocate, which could mean that there are no tangible benefits for these facilities to agglomerate with other types of industry. The analysis also revealed that there are some industries that do not colocate with any other industries (Table 5). Most of them are in the manufacturing domain (Main Category E), some of which require wide space to load and unload materials; and to establish an enterprise (Main Category R).

One of these non-colocating industries "Manufacture of food (Code 9)" is mapped in Fig. 4. The difference between this and "Miscellaneous retail trade (Code 60)" (Fig. 2) is obvious in that Manufacture of food shows clusters in a few specific places whereas miscellaneous retail trade is much more widely spread.

Table 5 Industries that do not colocate with other industries

Main category	Sub-category	Category of industries
E	9	Manufacture of food
E	13	Manufacture of furniture and fixtures
E	21	Manufacture of ceramic, stone and clay products
E	32	Miscellaneous manufacturing industries
H	43	Road passenger transport
H	48	Services incidental to transport
R	89	Automobile maintenance services
R	90	Machine, etc. repair services, except otherwise classified

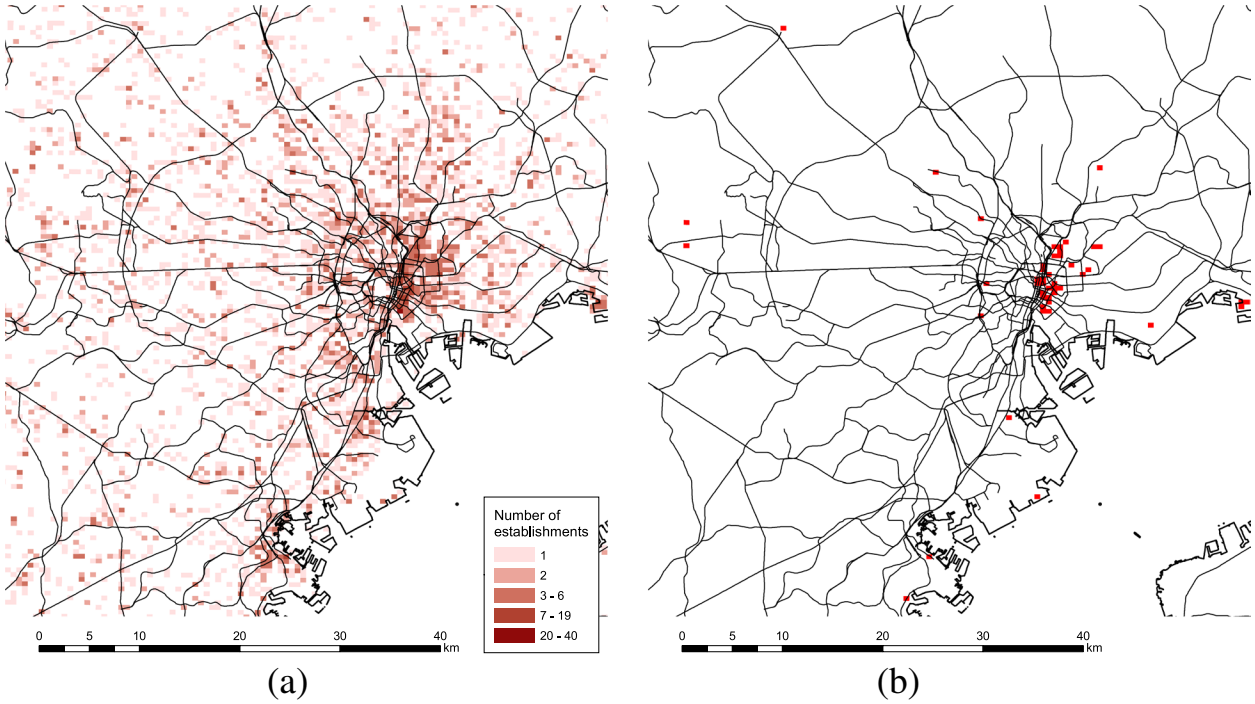

Fig. 4 Cluster mapping of Manufacture of food (Code 9): (**a**) clusters by the density distribution, and (**b**) detected clusters

4.4 Sensitivity of the clustered colocations

As with all statistical procedures, sensitivity of the cluster detection and the colocation frequency is determined by the significance level of the FDR-controlling procedure and the minimum support level of the frequent patterning mining. The above study was conducted with the conditions that the FDR value remains below the significance level $\alpha = 0.01$, and the minimum support at 1% of all grids that contain a valid cluster. To assess the sensitivity of the outcomes, we also carried out the same analysis using different combinations of the significance level and minimum support, as illustrated in Table 6.

These results confirm that the overall results are stable. Change in the significance level of FDR generally affects the frequency of the largest set of colocated industries being detected (*item b* in Table 6). Interestingly, depending on the significance level of the FDR control, the combination of industries comprising the most frequent, largest set varies slightly. This variation in the membership of the colocation clusters is nonetheless minimum, and those omitted from the list tend to return as a component of the second or the third most frequent set of the largest set of colocation (*item d* in Table 6 shows the number of different combinations of industries detected with the largest set of colocation for the respective minimum-support level). It shows that the outcomes remain robust, but they are also subject to the sensitivity of the significance level. Change in the minimum support of FP-growth truncates the largest colocated set (*item a* in Table 6, with the specific industries shown in *item c*). Naturally, a tighter constraint on the frequency would reduce the size of the largest set of colocated industries and, thereby, the results in extracting a subset of that set. Given these outcomes, the method can be considered to offer robust outcomes where the core set of colocations are consistent regardless of the threshold values for significance, and the only variations arises when the most frequent combinations are extracted for different permutation of the thresholds.

5 Discussion

Outcomes from the analysis demonstrate that the proposed method has the following advantages over its existing counterparts. First, it offers a much clearer representation of colocation. As our method uses *clusters of* point features as the unit of colocation, rather than *individual* point features for identifying colocations, it prevents the detection of by-chance colocations whilst also facilitating clearer representation and interpretation of their colocation patterns. Empirical analysis investigated over 6 million industry location data in Japan for the formation of clusters by 86 industry types aggregated to 500 m-square grid units. Results suggest that part of the industry form clear cluster colocations that have a multi-layered structure (Fig. 3 and Table 4). The proximity between the colocated clusters of industries is

Table 6 Sensitivity of cluster colocations

FDR \ Min Support		0.005	0.01	0.02
0.005	a	18	12	8
	b	122	254	683
	c	(39,54,55,57,58,60,68,69,72, 74,76,78,80,82,83,91,92,93)	(39,57,58,60,68,72, 74,76,80,82,83,92)	(57,58,60,68, 76,78,82,83)
	d	10	16	11
0.01	a	18	12	8
	b	132	277	733
	c	(39,54,55,57,58,60,68,69,72, 74,76,78,80,82,83,91,92,93)	(57,58,60,68,72,74, 76,78,80,82,83,92)	(57,58,60,68, 76,78,82,83)
	d	5	6	10
0.02	a	18	12	8
	b	148	296	834
	c	(39,54,55,57,58,60,68,69,72, 74,76,78,79,80,82,83,91,92)	(57,58,60,68,72,74, 76,78,80,82,83,92)	(57,58,60,68, 76,78,82,83)
	d	9	8	10

where

a: Number of industries in the most frequent, largest set of colocated industries,

b: Number of frequency for the most frequent, largest set of colocated industries,

c: The set of industries comprising the most frequent, largest set of colocated industries, and

d: Number of all observed combinations of the largest set of colocated industries

considered as a product of either some form of interaction (cooperative or competitive) among them, a shared customer base, or a common underlying factor such as the need to service a small community in the area. The multi-layered structure of the industry colocated clusters seems to reflect the scale and the structure of urban hierarchy in that small-sized colocations of the primary layer appear more frequently, including those embedded within small communities but they are likely to attract only local customers, whereas the larger-sized colocations across multi-layered structure would appeal to a larger community or a city and can be sustained by the larger customer base. Figure 5 compares the outcomes of cluster colocation analysis for different colocation sizes, ranging from a compact but frequent colocation between two types of industries to a larger but less common colocations between 10 industries, which respectively reflect the different layers of urban hierarchical structure.

The multi-layered structure derived in this study, along with the specific combinations of industry types detected at different colocation size, offers interesting insights into the relationship between the industries in Japan that were identified as forming colocations with one another. At the same time, these insights could well be unique to the situation in Japan and findings may not be directly applicable in the context of other countries and regions, as the local industry structures and urban dynamics tend to reflect their own economical, cultural and environmental conditions. While this presents a limitation of this study, the proposed methodology itself is transferrable and can be applied for discovering colocated clusters in other study areas. Most of the existing studies on spatial colocation focus on the algorithmic improvements, which is an important area of progress, but they tend not to investigate the colocations discovered (e.g. look into their context or their geographical distributions). Such pursuit would give us a clue for a better understanding of the urban and regional structure, and characterisation of the areas. In this sense, this study not only proposed a method but also investigated what the discovered colocations tell us.

Another limitation is the arbitrary nature of the choice of the threshold value for assessing the frequency of colocation. While it may not cause a significant impact on the detection of highly frequent colocations, it could affect the detection of less frequent colocations. Too few detection would mean truncating too much information, while too many detection could mean that not all data is meaningful. The larger the dataset, the more difficult it becomes to interpret the less frequent patterns of colocation. In other words, the method extracts the clustered colocations that are frequent and strong, while the ones that are eliminated in the process are less frequent and would not help interpret the association between the colocated industries.

Thirdly, our proposed method uses aggregated areal data, which is subject to the modifiable area unit

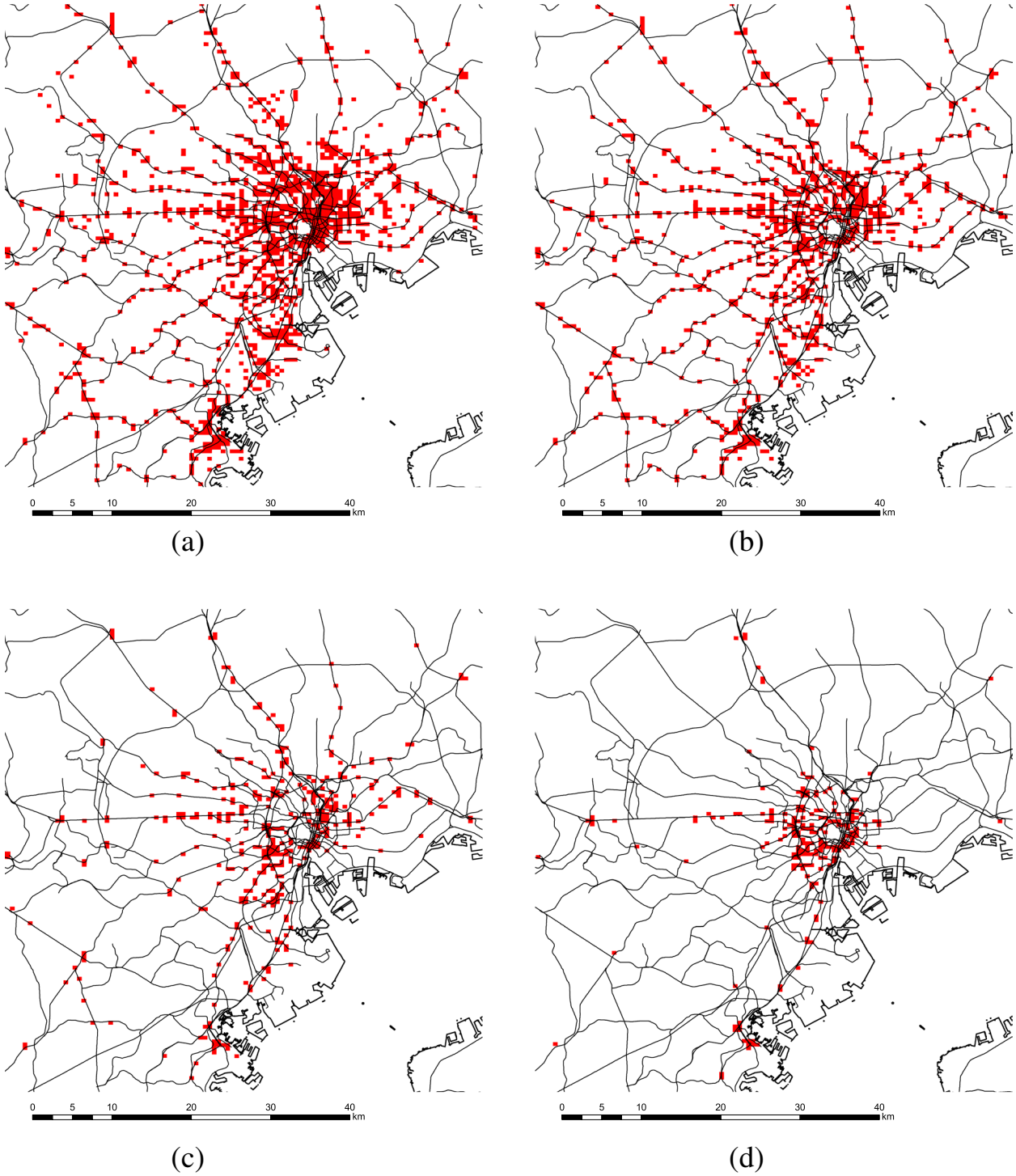

Fig. 5 Location of most frequently colocated clusters of industries between: (**a**) 2 types of industries; (**b**) 5 types of industries; (**c**) 7 types of industries, and (**d**) 10 types of industries

problem. In other words, the results would be affected if we changed the areal units used for data aggregation. The outcome of the analysis would be determined by what spatial granularity was used. As the aim of this study was to investigate local colocation of clusters at the community level, the 500m grid mesh seemed to be a logical choice. However, depending on the scale of the analysis, another unit of analysis may suit better.

Finally, this study does not account for the weight of individual facility and, instead, extracts the clusters and colocations by the volume of facilities only. A non-weighted extraction is useful for processing a large volume of data with a range of possible combinations between different entities, it is ill suited for specific, unusual cases where a single significant facility (e.g. a railway terminal station, a large football stadium or an upmarket department store attracting food outlets, retail and service industry to its vicinity). While such a significant facility will likely attract multiple industry types and form a sufficient cluster colocation to be extracted with the current setting of our method, assigning a large weight value to a significant facility could help clarify the cause for multiple industries to concentrate. It would be one of our future research directions.

6 Conclusion

This study proposed a new approach for identifying spatial colocations of the concentration of geographic features in the form of detecting the colocation of clusters (i.e. clusters of clusters) among point-type features. The proposed approach can (1) detect multiple clusters whilst avoiding multiple testing problems; (2) detect colocation of clustered features, which would lead to a variety of applications in the real world; and (3) be applied to large data sets because of an area-based approach (rather than distance-based approach).

As discussed in the above, the empirical analysis encompassing over 6 million industry locations in Japan helped illustrate the formation of colocated clusters in a multi-layered structure (Fig. 3 and Table 4). To our knowledge, a systematic extraction of the association between colocated industry clusters remain understudied. In this sense, we believe the diagram (Fig. 3) offers a novel contribution that captures the hierarchy of industrial activities and their agglomeration, akin to the classic urban hierarchy of Christaller's central place theory (Christaller, 1933; Openshaw & Veneris, 2003). It opens up an avenue for future research to establish a wider understanding on how different industries form colocated clusters across different scales, and how that would link to the notion of urban hierarchies.

In terms the application in other contexts, the proposed method may offer insights into the colocations of event-type features that change over time, rather than the static features such as the ones addressed in this study. This includes the colocation of crime incidents of different types. It is known that some crime types are spatially associated with some other crime types. By having more detailed information on which combinations of crime types tend to form colocations, we can better understand the associations between different crime types. Another possible application is the colocation of diseases. Many studies investigate clusters of a single type of diseases, but many areas in the world may suffer from different types of diseases. Colocation patterns that can be extracted from a large data set will give a wealth of new information that is otherwise difficult to obtain, including the common underlying ground that causes a specific set of diseases, and it may help predict what kind of diseases will likely happen (or their spatial diffusion in the future) in a specific area.

Finally, the methodology proposed here can also be used for comparative analysis of different colocation types across different regions and nations. The fundamental interest in such a study would be to unravel what makes difference in colocation patterns and more broadly how these colocations develop over time, and in this sense, this type of colocation study constitutes a background exploration on the areal structure. To gain a solid understanding on this point, the inquiries may benefit from colocation pattern analysis from multiple time points.

Appendix

The following list shows the main categories (denoted by the alphabets) and the sub-categories (denoted by numeric codes) of Japan Standard Industrial Classification.

III. Mining and Quarrying of Stone and Gravel
5 Mining and quarrying of stone and gravel
IV. Construction
6 Construction work, general including public and private construction work
7 Construction work by specialist contractor, except equipment installation work
8 Equipment installation work
V. Manufacturing
9 Manufacture of food
10 Manufacture of beverages, tobacco and feed
11 Manufacture of textile mill products
12 Manufacture of lumber and wood products, except furniture
13 Manufacture of furniture and fixtures
14 Manufacture of pulp, paper and paper products
15 Printing and allied industries
16 Manufacture of chemical and allied products
17 Manufacture of petroleum and coal products
18 Manufacture of plastic products, except otherwise classified
19 Manufacture of rubber products
20 Manufacture of leather tanning, leather products and fur skins

21 Manufacture of ceramic, stone and clay products
22 Manufacture of iron and steel
23 Manufacture of non-ferrous metals and products
24 Manufacture of fabricated metal products
25 Manufacture of general-purpose machinery
26 Manufacture of production machinery
27 Manufacture of business oriented machinery
28 Electronic parts, devices and electronic circuits
29 Manufacture of electrical machinery, equipment and supplies
30 Manufacture of information and communication electronics equipment
31 Manufacture of transportation equipment
32 Miscellaneous manufacturing industries

VI. Electricity, Gas, Heat Supply and Water
33 Production, transmission and distribution of electricity
34 Production and distribution of gas
35 Heat supply
36 Collection, purification and distribution of water, and sewage collection, processing and disposal

VII. Information and Communications
37 Communications
38 Broadcasting
39 Information services
40 Internet based services
41 Video picture, sound information, character information production and distribution

VIII. Transport and Postal Activities
42 Railway transport
43 Road passenger transport
44 Road freight transport
45 Water transport
46 Air transport
47 Warehousing
48 Services incidental to transport
49 Postal activities, including mail delivery

IX. Wholesale and Retail Trade
50 Wholesale trade, general merchandise
51 Wholesale trade (textile and apparel)
52 Wholesale trade (food and beverages)
53 Wholesale trade (building materials, minerals and metals, etc)
54 Wholesale trade (machinery and equipment)
55 Miscellaneous wholesale trade
56 Retail trade, general merchandise
57 Retail trade (dry goods, apparel and apparel accessories)
58 Retail trade (food and beverage)
59 Machinery and equipment
60 Miscellaneous retail trade

X. Finance and Insurance
62 Banking
63 Financial institutions for cooperative organizations

XI. Real Estate and Goods Rental and Leasing
68 Real estate agencies
69 Real estate lessors and managers
70 Goods rental and leasing

XII. Scientific Research, Professional and Technical Services
71 Scientific research and development institutes
72 Professional services, n.e.c.
73 Advertising
74 Technical services, n.e.c.

XIII. Accommodations, Eating and Drinking Services
75 Accommodations
76 Eating and drinking places
77 Food tale out and delivery services

XIV. Living-Related and Personal Services and Amusement Services
78 Laundry, beauty and bath services
79 Miscellaneous living-related and personal services
80 Services for amusement and hobbies

XV. Education, Learning Support
81 School education
82 Miscellaneous education, learning support

XVI. Medical, Health Care and Welfare
83 Medical and other health services
84 Public health and hygiene
85 Social insurance and social welfare

XVII. Compound Services
86 Postal services
87 Cooperative associations, n.e.c.

XVIII. Services, n.e.c.
88 Waste disposal business
89 Automobile maintenance services
90 Machine, etc. repair services, except otherwise classified
91 Employment and worker dispatching services
92 Miscellaneous business services
93 Political, business and cultural organizations
94 Religion
95 Miscellaneous services

The following main categories and sub-categories have been excluded from this list:

Main categories —"A. Agriculture and Forestry,", "B. Fisheries," "L. Scientific Research, Professional and Technical Services," and "S. Government, Except Elsewhere Classified;" and

Sub-categories—"61. Nonstore retailers," "64. Nondeposit money corporations, including lending and credit card business," "65. Financial products transaction dealers and futures commodity transaction dealers," "66. Financial auxiliaries," and "67. Insurance institutions, including insurance agents, brokers and services."

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization, R.I.; methodology, R.I.; validation, R.I., N.S. and S.S.; resources, R.I.; writing—original draft preparation, R.I., N.S. and S.S.; writing—review and editing, N.S. and S.S.; visualization, R.I. All authors have read and agreed to the manuscript.

Funding
No funding was received for conducting this study.

Availability of data and materials
All industry data used in this study are publicly available and can be accessed through https://www.e-stat.go.jp/en/.

Declarations

Competing interests
We have no interests, financial or otherwise, that are directly or indirectly related to the work submitted for publication.

Received: 18 July 2023 Revised: 11 September 2023 Accepted: 20 September 2023
Published online: 06 November 2023

References

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, Santiago, Chile. Morgan Kaufmann 1994, pp.487–499. ISBN 1-55860-153-8.

Aldstadt, J., & Getis, A. (2006). Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis, 38*, 327–343.

Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis, 27*(2), 93–115.

Anselin, L., Syabri, I., & Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, Santa Barbara. Edited by L. Anselin and S. Rey. Santa Barbara, CA: Center for Spatially Integrated Social Science, University of California, CD-ROM.

Barua, S., & Sander, J. (2014). Mining statistically significant co-location and segregation patterns. *IEEE Transactions on Knowledge and Data Engineering, 26*(5), 1185–1199.

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.

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

Bonetti, M., & Pagano, M. (2005). The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Statistics in Medicine, 24*(5), 753–773.

Boots, B. N., & Getis, A. (1988). *Point Pattern Analysis*. Sage Publications.

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. (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.

Celik, M. (2015). Partial spatio-temporal co-occurrence pattern mining. *Knowledge and Information Systems, 44*, 27–49.

Christaller, W. (1933). *Die Zentralen Orte in Süddeutschland*. Jena: Gustav Fischer.

Clark, P. J., & Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology, 35*, 445–453.

Cuzick, J., & Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society B, 52*, 73–104.

Deng, M., Cai, J., Liu, Q., He, Z., & Tang, J. (2017). Multi-level method for discovery of regional co-location patterns. *International Journal of Geographical Information Science, 31*(9), 1846–1870.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Oxford University Press.

Diggle, P. J., & Chetwynd, A. D. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics, 47*, 1155–1163.

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

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. (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. *American Economic Review, 100*, 1195–1213.

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

Gatrell, A. C. (2002). *Geographies of Health*. Blackwell Publishing.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*(3), 189–206.

Han, J., Pei, J., & Yiwen, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record, 29*(2), 1–12.

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

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering, 16*(12), 1472–1485.

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

Knox, E. G. (1989). Detection of clusters. In P. Elliott (Ed.), *Methodology of Enquiries into Disease Clustering* (pp. 17–20). Small Area Health Statistics Unit.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods, 26*, 1481–1496.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine, 14*, 799–810.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., & Key, C. (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health, 88*, 1377–1380.

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

Lawson, A. B. (2006). *Statistical Methods in Spatial Epidemiology* (2nd ed.). John Wiley & Sons.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society B, 10*, 243–251.

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

Morimoto, Y. (2001). Mining Frequent Neighboring Class Sets in Spatial Databases. Proc. ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, pp. 353–358.

Morioka, W., Kwan, M.-P., Okabe, A., & McLafferty, S. L. (2022a). A statistical method for analyzing agglomeration zones of co-location between diverse facilities on a street network. *Transactions in GIS, 00*, 1–22. https://doi.org/10.1111/tgis.12969

Morioka, W., Kwan, M.-P., Okabe, A., & McLafferty, S. L. (2022b). Local indicator of spatial agglomeration between newly opened outlets and existing competitors on a street network. *Geographical Analysis, 00*, 1–16. https://doi.org/10.1111/gean.12343

Morioka, W., Okabe, A., Kwan, M.-P., & McLafferty, S. L. (2022c). An exact statistical method for analyzing co-location on a street network and its computational implementation. *International Journal of Geographical Information Science, 36*(4), 773–798.

Openshaw, S., & Veneris, Y. (2003). Numerical experiments with central place theory and spatial interaction modelling. *Environment and Planning A, 35*(8), 1389–1403.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis, 27*(4), 286–306.

Ouyang, Z., Wang, L., & Wu, P. (2017). Spatial co-location pattern discovery from fuzzy objects. *International Journal of Artificial Intelligence Tools, 26*(2), 1750003.

Patil, G. P., & Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics, 11*, 183–197.

Ripley, B. D. (1976). The second-order analysis of stationary point process. *Journal of Applied Probability, 13*, 255–266.
Ripley, B. D. (1981). *Spatial Statistics*. John Wiley & Sons.
Rushton, G., & Lolonis, P. (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine, 15*, 717–726.
Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. *Proceedings of the International Symposium on Spatial and Temporal Databases* (pp. 236–256). Berlin: Springer.
Shekhar, S., & Chawla, S. (2003). *Spatial Databases: A Tour*. Prentice Hall.
Shiode, S., & Shiode, N. (2020). A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets. *Computers, Environment and Urban Systems, 83*, e101500.
Shiode, S., & Shiode, N. (2022). Network-based space-time Scan Statistics for detecting micro-scale hotspots. *Sustainability, 14*(24), 16902.
Shiode, S., Shiode, N., & Inoue, R. (2023). Measuring the colocation of crime hotspots. *GeoJournal, 88*, 3307–3322.
Takahashi, K., Kulldorff, M., Tango, T., & Yih, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics, 7*, 14.
Tango, T. (1999). Comparison of general tests for spatial clustering. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, L. Viel, & R. Bertollini (Eds.), *Disease Mapping and Risk Assessment for Public Health* (pp. 111–117). John Wiley & Sons.
Tango, T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine, 19*, 191–204.
Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics, 4*, 11.
Tonkin, M., Woodhams, J., Bull, R., Bond, J. W., & Palmer, E. J. (2011). Linking different types of crime using geographical and temporal proximity. *Criminal Justice and Behavior, 38*, 1069–1088.
Turnbull, B., Iwano, E. J., Burnett, W. S., Howe, H. L., & Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in Upstate New York. *American Journal of Epidemiology, 132*, 136–143.
Wang, X., Lei, L., Wang, L., Yang, P., & Chen, H. (2022). Spatial colocation pattern discovery incorporating fuzzy theory. *IEEE Transactions on Fuzzy Systems, 30*(6), 2055–2072.
Xiao, X., Xie, X., Luo, Q., & Ma, W. (2008). Density based co-location pattern discovery. *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 250–259). Irvine: (ACM-GIS).
Yoo, J.S., & Shekhar, S. (2004). A partial join approach for mining co-location patterns. In: *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems* (ACM-GIS).
Yoo, J. S., & Bow, M. (2012). Mining spatial colocation patterns: A different framework. *Data Mining and Knowledge Discovery, 24*(1), 159–194.
Yoo, J. S., & Shekhar, S. (2006). A join-less approach for mining spatial colocation patterns. *IEEE Transactions on Knowledge and Data Engineering, 18*(10), 1323–1337.

Publisher's Note