Spatial Co-location Pattern Mining

Venkata M. V. Gunturi[(✉)]

IIT-Ropar, Rupnagar, India
gunturi@iitrpr.ac.in

Abstract. Given a spatial dataset containing instances of a set of spatial Boolean feature-types, the problem of spatial co-location pattern mining aims to determine a subset of feature-types which are frequently co-located in space. Spatial Co-location patterns have a wide range of applications in the domains such as ecology, public health and public safety. For instance, in an ecological dataset containing event instances corresponding to different bird species and vegetation types, spatial co-location patterns may revel that a particular species of birds prefer a particular kind of trees for their nests. Similarly, in a crime dataset, spatial co-location may revel a pattern that drunk-driving cases are co-located with bar locations. This article presents a gentle introduction to spatial co-location pattern mining. It introduces a well studied interest measure called *participation index* for co-location mining and, then discusses an algorithm to determine patterns having high participation index in a spatial dataset.

1 Introduction

Widespread use of spatial computing technologies [10,19–21,24] has lead to increasing interest in mining interesting and non-trivial patterns from spatial data. Over the years several works have made progress towards this end (e.g., [5,7,9,11–13,15,22,23]) by exploring different aspects of the problem of finding patterns from data which is embedded in geographic space.

One of influential results in the area of pattern mining for geographic space includes the work done in the area of *Spatial Co-Location* pattern mining. Given a spatial dataset containing feature-instances of spatial boolean feature-types, the problem of spatial co-location pattern mining aims to determine a subset of feature-types whose instances occur together frequently. Mining spatial co-location patterns has a wide range of applications in domains such as ecology, criminology and public health. Table 1 illustrates some sample patterns.

This article presents a gentle introduction to the problem of spatial co-location pattern mining. Section 2 introduces some basic concepts and presents the problem in a formal manner. In Sect. 3, we present a traditional algorithm for mining co-location patterns. Section 4 formally establishes the anti-monotonicity property of the participation index, a popular interest measure used in co-location pattern mining. Finally in Sect. 5, we summarize some of the recent results in the area and conclude in Sect. 6.

Table 1. Sample spatial co-location patterns in various spatial datasets.

| Sno	Feature-type A	Feature-type B	Interpretation
1	Nile Crocodile location	Egyptian Plover bird location	Nile crocodile and the plover birds have a symbiotic relationship
2	Bar locations	Reports of Drunk-driving	This pattern may be observed near the bar closing times [15]. Perhaps the bar customers drive back after bar closing
3	Locations of patients with respiratory problems	High particulate matter locations	Perhaps the air pollution contributes to the respiratory problems

2 Basic Concepts and Problem Definition
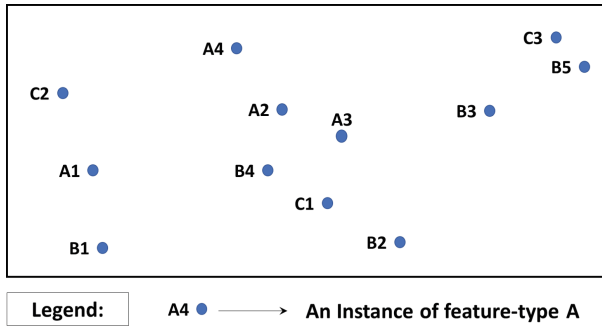
Fig. 1. A sample spatial dataset.

Definition 1. *Spatial Boolean Feature (f^i): comprises of those features which can be precisely defined in boolean form (i.e., present or absent) for any given location (x, y). Examples include, ecological features such as presence of bird nest, a particular type of tree, a particular animal species, etc. Spatial boolean features can also be used to define a particular type of spatial event. Examples of this category include, incidence locations of events such as forest fires, a particular disease, etc. In case one is interested in studying spatial aspect of diseases, the event location is the spatial location (home or office) of the patient. Note that in a dataset, each of the feature types f^i's would be associated a set of locations $\mathcal{L}_{f^i}$ where the particular feature type f^i occurs.*

Definition 2. *Spatial Co-Location Pattern (C_i): Given a set of Spatial Boolean feature-types $\mathcal{F} = \{f^1, f^2, f^3, \ldots, f^i\}$, a Spatial co-location C_i is defined as a non-empty subset of given feature-types $\mathcal{F}$, i.e., $C_i = \{f^1, f^2, \ldots, f^\gamma\}$ and $C_i \subseteq \mathcal{F}$.*

Definition 3. *Row instance of Spatial Co-location pattern C_i: Given a spatial[1] Co-location pattern $C_i = \{f^1, f^2, \ldots, f^\gamma\}$, a row instance of C_i is defined a particular collection of the instances (occurrences) of the feature types in C_i. In other words, a row instance of C_i is a collection of some specific features-instances of the feature-types in C_i. The key aspect to note is that, for any particular collection of feature-instances $\mathcal{G}(C_i) = \{G_{f^1}, G_{f^2}, \ldots, G_{f^\gamma}\}$ to become a valid row instance of C_i, each of the feature-instances in $\mathcal{G}(C_i)$ should be a neighbor of all the other feature-instances in $\mathcal{G}(C_i)$. In other words, $\forall$ p and q where $G_{f^p} \in \mathcal{G}(C_i)$ and $G_{f^q} \in \mathcal{G}(C_i)$, the feature-instance G_{f^p} and G_{f^q} are spatial neighbors (as per a given relation $\mathcal{R}$). Note that the notion of spatial neighbor is formally defined along with the input of the problem. Typically, they are defined using Euclidean, Geodetic or Network distance thresholds.*

Definition 4. *Table instance of Co-location pattern C_i: A collection of all the valid row instances of a Co-location pattern C_i are called as a table instance of C_i.*

Definition 5. *Participation ratio $Pr(f^\alpha, C_i)$ of a feature type f^α in a Co-location pattern C_i is the fraction of instances of the feature-type f^α which participate in any row instance of the spatial Co-location pattern C_i. More formally, $Pr(f^\alpha, C_i)$ is represented by the following equation:*

$$Pr(f^\alpha, C_i) = \frac{\#\ distinct\ instances\ of\ f^\alpha\ participating\ in\ C_i}{\#\ instances\ of\ f^\alpha}$$

Definition 6. *Participation index of a Co-location $C_i = \{f^1, f^2, \ldots, f^\gamma\}$ ($C_i \subseteq \mathcal{F}$) is defined as $\min_{f^j \in C_i}\{Pr(f^j, C_i)\}$. In other words, participation index of a Co-location pattern C_i is the minimum of all participation ratios' of its component feature-types.*

2.1   Problem Definition

The problem of spatial co-location pattern mining is defined as follows:
Input:

1. A set of N Spatial Boolean feature-type $\mathcal{F} = \{f^1, f^2, f^3, \ldots, f^N\}$.
2. A set of M feature-instances $P = \{p_1, p_2, \ldots, p_M\}$. Each $p_i \in P$ is a tuple <instance-id, spatial feature-type f^θ, location l>. Here, the feature-type $f^\theta \in \mathcal{F}$.
3. A neighbor relation $\mathcal{R}$. Given any two feature-instances p and q, the neighbor relation outputs "1" or "0" depending on whether p and q are deemed to be spatial neighbors or not. $\mathcal{R}$ is symmetric and reflexive.
4. Minimum prevalence threshold θ of the participation index.

Output:

1. Co-locations patterns whose participation index value is greater than θ.

---

[1] Whenever the context is clear, we drop the keyword "Spatial" from "Spatial Co-location" to maintain clarity of text.

3   Co-location Miner Algorithm

This section describes an elemental algorithm for determining spatial co-locations which internally uses spatial join operations. In its most basic form spatial co-location miner is quite similar to the traditional Apriori algorithm [2,3]. It generates candidate patterns in increasing size and at each stage, it uses patterns of size $k-1$ to generate patterns of size k. Algorithm 1 illustrates a pseudo-code of the algorithm.

Algorithm 1. Spatial Co-Location Miner

1: $P \leftarrow$ Prevalent size-1 co-location set along with their table instances.
2: Generate size-2 co-location rules
3: **for** size k co-location pattern $kin\{3, 4, \ldots, N\}$ **do**
4:    Generate candidate prevalent patterns of size k from the candidate patterns of size $k-1$.
5:    Generate table instances and prune based on neighborhood.
6:    Prune based on participation index threshold.
7: **end for**
8: Output the maximal sized co-location patterns.

In the first step (refer Algorithm 1), the algorithm generates spatial co-location patterns of size-1. These would be trivially be all the feature-types (singleton) present in the input dataset. For instance, consider the sample spatial dataset illustrated in Fig. 1. For this input dataset, the size-1 co-location patterns would simply be all the individual feature-types, A, B and C. Figure 2 illustrates the table instances of these trivial co-location patterns.

Candidate Spatial Co-location pattern		
A	**B**	**C**
A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	
	B5	

Instances

Fig. 2. Table instances of size-1 co-locations patterns.

Following this, in the second step (refer Algorithm 1), the algorithm generates size-2 candidate patterns. Table instances of these size-2 candidate patterns are generated by computing the spatial inner join of all the instances of all spatial features. The output of the spatial inner join procedure would be the pairs of feature-instances which satisfy the neighbor relation $\mathcal{R}$ (given in the input). In other words, it returns all the pairs of feature-instances which are deemed to be "neighbors" according to the neighbor relation $\mathcal{R}$. For this step, one may use sweeping based spatial join algorithms such as the one proposed in [4]. Figure 3 illustrates a sample result of the spatial join algorithm on the dataset shown in Fig. 1. Figure 4 shows the result in a tabular form along with the table instances of all the size-2 patterns ($< A, B >$, $< B, C >$, $< A, C >$). At this stage, the algorithm computes the participation indices of all each of candidate size-2 patterns. Figure 4 reports these indices at the bottom. Following this, any size-2

pattern whose participation index is less than the given threshold θ is pruned out.
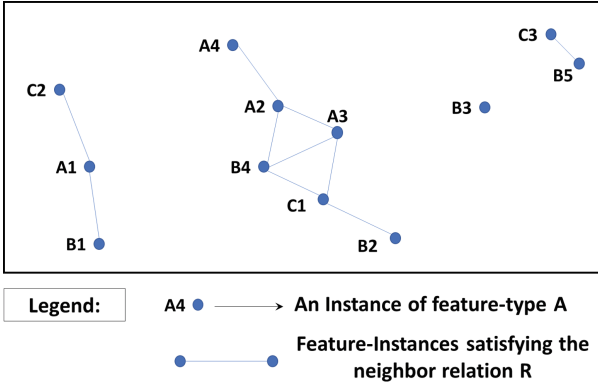

Fig. 3. Size-2 co-locations pattern instances.

Following the generation of size-2 candidate patterns, the algorithm enters a loop (line 3 Algorithm 1). In each iteration of the loop, at step-4, the algorithm generates candidate patterns of size k ($k = 3, 4, \ldots, N$) using the patterns of size $k-1$. This procedure is explained next through our example size-2 patterns shown in Figs. 3 and 4.

Candidate Spatial Co-location pattern		
A-B	**B-C**	**A-C**
A1-B1	B4-C1	A1-C2
A2-B4	B2-C1	A3-C1
A3-B4	B5-C3	
0.4	**0.50**	**0.60**

Participation Index

Fig. 4. Table instances of size-2 co-locations patterns.

Consider our size-2 patterns illustrated in Fig. 4. As can be noted, we have three size-2 patterns, viz., $< A, B >$, $< B, C >$, $< A, C >$. The algorithm takes pairs of size-2 patterns which have a common feature-type to create size-3 patterns. In our example, it would take $< A, B >$ and $< B, C >$ (or $\{< A, B > < A, C >\}$ or $\{< B, C > < A, C >\}$). Note that in any particular iteration of the loop, step-4 would take two size $k-1$ patterns p and q having $k-2$ common features. Following this the algorithm creates a size k pattern comprising of the following: (a) $k-2$ features common to both p and q, (b) $k-1$th feature of p, (c) $k-1$th feature of q. While creating a pattern of size k, it verifies whether all of its subsets of size $k-1$ have participation index greater than θ (minimum prevalence threshold) or not. For instance, in our example, while creating the pattern $< A, B, C >$, it checks if all size-2 subsets of $< A, B, C >$, i.e., $< A, B >$, $< B, C >$, $< A, C >$, have participation index greater θ. In case any of the subsets does pass the threshold, then the pattern $< A, B, C >$ would have been pruned. In our example, θ is set to 0.20.

After generating size k patterns from size $k-1$ patterns, the algorithm computes (at step-5) the table instances of all size k patterns. In our example, we have only one size k pattern, viz., $< A, B, C >$. For creating the table instances of size k patterns, the algorithms *joins* the table instances of size $k-1$ patterns. For our pattern $< A, B, C >$, these would be the table instances of the $< A, B >$ and $< B, C >$ (any two table instances can be joined). The row instances of $< A, B >$ and $< B, C >$ are processed in the followed way. Assume a row instance $< A_x, B_y >$ of $< A, B >$ and a row instance $< B_y, C_z >$ of $< B, C >$. These row instances would be joined on the value on the common attribute B to form a row instance of $< A, B, C >$. However, before accepting $< A_x, B_y, C_z >$ as a valid row instance of $< A, B, C >$, the algorithm tests if A_x and C_z adhere to the definition of spatial neighbor $\mathcal{R}$. If not, then the row instance $< A_x, B_y, C_z >$ is discarded. Figure 5 illustrates the size-3 pattern and its table instance for our example. In our example, the row instances $< A3, B4 >$ and $< B4, C1 >$ joined to form the row instance $< A3, B4, C1 >$ of $< A.B, C >$. Note that the join of row instances $< A2, B4 >$ and $< B4, C1 >$ would not be accepted as the $A2$ and $C1$ are not spatial neighbors (refer Fig. 3).
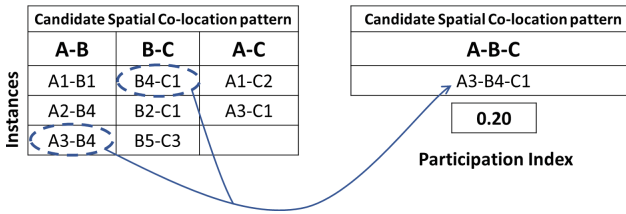

Fig. 5. Size-3 co-locations pattern instances.

Following the computation of the table instances of size k patterns, the algorithm determines its participation index. Any pattern with a participation index less that θ is pruned out. The loop on line 3 of Algorithm 1 iterates until the size of the pattern reaches the maximum possible of N (i.e., the number of feature-types in the dataset). In our example shown in Fig. 1, the maximum possible size of co-location pattern can be 3 ($< A, B, C$).

4 Theoretical Analysis

Lemma 1. *Participation index of a Co-location pattern $C_i = \{f^1, f^2, \ldots, f^\gamma\}$ is anti-monotonic in number of features.*

Proof. Consider a spatial Co-location pattern $C_i = \{f^1, f^2, \ldots, f^\gamma\}$. Participation ratio $Pr(f^k, C_i)$ of a feature type f^k in the Co-location pattern C_i is given by the following equation:

$$Pr(f^k, C_i) = \frac{\#\ distinct\ instances\ of\ f^k\ participating\ in\ C_i}{\#\ instances\ of\ f^k}$$

Participation index of C_i is defined as $\min_{f^j \in C_i} \{Pr(f^j, C_i)\}$. In other words, participation index of the Co-location pattern C_i is the lowest participation ratio among all its component feature-types. For establishing anti-monotonicity, we need to prove that participation index of another pattern $C_j = \{f^1, f^2, \ldots, f^\gamma, f^\beta\}$ which contains one additional feature-type f^β (in addition to $f^1, f^2, \ldots, f^\gamma$) is less than (or equal to) the participation index of C_i.

Without loss of generality, consider any feature type f^α in C_i. Now, a new feature-type f^β is being included in C_i (to create C_j). Consequently, the participation ratio of f^α would either decrease further or remain the same (at best). This happens because numerator in the participation ratio is likely to decrease or (at best) remain the same. Basically, now for a feature-instance of f^α to be included in a row instance of C_j it would have to be a spatial neighbor of one additional feature-type f^β (in addition to features $f^1, f^2, \ldots, f^\gamma$). So the number of distinct feature-instances of f^α in the numerator can either remain same or decrease. In other words, $Pr(f^\alpha, C_j) \leq Pr(f^\alpha, C_i)$. Using a similar argument, one can also show $\forall f^j \in C_i \; Pr(f^j, C_j) \leq Pr(f^j, C_i)$.

Now, consider the newly added feature-type f^β. We have the following two possibilities: (i) $Pr(f^\beta, C_j) < Pr(f^j, C_j) \forall f^j \in C_i$ or, (ii) there exists some $f^j \in C_i$ such that $Pr(f^\beta, C_j) > Pr(f^j, C_j)$. In the first case, clearly the participation index of C_j would be less than (or equal to) that of C_i. This is because $Pr(f^j, C_j) \leq Pr(f^j, C_i) \; \forall f^j \in C_i$. In the second case, the minimum of the participation ratios would now come from one of the features originally present in C_i. However, as we know $Pr(f^j, C_j) \leq Pr(f^j, C_i) \; \forall f^j \in C_i$, the participation index of C_j would again be less than (or equal to) that of C_i.

5   Other Works in Spatial Co-Location Pattern Mining

Spatial Co-locations were originally proposed in [22,25,26]. Over the years, several researchers have extended the traditional definition of co-location patterns along multiple aspects. Following is brief summary of some of those works.

[8,9,14,15] brought in the notion of time into co-location patterns. In these works, feature-instances have both spatial and temporal co-ordinates. [14,15] defined the co-location patterns as partially ordered sets. More specifically, the event instances participating in the co-location pattern were now related using a spatio-temporal neighbor relation. On the other hand, [8,9] explored co-location patterns in the domain of moving objects. Here, the input to the problem was a series of snapshots of locations of feature-types over a time window. Each snapshot was a collection of feature-instances at a particular time point. Given such a dataset, the goal was to determine co-location patterns which appear persistently over a set of time frames (but not necessarily in consecutive snapshots). In other words, for a set of feature-types to be considered as a valid pattern, they must form a valid co-location pattern (as per a certain threshold on participation index) at-least a certain number (given as threshold) of snapshots in a given time-window.

[5,6] explored the statistical significance aspect of co-location pattern discovery. The notion of statistical significance is of particular importance in co-location pattern mining. For instance consider the following scenario. We have two features α and β in the dataset which are independent of each other but abundant in quantity. In such a case, feature instances of α and β are likely to fall next to each other and the co-location pattern mining algorithm would return $\{\alpha, \beta\}$ as a valid pattern. Such spurious patterns can also arise if the feature-types α and β are auto-correlated (i.e., feature instances have a natural tendency to cluster). In such a case, if clusters of two independent features happen to overlap with each other, the participation index of the pattern $\{\alpha, \beta\}$ would increase dramatically. To this end, [5,6] proposed a novel Null model design where they take the underlying data-distribution characteristics of features to create a random dataset. Following this they compare (using p-values) the participation index of the pattern obtained in the real dataset against the random dataset to check if the obtained pattern was statistically significant. Basically, their approach creates several random datasets (having the same data distribution as the original real dataset), and counts the number random datasets which had a higher participation index for the pattern C_i generated by the algorithm on the real dataset. The higher the number, the greater is the chance that C_i was a random pattern.

[1] explored use of cross-k function [16] as an interest measure (instead of participation index) in co-location pattern mining. They extended the traditional co-location pattern mining by attempting to determine the *sub-region* of the given study region where the given two feature-types are associated to the maximum extent. The primary challenge addressed in this paper was the non-monotonic nature the cross-k function, whereby a larger area may have a smaller value of the interest measure but its subset (a smaller area inside the large area) may have a larger value of the interest measure.

[17,18] proposed algorithms which exploited high performance computing (GPUs) for determining spatial co-location patterns.

6   Conclusion

Spatial Co-location pattern mining delves into the problem of determining frequently co-occurring spatial feature-types in spatial dataset. These patterns are fundamentally different from the traditional association rules defined for transactions databases. Spatial co-locations have use-cases in several application domains such as ecology, public health and public safety. Several works explored this problem from multiple aspects. Participation index was one of the popular interest measures used in co-location pattern mining.

References

1. Agarwal, P., Verma, R., Gunturi, V.M.V.: Discovering spatial regions of high correlation. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 1082–1089 (2016)
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, SIGMOD 1993, pp. 207–216 (1993)
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994, pp. 487–499 (1994)
4. Arge, L., Procopiuc, O., Ramaswamy, S., Suel, T., Vitter, J.S.: Scalable sweeping-based spatial join. In: Proceedings of the 24th International Conference on Very Large Data Bases, VLDB 1998, pp. 570–581 (1998)
5. Barua, S., Sander, J.: Mining statistically significant co-location and segregation patterns. IEEE Trans. Knowl. Data Eng. **26**(5), 1185–1199 (2014)
6. Barua, S., Sander, J.: SSCP: mining statistically significant co-location patterns. In: Pfoser, D., et al. (eds.) SSTD 2011. LNCS, vol. 6849, pp. 2–20. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22922-0_2
7. Cao, H., Mamoulis, N., Cheung, D.W.: Mining frequent spatio-temporal sequential patterns. In: Fifth IEEE International Conference on Data Mining (ICDM 2005), pp. 82–89 (2005)
8. Celik, M., Kang, J.M., Shekhar, S.: Zonal co-location pattern discovery with dynamic parameters. In: Seventh IEEE International Conference on Data Mining (ICDM), pp. 433–438 (2007)
9. Celik, M., Shekhar, S., Rogers, J.P., Shine, J.A.: Mixed-drove spatiotemporal co-occurrence pattern mining. IEEE Trans. Knowl. Data Eng. **20**(10), 1322–1335 (2008)
10. Güting, R.H.: An introduction to spatial database systems. VLDB J. **3**(4), 357–399 (1994)
11. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In: Egenhofer, M.J., Herring, J.R. (eds.) SSD 1995. LNCS, vol. 951, pp. 47–66. Springer, Heidelberg (1995). https://doi.org/10.1007/3-540-60159-7_4
12. Li, X., Han, J., Lee, J.-G., Gonzalez, H.: Traffic density-based discovery of hot routes in road networks. In: Papadias, D., Zhang, D., Kollios, G. (eds.) SSTD 2007. LNCS, vol. 4605, pp. 441–459. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73540-3_25
13. Liu, Z., Huang, Y.: Mining co-locations under uncertainty. In: Nascimento, M.A., Sellis, T., Cheng, R., Sander, J., Zheng, Y., Kriegel, H.-P., Renz, M., Sengstock, C. (eds.) SSTD 2013. LNCS, vol. 8098, pp. 429–446. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40235-7_25
14. Mohan, P., Shekhar, S., Shine, J.A., Rogers, J.P.: Cascading spatio-temporal pattern discovery: a summary of results. In: Proceedings of the SIAM International Conference on Data Mining (SDM), pp. 327–338 (2010)
15. Mohan, P., Shekhar, S., Shine, J.A., Rogers, J.P.: Cascading spatio-temporal pattern discovery. IEEE Trans. Knowl. Data Eng. **24**(11), 1977–1992 (2012)
16. Ripley, B.D.: The second-order analysis of stationary point processes. J. Appl. Probab. **13**(2), 255–266 (1976)

17. Sainju, A.M., Aghajarian, D., Jiang, Z., Prasad, S.K.: Parallel grid-based coloca-tion mining algorithms on GPUs for big spatial event data. IEEE Transactions on Big Data (2018). https://doi.org/10.1109/TBDATA.2018.2871062
18. Sainju, A.M., Jiang, Z.: Grid-based colocation mining algorithms on GPU for big spatial event data: a summary of results. In: Gertz, M., et al. (eds.) SSTD 2017. LNCS, vol. 10411, pp. 263–280. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64367-0_14
19. Shekhar, S., Chawla, S.: Spatial Databases: A Tour. Prentice Hall (2003). (ISBN 013-017480-7)
20. Shekhar, S., Chawla, S., Ravada, S., Fetterer, A., Liu, X., Lu, C.T.: Spatial databases - accomplishments and research needs. IEEE Trans. Knowl. Data Eng. **11**(1), 45–55 (1999)
21. Shekhar, S., Feiner, S.K., Aref, W.G.: Spatial computing. Commun. ACM **59**(1), 72–81 (2015)
22. 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). https://doi.org/10.1007/3-540-47724-1_13
23. Wang, S., Huang, Y., Wang, X.S.: Regional co-locations of arbitrary shapes. In: Nascimento, M.A., et al. (eds.) SSTD 2013. LNCS, vol. 8098, pp. 19–37. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40235-7_2
24. Worboys, M., Duckham, M.: GIS: A computing perspective. CRC (2004). ISBN: 0415283752
25. Yoo, J.S., Shekhar, S., Celik, M.: A join-less approach for co-location pattern min-ing: a summary of results. In: Fifth IEEE International Conference on Data Mining (ICDM) (2005)
26. Yoo, J.S., Shekhar, S., Smith, J., Kumquat, J.P.: A partial join approach for min-ing co-location patterns. In: Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems, GIS 2004, pp. 241–249 (2004)