Discovery of Regional Co-location Patterns with k-Nearest Neighbor Graph

Feng Qian[1], Kevin Chiew[2], Qinming He[1], Hao Huang[1], and Lianhang Ma[1]

[1] College of Computer Science and Technology, Zhejiang University, P.R. China
[2] School of Engineering, Tan Tao University, Vietnam

Abstract. The spatial co-location pattern mining discovers the subsets of features of which the events are frequently located together in a geographic space. The current research on this topic adopts a distance threshold that has limitations in spatial data sets with various magnitudes of neighborhood distances, especially for mining of regional co-location patterns. In this paper, we propose a hierarchical co-location mining framework by considering both varieties of neighborhood distances and spatial heterogeneity. By adopting k-nearest neighbor graph (kNNG) instead of distance threshold, we propose "distance variation coefficient" as a new measure to drive the mining process and determine an individual neighborhood relationship graph for each region. The experimental results on a real world data set verify the effectiveness of our framework.

Keywords: co-location pattern, kNNG, variation coefficient.

1 Introduction

The spatial co-location pattern mining [1] discovers the subsets of features (co-locations) of which the events are frequently located together in a geographic space. It has been applied to many areas like mobile commerce, earth science, biology, public health, and transportation [2]. Figure 1(a) shows a sample data set containing instances of six spatial features represented by distinct shapes. The instances of features describe the presence of their instances at different locations in a 2D or 3D space. A careful review reveals four co-location patterns as illustrated in Figure 1(c). To discover these patterns, the current research (see e.g., [2,3]) adopts an approach with two phases, namely (1) converting the spatial data set into a neighborhood relationship graph (NRG for short) using a distance threshold as illustrated in Figure 1(b) in which the distance threshold is defined as the maximal distance allowed for two events to be neighbors; and (2) finding prevalent co-locations based on their clique instances in the derived graph.

The first phase implicitly assumes the normal distances of neighbors being smaller than the predefined distance threshold. This requires an approximately uniform distribution of spatial events across the space, as well as the joint distributions of features. In real life however, the data density often varies across different areas, leading to more complex joint distributions. For such data sets with various densities, an improper distance threshold may severely affect the mining results due to two reasons. (1) First, a small distance threshold may ignore many clique instances of prevalent co-locations in sparse areas, resulting in these co-locations being under-estimated. (2) Second, a large

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 174–186, 2013.

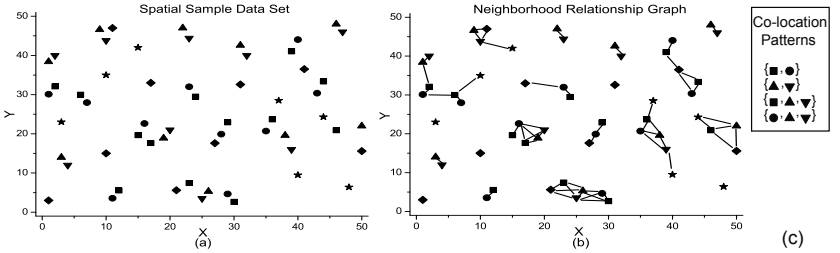

Fig. 1. An illustrative example of spatial co-location pattern discovery

distance threshold may introduce irrelevant clique instances of candidate co-locations in density areas, resulting in these co-locations being over-estimated.

Given the above concerns, we propose to find the regional co-location patterns as an extension to the conventional two-phase approaches. Our motivation comes from (1) the inconsistency of neighborhood distances in a data space; that is, the distance between any two neighboring events varies across the space since the data densities vary from region to region in the space. This inconsistency motivates us to investigate a new measure "distance variation coefficient" (see Section 3) rather than distance threshold to drive the mining process; and (2) the existence of spatial heterogeneity which demonstrates that most geographic processes vary by locations [4], and indicates that the inconsistent co-location sets may be found from different regions [5]. For instance, back to our example of mobile commerce, the requesting patterns in business regions are usually different from those in tourist regions.

As a mining strategy, we propose to hierarchically merge spatial events into local regions in the form of NRGs followed by passing them to the second phase of the conventional approaches. To enable the regional co-location pattern discovery meaningful, we partially inherit from the conventional approaches the assumption that a reasonable region is supposed to have relatively consistent neighborhood distances. This means the distances among the data points and their neighbors inside the region vary within a small range. Besides, we assume that different regions have inconsistent co-locating information in terms of spatial heterogeneity. Given these, this paper addresses two problems: (1) identifying regions with consistent neighborhood distances and co-locating information; and (2) specifying an individual NRG for each identified region.

We adopt k-nearest neighbor graphs (kNNG) that capture more natural neighborhoods [6] to describe the consistency of neighborhood distances within a region. Similar to distance thresholds, predefining k value for each region may lead to under-estimation or over-estimation of co-locations. Instead, we introduce a new measure "distance variation co-efficient" to control the range of distance varying and automatically determine k value for each region. With kNNG, the regions' NRGs are naturally prepared.

We then are able to define the similarity of co-location information between adjacent regions by passing these graphs to the second phase of the conventional approaches. The definition of similarity is a measure about whether the corresponding regions share consistent co-location information and are qualified to be merged. Analogous to kNNG, the mutual k-nearest neighbors (MkNN) naturally capture the inter-connectivity of adjacent regions [7,8]. We adopt MkNN to exclude the real region boundaries when hierarchically merging the regions.

In summary, our contributions are as follows. (1) We propose a hierarchical mining framework to discover regional co-locations by considering both varieties of neighborhood distances and spatial heterogeneity. (2) We propose a novel "distance variation coefficient" to drive the mining process and determine an individual NRG for each region. (3) We evaluate our mining algorithm with experiments on a real world data set by comparing against the conventional approaches using distance thresholds.

The remaining sections are organized as follows. We first review the current research results on spatial co-location pattern mining in Section 2, followed by giving a formal description for the research problem and some basic definitions in Section 3. We then present our hierarchical mining framework to discover regional co-location patterns in Section 4, together with the experimental evaluation in Section 5 before concluding our work in Section 6.

2 Literature Review

Various algorithms have been proposed for spatial co-location pattern mining. They can be classified into four types as reviewed in the following.

One main type of these techniques are the aforementioned two-phase approaches, by which the NRG is determined by a predefined distance threshold. This general framework was first proposed by Shekhar *et al.* [9]. Within the framework, different algorithms such as joinless algorithm [2], synchronic sweep algorithm [10], and density-based algorithm [3], were proposed to improve the performance of mining process, especially the efficiency of collecting clique instances. For example, Xiao *et al.* [3] proposed to search the dense areas with high priorities so as to speed up the decision making of the algorithm. The approach was also used in spatio-temporal data sets by introducing a time factor as the time interval threshold [11].

As a distortion of the first type of approaches, the second type diversifies the objective of spatial co-location pattern mining. For example, it was extended to mine complex spatial co-location patterns (e.g., one-to-many, self-colocating, self-exclusive, and multi-feature exclusive) [12] and maximal co-location patterns [13]. Huang *et al.* [14] also adjusted the interest measure to treat the case with rare events. Yoo *et al.* [15] proposed to find the N-most prevalent co-location patterns.

The third type replaces the usage of distance threshold in the first phase. Huang *et al.* [16] proposed to use density ratio of different features to describe the neighborhood relationship together with a clustering algorithm. A buffer-based model [17] was also proposed to describe the neighborhood relationship for dealing with extended spatial objects such as lines and polygons. Sheng *et al.* [18] used the influence functions to model the spatial features. Among these work, a similar neighborhood related threshold or function has to be predefined by users.

These three types of techniques focus on the global co-location patterns. That is, the first and second types adopt a predefined distance threshold to determine the NRG, while the third type replaces it with a similar neighborhood related threshold. The fourth type assumes that the neighborhood distances are consistent. The data sets with various densities are not sophisticatedly treated in these work.

The fourth type of techniques discovers the regional co-location patterns. Celik *et al.* [19] straightforwardly applied the conventional approaches to a set of zones, where

a zonal space has to be specified by users. Eick *et al.* [5] adopted the prototype-based clustering [20] to find regional co-location patterns. The interestingness of co-locations was scored in its fitness function. As an input of this approach, every event has a vector value of all the features, in which each item needs to be a continuous type. However, this work did not explore the monotonic property of the interesting measure proposed by Shekhar *et al.* [9] which introduces the pruning techniques to the mining process. Moreover, this approach may not be applicable to the discrete type of inputs.

3 Problem Formulation

In this section, we will present the problem statement after some basic definitions related to regional co-location mining.

3.1 Basic Definitions

A spatial data set is an input of the spatial co-location pattern mining algorithm.

Definition 1 (Spatial data set). *A spatial data set has a set of non-spatial features* $F = \{f_1, f_2, \ldots, f_n\}$*, and consists of a set of spatial events* $E = \{E_1, E_2, \ldots, E_n\}$*, where* E_i $(1 \leqslant i \leqslant n)$ *is a set of events of feature* f_i*. Every event* $e_j \in E_i$ $(1 \leqslant j \leqslant |E_i|)$ *has a vector information of* $\langle$*feature type* f_i*, event ID* j*, spatial location (x, y)*$\rangle$*.* ∎

Given the neighborhood constraint, a spatial data set or part of it (one of its regions) can be converted into an NRG as the foundation for co-location discovery. The conventional approaches adopt the distance threshold to describe this neighborhood constraint which may lead to limitations as discussed in Section I. To get rid of those limitations, we adopt the kNNG to define the NRG in the following.

Definition 2 (Neighborhood relationship graph (NRG)). *The neighborhood relationship graph* $\mathcal{G}$ *is implemented by* k*NNG, in which each vertex represents a spatial event, and there is an edge connecting two vertices if they have different features and either of them is among the* k*NNs of the other one. The graph's vertices and edges are denoted as* $V(\mathcal{G})$ *and* $E(\mathcal{G}) = k$*NNG(*$V(\mathcal{G})$*), and each edge is assigned with a weight that is the Euclidean distance between two spatial events connected by an edge.* ∎

In our approach, each NRG corresponds to an individual region. In practice, kNNG can be calculated by firstly finding the kNN of spatial events and then filtering out the edges whose end points share the same feature type. In the context of NRG represented by kNNG, the neighborhood distance is the weight of an edge in the graph.

Given the above neighborhood constraint, the interestingness of prevalent co-locations within a region is defined by an interesting measure known as participation index which is in turn defined from participation ratio. We borrow the definitions [1,2] as follows.

Definition 3 (Participation ratio). *Given the co-location* C $(C \subset F)$*, its participation ratio of* f_i $(f_i \in C)$ *is defined as* $Pr(C, f_i) = \frac{\pi_{f_i}(table\_instance(C))}{table\_instance(f_i)}$*, where* π *is the relational projection operation with duplication elimination and* $table\_instance$ *is the collection of clique instances of co-locations or features, in each instance of co-locations the spatial events are neighbors to each other.* ∎

Definition 4 (Participation index). *The participation index of co-location C is $Pi(C) = \min_{f_i \in C}\{Pr(C, f_i)\}$, which measures the prevalence of C.* ∎

With a predefined prevalence threshold θ, the second phase of the conventional algorithms such as join [1] or joinless algorithm [2] can wisely discover for each region a set of co-locations of which the value of participation index is not less than threshold θ (i.e., $Pi(C) \geqslant \theta$). Moreover, due to the monotonic property of participation index, i.e., $Pi(C') \geqslant Pi(C) \forall\, C' \subset C$, pruning techniques such as *apriori* [21] may be introduced into the mining process.

As foregoing, we assume that a reasonable region has relatively consistent neighborhood distances, meaning that the edge weights in an NRG have small variation. In a usual application domain of co-location discovery, geographers and biologists care about the spatial patterns under specific spatial frameworks following clumped, random or uniform distribution [22]. In clustering applications, classic algorithms (e.g., k-means algorithm [20]) often assume that the clumped clusters are compact, implying that they have consistent neighborhood distances inside. As for the data sets with random or uniform distribution, the local regions are also reasonable to hold the consistency. Based on the above applications, we define the distance variation coefficient to investigate the neighborhood distances of regions as follows.

Definition 5 (Distance variation coefficient). *The distance variation coefficient of the NRG $\mathcal{G}$ is defined as $\Omega(\mathcal{G}) = \frac{\sigma(E(\mathcal{G}))}{\mu(E(\mathcal{G}))}$, where $\mu(\cdot)$ and $\sigma(\cdot)$ are statistical operations for calculating the mean value and standard deviation of the weights of all edges in $\mathcal{G}$.* ∎

Our mining framework allows us to hierarchically merge regions followed by finding their prevalent co-location sets. During the merging process, the range of distance varying is controlled by a corresponding distance variation threshold ϵ. The algorithm does not stop until the distance variation of every newly merged region is greater than ϵ, i.e., $\Omega(\mathcal{G}) > \epsilon$.

As another assumption w.r.t. spatial heterogeneity, we define the similarity of regions in the following. By gradually combining the most similar regions under the distance variation constraint, our algorithm guarantees the maximum consistency of co-locating information inside the regions.

Definition 6 (Similarity of NRGs). *Given the prevalence threshold θ and two NRGs $\mathcal{G}_1$ and $\mathcal{G}_2$, their similarity is defined as $\mathcal{R}(\mathcal{G}_1, \mathcal{G}_2) = \frac{1}{\mathcal{L}-1} \sum_{i=2}^{\mathcal{L}} J(\mathcal{C}_{1i}, \mathcal{C}_{2i})$, where $\mathcal{C}_{1i}$ is the set of size i co-locations of $\mathcal{G}_1$ each of which is prevalent (i.e., $Pi(\mathcal{C}_{1i}) \geqslant \theta$), $\mathcal{C}_{2i}$ the set of size i co-locations of $\mathcal{G}_2$ each of which is prevalent (i.e., $Pi(\mathcal{C}_{2i}) \geqslant \theta$), $J(\cdot)$ the Jaccard index defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and $\mathcal{L}$ the maximum size of the prevalent co-locations of either $\mathcal{G}_1$ or $\mathcal{G}_2$.* ∎

The size of a co-location C is the number of distinct features it contains, namely $|\{f_i | f_i \in C\}|$ [1]. We indicate $Pi(C_{1i}) \geqslant \theta$ if $Pi(C) \geqslant \theta \;\forall\; C \subset C_{1i}$. The Jaccard index is a statistic commonly used for comparing the similarity of data sets. We adopt it to investigate the intersection rate of prevalent co-locations between regions. The similarity function helps us decide the priority of candidates to be merged. Finally, we give the following definition to identify the regions with rare events as anomalies.

Definition 7 (Significant NRG). *Given a ratio threshold α, an NRG $\mathcal{G}$ and its corresponding region are significant if $|V(\mathcal{G})| \geqslant \alpha|E|$, where $|E|$ is the total number of spatial events.* ∎

3.2 Problem Statement

With the above definitions, we give a formal description of regional co-location pattern discovery in the following.

Given: (1) A spatial data set including a set of features F and a set of their spatial events E; (2) a prevalence threshold θ; (3) a distance variance threshold ϵ; and (4) a ratio threshold α.

Find: Regional co-location patterns, i.e., (1) a set of regions $\Gamma = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m\}$ in which each element is represented as a kNNG, where $V(\mathcal{G}_i) \subset E$, $E(\mathcal{G}_i) = k\text{NNG}(V(\mathcal{G}_i))$ $(1 \leqslant i \leqslant m)$; and (2) a set of prevalent co-locations C for each region $\mathcal{G}_i$ where $Pi(C) \geqslant \theta$.

Constraints: Consistent neighborhood distances and consistent co-locating information within regions, i.e., (1) $\Omega(\mathcal{G}_i) \leqslant \epsilon$ and $|V(\mathcal{G}_i)| \geqslant \alpha|E|$ $(1 \leqslant i \leqslant m)$; and (2) minimizing the similarity between adjacent regions $\mathcal{R}(\mathcal{G}_i, \mathcal{G}_j)$ $(1 \leqslant i, j \leqslant m)$.

4 Regional Co-location Mining Algorithm

In the next, we present our algorithm for mining regional co-location patterns. The algorithm is carried out by three steps as follows. (1) In Step 1, a set of initial regions are formed by assigning each of them an MkNN edge with $k = 1$ or a single event that does not participated in those mutual edges. (2) In Step 2, the algorithm iteratively merges the similar regions under the constraint of distance variation. Step 2 is achieved by two sub-steps as follows. (2.1) In Step 2.1, in each iteration, k is increased by one and every newly generated MkNN edge links two of the current regions which compose a merging candidate; and (2.2) in Step 2.2, we decreasingly sort the merging candidates by their similarity values and sequentially merge them if the constraint of distance variation is satisfied. (3) In Step 3, the significant regions represented as kNNGs are returned together with their prevalent co-locations which are discovered by the second phase of join algorithm [1] in each region.

We illustrate the process of each iteration in Figure 2, in which Figure 2(a) shows four significant regions as indicated by circles each of which has a prevalent co-location in the $(k-1)$th iteration, and three MkNN edges found indicated by solid lines in the kth iteration; and Figure 2(b) shows a new region 5 which is the merge result of regions 1 and 3. Algorithm 1 presents the pseudo code of the mining process.

Step 1. Algorithm 1 first initializes the value of k as one, and sets the graph set Γ and the candidate set S empty (lines 1–2). It then finds a set of MkNN edges with $k = 1$, which are disjoint to each other since any event has at most one mutual neighbor. For these mutual edges, we also require that the end points of them have distinct feature types (line 3). Naturally, a part of initial regions are formed by assigning a found edge

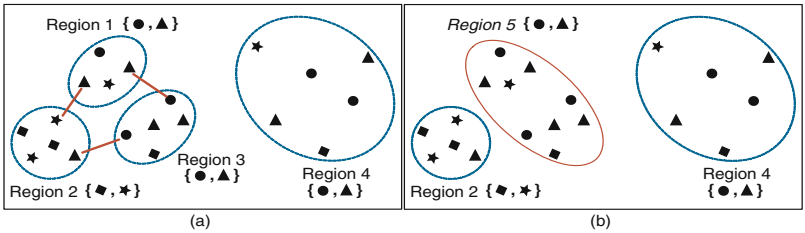

Fig. 2. An example of the mining process in each iteration

Inputs: A data set including E and F, parameters θ, ϵ and α.
Outputs: A set of kNNGs Γ, a set of corresponding co-locations C.
Method:
1: $k = 1$; //k of kNN
2: $\Gamma = \varnothing$; $S = \varnothing$; //S is a set of merging candidates
3: N_k = MkNN(E, F, k); //N_k is a set of MkNN edges of E
4: Initialize each edge $\ell \in N_k$ as an individual graph $\mathcal{G}$ in Γ;
5: Initialize each $e \in E \backslash V(N_k)$ as an individual graph $\mathcal{G}$ in Γ;
6: **while** $|\{\mathcal{G} | \mathcal{G} \in \Gamma$ **and** $V(\mathcal{G}) \geqslant \alpha |E|\}|$ is changed
7: $k = k + 1$; // increase k by one
8: N_k = MkNN(E, F, k); //update the mutual k-nn edges
9: **for** each $\ell_{e_1,e_2} \in N_k \backslash N_{k-1}$ //e_1 and e_2 are end points of edge ℓ
10: **if** $e_1 \in V(\mathcal{G}_1)$ **and** $e_2 \in V(\mathcal{G}_2)$ //link two regions
11: **then** $S = S \cup \mathcal{R}(\mathcal{G}_1, \mathcal{G}_2)$; //append to candidate set without duplication
12: **end for**
13: sort S decreasingly by similarity value R;
14: **for** each $\mathcal{R}(\mathcal{G}_1, \mathcal{G}_2) \in \mathcal{S}$ //traverse candidates by sorted order
15: $V(\mathcal{G}) = V(\mathcal{G}_1) \cup V(\mathcal{G}_2)$; $E(\mathcal{G}) = k$NNG($V(\mathcal{G})$); //test merging
16: **if** $\Omega(\mathcal{G}) \leqslant \epsilon$ **or** $(V(\mathcal{G}_1) < \alpha)$ **and** $(V(\mathcal{G}_2) < \alpha)$ //variation not large
17: **then** $\Gamma = ((\Gamma \backslash \mathcal{G}_1) \backslash \mathcal{G}_2) \cup \mathcal{G}$ //replace $\mathcal{G}_1$, $\mathcal{G}_2$ with their merged region
18: **end for**
19: **end while**
20: Calculate co-locations C for each $\mathcal{G} \in \Gamma$ with prevalence threshold θ;
21: **return** Γ & C; //output significant regions and their co-locations

Algorithm 1. Discovery of regional co-location patterns

to each of them (line 4); whereas the rest part is formed by assigning each region with a single event that does not participate in any of those edges (line 5).

Step 2.1. It then starts to merge the current regions iteratively. The process does not terminate until the significant region can be merged, which means that the number of significant regions does not change (line 6). In each iteration, the algorithm increases k by one and updates the set of MkNN edges (lines 7–8). For each newly generated edge, a merging candidate is formed if it connects two of the current regions and the similarity is calculated (lines 9–12).

Step 2.2. Given a set of merging candidates, the algorithm decreasingly sorts them by their similarity values (line 13), and sequentially investigates each candidate whether

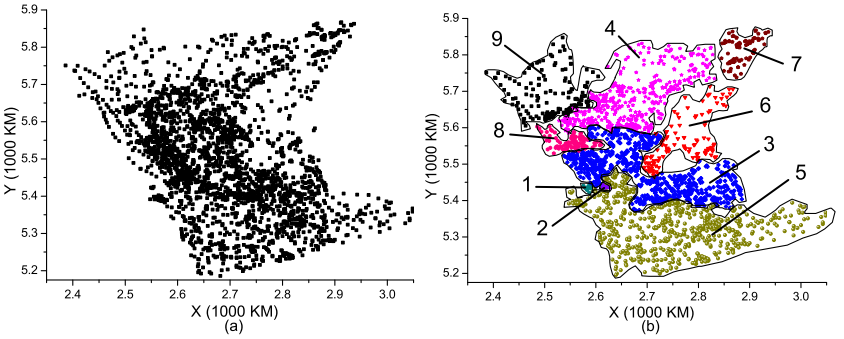

Fig. 3. An example of regional co-location pattern mining

its two regions have not yet been merged (lines 14–15). Each candidate is investigated by a test merge and an evaluation of the merged region with the constraint of distance variation. If the constraint is satisfied or both of two regions of the candidates are still insignificant, the algorithm replaces them with the newly merged region (lines 16–19).

Step 3. Based on a set of final regions, the algorithm calculates a set of co-location patterns for each of them if it is significant with small variation, and uses the second phase of join algorithm to find prevalent co-locations (line 20). Otherwise, the region is discarded since it has rare events which can be regarded as an anomaly. Finally, a set of significant regions and their prevalent co-location sets are returned to users (line 21).

5 Experimental Evaluation

Based on a real world data set, we evaluate the mining results of our approach against the conventional ones, and study the trends of several statistics in our mining process.

5.1 Description of Real Data Set

Figure 3(a) shows a real world data set (details shown in Table 1) we use for experiments. It is available at the Digital Chart of the World (DCW) Data Server [23] for research on spatial co-location discovery (see e.g.,[10,18]). Its spatial events have location information of latitudes and longitudes. Moreover, they are classified by distinct types of landmarks, such as drainage, land cover, and populated place. It is obvious that the data set includes various densities. In the data set, the geographic coordinates are transferred to projection coordinates using Universal Transverse Mercartor projection.

5.2 Mining Results

In this set of experiments, we run both our regional mining algorithm and the join algorithm on the DCW data set with parameters $\epsilon = 0.6$, $\theta = 0.6$ and $\alpha = 0.005$. Figure 3(b) illustrates our mining results (details shown in Table 2) for the data set shown in Figure 3(a), in which we identify nine regions and some regional co-locations.

Table 1. The data set of US Minnesota state in DCW

No.	Landmark Type	abbr.	# of events	No.	Landmark Type	abbr.	# of events
1	Aeronautical Point	Ap	86	5	Hypsography	Hy	72
2	Cultural Landmark	Cl	103	6	Hypsography Supplemental	Hs	687
3	Drainage	Dr	6	7	Land Cover	Lc	28
4	Drainage Supplemental	Ds	1338	8	Populated Place	Pp	517

Table 2. Information of found regions in DCW data set

Regions	$\boldsymbol{R}_1$	$\boldsymbol{R}_2$	$\boldsymbol{R}_3$	$\boldsymbol{R}_4$	$\boldsymbol{R}_5$	$\boldsymbol{R}_6$	$\boldsymbol{R}_7$	$\boldsymbol{R}_8$	$\boldsymbol{R}_9$
Number of events	15	14	866	430	727	128	94	96	157
Average neighborhood distance (km)	10.54	8.12	16.74	22.19	24.91	3.63	3.49	2.78	4.22
Distance variation	0.51	0.40	0.58	0.60	0.57	0.60	0.55	0.59	0.60
Number of co-locations (sizes 2, 3, and 4)	3, 1, 0	1, 0, 0	3, 1, 0	10, 10, 3	7, 3, 0	15, 13, 4	8, 3, 0	10, 5, 1	9, 2, 0

Table 3. Value of participation index of co-locations in DCW data set

Algorithms	$Pi(\{\text{Ds,Hs,Pp}\})$	$Pi(\{\text{Ap,Ds,Pp}\})$	$Pi(\{\text{Hs,Cl,Pp}\})$	$Pi(\{\text{Ap,Cl,Pp}\})$
join(30km)	0.65	Null	Null	Null
join(35km)	0.72	0.70	Null	Null
join(40km)	0.77	0.78	0.64	Null
RCMA	0.98(R_3),0.96(R_4) 0.69(R_5),0.90(R_6)	0.66(R_5),0.71(R_6) 0.86(R_8),0.75(R_9)	0.71(R_4) 0.90(R_5)	0.81(R_6)

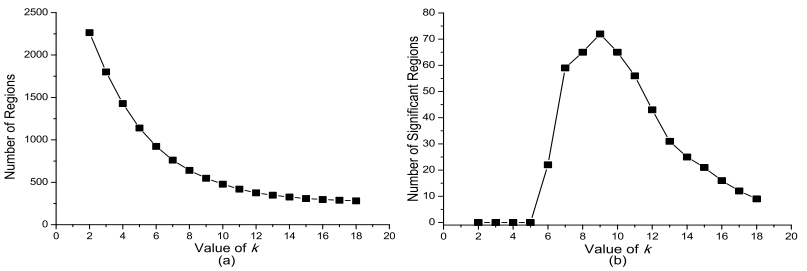

Fig. 4. Impact of k values to (a) the number of regions, and (b) the number of significant regions

In Table 3, we select four co-locations to demonstrate the difference between these two algorithms. The join algorithm discovers only the co-location {Ds, Hs, Pp} with a small distance threshold (30km). When the value of distance threshold increases, the other two co-locations ({Ap, Ds, Pp} and {Hs, Cl, Pp}) are detected. However, these regional co-locations are over-estimated to be globally prevalent. By contrast, our algorithm can detect them in their corresponding regions. We also find the co-location

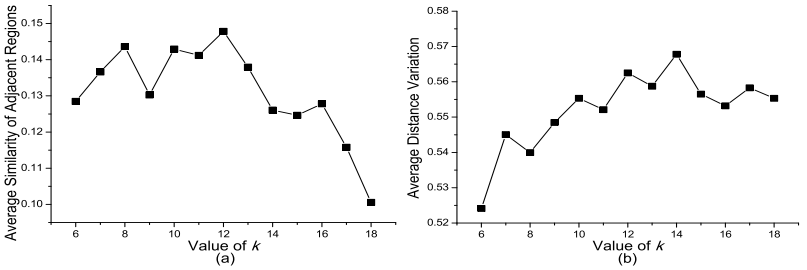

Fig. 5. Impact of k values to (a) the average similarity of adjacent regions, and (b) the average distance variation

{Ap, Cl, Pp} in a single region which is under-estimated by the conventional approaches. By comparing with the previous research results [18], our mining results include their discovered co-locations, and can assign them to the corresponding regions.

5.3 Evaluation of Regional Mining Process

In what follows, we study the trends of four statistics in our mining process on the DCW data set with the information of the number of regions, the number of significant regions, the average similarity of adjacent regions and the average distance variation.

Number of Regions. As can be seen from Figure 4(a), the number of regions decreases with the increase of k value. This is because the small regions are merged into larger ones. The decreasing is fast when k is small, while most of the regions are insignificant and the merging condition is easy to satisfy. The fast rate of decreasing also indicates that a relatively small k value can sufficiently describe the neighborhood relationship of spatial events, as well as the regional structure of space.

Number of Significant Regions. At the initial stage of mining process, there are hardly any significant regions as illustrated in Figure 4(b). Then the number of significant regions increases in a sudden when $k = 6$. This is because many insignificant regions are closing to the ratio line and become significant. The remaining set of insignificant regions naturally become anomalies which are hard to merge due to different neighborhood distance from their adjacent regions. After that, the number of significant regions decreases when the similar regions are sequentially merged.

Average Similarity of Adjacent Regions. To determine whether two regions are adjacent, we calculate each region a minimum bounding rectangle (MBR) and extend its width and height by 10%. Two regions are regarded as adjacent if they have an overlap between their MBRs. According to Definition 6, we calculate the average similarity of adjacent regions as shown in Figure 5(a). At the beginning of the process the similarity fluctuates, and then decreases rapidly. To explain the fluctuation, we tentatively test the relationship of k and the distance variation in Figure 6. We continually generate a set of kNNGs of the DCW data set by increasing k value from one to larger values, then calculate the value of distance variation for each generated graph. As can be concluded, the distance variation of the graph is large with small k. It decreases as k value

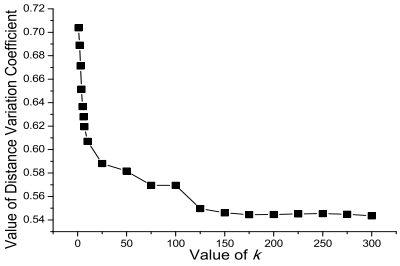

Fig. 6. Impact of k values to the distance variation of kNNG

increases. This is because small value of k indicates a large gap of the neighborhood distance between the events in dense and sparse areas. When k becomes larger, the edges with larger weights are involved and reduce the gap. Thus, at the beginning of the merging process, the regions which have different neighborhood distance are not merged even if they have consistent co-locating information. With larger k values, the distance variation of regions becomes smaller, while the consistency of co-locating information generally becomes a primary reference for merging. This explains the early fluctuation because the impact of distance variation and co-locating information match each other, and the later on drop of the similarity because the co-locating information competes as the primary reference.

Average Distance Variation. Figure 5(b) shows the implication between the average distance variation of regions and k values. With slight fluctuation, the average distance variation generally becomes large. The trend of curve verifies our explanation that the impact of distance variation and co-locating information match each other at the initial stage, and then the latter becomes the primary reference for merging.

6 Conclusion

We have discussed the limitations of conventional approaches to mining spatial co-locations using distance thresholds, especially for data sets with various magnitudes of neighborhood distances. To get rid of those limitations, we have proposed a hierarchical mining framework to discover regional co-locations accounting for both varieties of neighborhood distances and spatial heterogeneity. By adopting kNNG instead of distance thresholds, we have proposed a novel "distance variation coefficient" to drive the mining process and determine an individual NRG for each region. With rigorous experiments on a real world data set, we have demonstrated that our framework has been effective for the discovery of regional co-location patterns.

Acknowledgement. This work is partly supported by National Key Technologies R&D Program of China under Grant No. 2011BAD21B02 and MOE-Intel IT Research Fund of China under Grant No. MOE-INTEL-11-06, in which Chiew's work is partly supported by National Natural Science Foundation of China under Grant No. 61272303.

References

1. Huang, Y., Shekhar, S., Xiong, H.: Discovering colocation patterns from spatial datasets: A general approach. IEEE Transactions on Knowledge and Data Engineering 16(12), 1472–1485 (2004)
2. Yoo, J.S., Shekhar, S.: A joinless approach for mining spatial colocation patterns. IEEE Transactions on Knowledge and Data Engineering 18(10), 1323–1337 (2006)
3. Xiao, X., Xie, X., Luo, Q., Ma, W.Y.: Density based co-location pattern discovery. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, USA, November 5-7, pp. 1–10 (2008)
4. Miller, H.J., Han, J.: Geographic Data Mining and Knowledge Discovery. CRC Press, New York (2009)
5. Eick, C.F., Parmar, R., Ding, W., Stepinski, T.F., Nicot, J.P.: Finding regional co-location patterns for sets of continuous variables in spatial datasets. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Irvine, USA, November 5-7, pp. 30:1–30:10 (2008)
6. Karypis, G., Han, E.H., Kumar, V.: CHAMELEON: Hierarchical clustering using dynamic modeling. IEEE Computer 32(8), 68–75 (1999)
7. Brito, M.R., Chavez, E.L., Quiroz, A.J., Yukich, J.E.: Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Statistics & Probability Letters 35(1), 33–42 (1997)
8. Ding, C., He, X.: K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. In: Proceedings of the ACM Symposium on Applied Computing, Nicosia, Cyprus, March 14-17, pp. 584–589 (2004)
9. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (eds.) SSTD 2001. LNCS, vol. 2121, pp. 236–256. Springer, Heidelberg (2001)
10. Zhang, X., Mamoulis, N., Cheung, D.W., Shou, Y.: Fast mining of spatial collocations. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, August 22-25, pp. 384–393 (2004)
11. Yoo, J.S., Shekhar, S., Kim, S., Celik, M.: Discovery of co-evolving spatial event sets. In: Proceedings of the 6th SIAM International Conference on Data Mining, Bethesda, USA, November 20-22, pp. 306–315 (2006)
12. Munro, R., Chawla, S., Sun, P.: Complex spatial relationships. In: Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, USA, December 19-22, pp. 227–234 (2003)
13. Wang, L., Zhou, L., Lu, J., Yip, J.: An order-clique-based approach for mining maximal co-locations. Information Sciences 179(19), 3370–3382 (2009)
14. Huang, Y., Pei, J., Xiong, H.: Mining co-Location patterns with rare events from spatial data sets. GeoInformatica 10(3), 239–260 (2006)
15. Yoo, J.S., Bow, M.: Mining spatial colocation patterns: a different framework. Data Mining and Knowledge Discovery 24(1), 159–194 (2012)
16. Huang, Y., Zhang, P., Zhang, C.: On the relationships between clustering and spatial co-location pattern mining. International Journal on Artificial Intelligence Tools 17(1), 55–70 (2008)
17. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoo, J.S.: A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena, USA, April 22-24, vol. 89, pp. 78–89 (2004)

18. Sheng, C., Hsu, W., Li Lee, M., Tung, A.K.H.: Discovering spatial interaction patterns. In: Haritsa, J.R., Kotagiri, R., Pudi, V. (eds.) DASFAA 2008. LNCS, vol. 4947, pp. 95–109. Springer, Heidelberg (2008)
19. Celik, M., Kang, J.M., Shekhar, S.: Zonal co-location pattern discovery with dynamic parameters. In: Proceedings of IEEE International Conference on Data Mining, Omaha, USA, October 28-31, pp. 433–438 (2007)
20. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D., Steinberg, D.: Top 10 algorithms in data mining. Knowledge and Information Systems 14(1), 1–37 (2008)
21. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile, September 12-15, pp. 487–499 (1994)
22. Wang, J.: Spatial Analysis. Science Press, Beijing (2006)
23. Digital Chart of the World (2010), `http://www.maproom.psu.edu/dcw/`