An Image Processing based Algorithm for Discovering Co-Location Patterns

Shahbaz Ahmad
National Textile University
Faisalabad, Punjab
Pakistan

Muhammad Asif
National Textile University
Faisalabad, Punjab
Pakistan

ABSTRACT
Spatial co-location patterns represents the subset of Boolean spatial features (e.g. Frontage roads, freeways) whose instances are often located in close geographic proximity. For instance, stagnant water founts and west Nile ailments are often co-located. The co-location pattern can be defined as an undirected connected graph in which every node represents a feature and every single edge denotes relationship (neighbourhood) between connecting features. Literature provides different approaches (including transaction based, join and join-less approaches) to discover co-location patterns. This paper proposes, implements and tests an image processing based algorithm to discover these patterns. The algorithm inputs minimum confidence measure (for statistical significance), neighbourhood distance threshold and set of Boolean spatial features, whose instances are represented as an image. It converts the image into binary image and then uses the concept of neighbourhood relationship (materialized using distance threshold) and confidence measure to mine the patterns. Furthermore, this paper provides implementation and testing of proposed algorithm in terms of time and space complexity.

General Terms
Algorithms, Image Processing, Spatial Data Mining

Keywords
Association rule mining, Co-Location Pattern discovery, Collocation Pattern, Image Processing, Spatial Association Pattern, Spatial Data mining

1. INTRODUCTION
The spatial Co-Location Pattern (CLP) represents the subset of Boolean Spatial Features (BSFs) whose instances are often located in close geographic proximity. For example, along with growth in mobile computing (such as mobile phones and PDAs); e-services are growing. To provide location-aware market promotions, one need to know popular e-services that are often located together. As another example, in field of ecology, researchers are very interested in discovery of co-occurrences of diverse BSFs, for example EI Niño, drought, extremely high precipitation, and significant drop in vegetation etc. [1, 2].

The co-location pattern P can be defined as an undirected connected graph in which every node represents a feature and every single edge denotes relationship (neighbourhood) between connecting features. Consider an example pattern having three nodes which are labelled as weather, timetabling, and ticketing. Also, it has two edges which connects weather with timetabling and ticketing with timetabling. A set of objects is said to be instance of pattern P iff it satisfies binary (neighbourhood) and unary (feature) constraints that are specified by P's graph. The instance of P is a set $\{f_1, f_2, f_3\}$, where $label(f_1)$ = *weather*, $label(f_2)$ = *timetabling* and $label(f_3)$ = *ticketing* are known as unary constraints and $dist(f_1, f_2) \leq \varepsilon$, $dist(f_2, f_3) \leq \varepsilon$ are named as spatial binary constraint.[1, 3, 4, 5]

To represent CLPs as graphs, three pattern representation (i.e. star, clique and generic) can be used as shown in figure 1. A variable labelled with feature f_i is merely indorsed to take instance of that feature as values. Set of variable pairs that fulfil the spatial relationship (i.e. constraint) in an effective pattern instance are connected by edge. In figure 1 representation only one spatial constraint (i.e. close to) is specified. However in general, any spatial constraint and relationship might label each edge.

There might be different CLPs such as point CLPs and line string CLPs. In a point CLP each BSF is represented by unique points on the graph (such as +, - and *). Symbiotic species might be a decent case to demonstrate a point CLP. For example, the Egyptian plover and Nile crocodile provides an excellent graphic (as shown in Figure 1) to show a point spatial co-location representation. A vigilant scan of figure 2 reveals that sets {'+', 'x'} and {'o', '*'} always be apt to locate together. Line string representation is second type to demonstrate a CLP. Highways and frontage roads in an urbanite road map can be considered a good example of *line-string patterns*. Figure 3 shows such a co-location pattern. Highways (such as Hwy100) and frontage roads, for example Norman dale road, are co-located. This paper propose and implements an image processing based algorithm for discovery of CLPs. The algorithm efficiently mine the point CLPs and it is based on reference feature centric model (discussed in section 2.3). We also, implemented this algorithm in MATLAB 2013b and tested it on test data set.

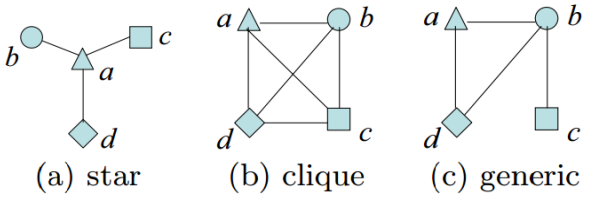

Fig 1. Co-location pattern graph representations

The rest of paper is organized as: section 2 describes brief background of problem and reviews some of methods to discover the CLP, section 3 describes the proposed algorithm and section 4 gives implementation and testing details of proposed algorithm.

2. BACKGROUND

The Co-Location Pattern (CLP) and rule discovery shares a part of Spatial Data Mining (SDM) process. The classical and SDM differs primarily on the basis of the data to be inputted, statistical base of the problems, output patterns, and computational process. The research activities in this field are generally focused on the output pattern category; explicitly spatial outliers, predictive models, clusters and spatial co-location rules.

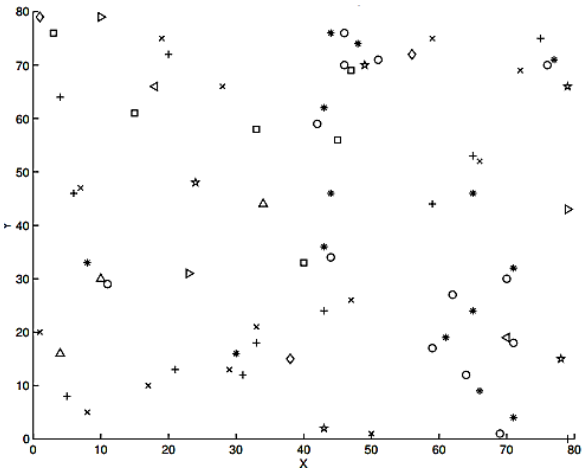

Fig 2: Point Co-location pattern

The association rule mining problems was initially discussed in [6]; which primarily deals with association rules that are established on spatial relationships (such as proximity, adjacency) of objects or events. To mine these rules, spatial dataset is transformed to transactional data using different algorithms such as reference feature centric model (discussed in section). Later, research activities in this field was shifted to mining *co-location patterns* that are feature centric sets having instances which are located in same geographic neighbourhood [1, 3, 7, 8]. Wherein [3, 7] focused on patterns where a complete graph is formed by the closeness relationships between features. While, [1] protracted this model to feature-sets with closeness relationships among any random pairs and recommended an effective algorithm for mining such patterns. [4, 9] extended the concept of co-locations for objects with extended shapes and objects such as polygons. Also, [10] deliberated the mining of co-location patterns which involve spatio-temporal topological constraints.

2.1 Basic Concepts and terms

Boolean Spatial Features[11] are geographical object types that are either present or absent at different localities in a two or higher (three) dimensional metric space, such as surface of earth. Some common examples of BSFs are classifications such as animal species, plant species, cancers, crimes, drought, business and types of roads.

Spatial association rules or Co-location rules are representations to deduce the occurrence of BSF in the neighbourhood of occurrences of other BSF(s). These rules are in the form of $X \rightarrow Y$ *(Z %)* in which X and Y are subset of Boolean spatial features and Z% shows the rule's conditional probability.

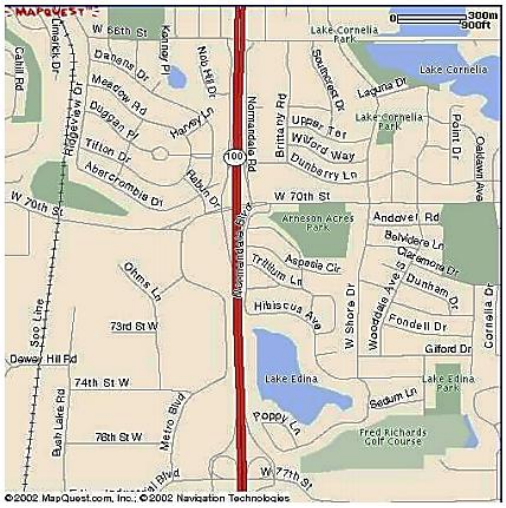

Fig 3: Line Co-location pattern

To measure statistical significance and value of discovered patterns, [3, 7] described some useful measure which are described below.

Participation ratio B of a feature f_j in pattern P is represented as *pr (P, f_j)* is fraction of instances of f_j. Conventionally, *pr(P, f_j)* = $|\varphi f_j\ (All\ instances\ of\ P)| \Big/ |instances\ of\ f_j|$. Here φ is the relational projection with duplicate elimination. For instance, let a co-location pattern P = {W, X, Y, Z} and W, X, Y, and Z have n_w n_x, n_y and n_z instances respectively. If n_W^P , n_X^P , n_Y^P and n_Z^P are the distinct instances of W, X, Y and Z respectively which participate in P, then the participation ratio of W, X, Y and Z will be $n_W^P \Big/ n_W$, $n_X^P \Big/ n_X$, $n_Y^P \Big/ n_Y$ and $n_Z^P \Big/ n_Z$ respectively.

Using this ratio, it can be said that f_j participates in pattern P's instance with confidence measure of *pr(P, f_j).*

Participation Index (PI) of a CLP *P* is defined as *PI(P)* = $min_k \{pr(P, f_k)\}$. For example P = {W, X, Y, Z} where participation ratios of W, X, Y and Z are $2/6, 2/5, 2/7$ and $1/9$ respectively. Then *PI(P)* = $1/9$.

PI shows the least probability that each time an instance $f_j \in P$ is present on map, and then it will participate in instance of P. Therefore, it could be utilized to describe the vigor of the pattern in inferring co-location of features.

Confidence of pattern P denoted as *conf(P)* is stated by the equation given below

$$conf(P) = maximum\ \{pr(f_j,P), f_j \in P\}$$

Confidence describes the ability of a P to drive co-location rule by participation ratio. If P's confident is at least the minimum confidence threshold, then at least on co-location rule (for f_j with *PI(f_j,p) = conf(P)*) can be generated.

2.2 Related Algorithms

There are different models that can be applied to investigate diverse field problems. Multiple algorithms are also used in discovery process. Algorithms and approaches available to discover CLP and rules possibly categorized into two main

classes or categories named as data mining and spatial statistics approaches. Which are discussed below

2.2.1 Spatial Statistics based approaches

In these type of approaches, relationship between various kinds of spatial features is described using measures of spatial correlation. Spatial correlation processes take account of the cross-K function with Monte Carlos simulation, spatial regression models and mean nearest-neighbour distance. It could be a computationally expensive to compute spatial correlation measures for all CLPs due to the reason that candidate subsets that are extracted from a huge collection of spatial BSFs is in exponential number.

2.2.2 Data mining approaches [2, 11, 14]

These approaches and algorithms can be sub divided in two classes namely, association rule-based method and clustering-based map overlay methods.

Clustering-based map overlay methods keeps each spatial attribute by way of map layer and ruminates spatial clusters (also called regions) of point data in every layer as aspirants for mining association among them.
Association rule based methods can be again subdivided into two types namely: distance based methods and transaction based methods.

Transaction based methods: These type of methods convert spatial datasets into transactional datasets. Then, it uses an A priori-like algorithms to mine spatial association rules just like association rule mining process. Transactions over spatial datasets might be defined using reference centric model (discussed in section 2.3) which enables to run A priori algorithm to mine association rules. There are some drawbacks of this approach, such as,

- Generalization of this model is non-trivial in the case where none of the reference feature is indicated.
- When making transactions about localities of all feature occurrences, duplicate counts for several candidate association may possibly result.
- When making transactions about localities of all feature occurrences, may result in duplicate counts for several aspirant associations.

Distance-based approaches are relatively novel. A couple of different approaches have been presented by different research groups. One proposes the participation index as the prevalence measure, which possesses a desirable anti-monotone property. Thus a unique subset of colocation patterns can be specified with a threshold on the participation index without consideration of detailed algorithm applied such as the order of examination of instances of a co-location. Another advantage of using the participation index is that it can define the correctness and completeness of co-location mining algorithms

2.3 Models to Discover CLPs

Depending on the focus of search [11] classifies the methods to find CLP in spatial datasets into three categories. These categories are *reference feature centric model*, *window centric model* and *event centric model.*
Reference feature centric model [6] The application domains that focus on specific BSF such as cancer are more suitable to reference feature centric model. Researchers in this field tries to find CLPs between this BSP and other task related features like asbestos or other substances. This model is mainly based on the concept of neighbourhood relationship to construct the transactions from provided datasets. It also use the measurements such as support and confidence to show the degree of interestingness.

For more elaboration, consider two features X and Y, and if X is taken as reference feature, and Y is said to be neighbour of X if it is close it. But the question that arises here is; how to state that Y is neighbour of X? For this purpose depending on type of application domain that is being investigated, Manhattan or Euclidean distance can be used. Then using any of the definition of distance, it could be declared that the features are neighbour or not. Thus by taking X (reference feature) all the other BSFs adjacent to X are used as transactions. Once dataset is materialized in transactional form as stated above; support and confidence of rule is calculated to extent of interestingness.

Window Centric model[3] Window centric model also called *data partitioning model* outlines apposite sized windows and then itemizes all potential windows as transactions. Essentially, whole large space is partitioned into small size windows and then rule discovery process focus on local CLPs which are confined by the boundaries of current window. This model take no concern about the patterns that are across multiple windows. Each window acts as a transaction and in this way it forms a new transactional dataset. The window centric process tries to discover features that seems together the maximum number of times in these transactions, alias, windows, i.e., using support and confidence measurements.
Event Centric Model[7, 15] This model is commonly associated to ecology domain problems, where researchers want to inspect explicit events such as EI Nino, drought etc. This model is aimed to discover the subsets of spatial features that are possible to happen in vicinity of a particular event type. A key assumption of this algorithm is that the neighbours are reflexive, that is, interchangeable. For example, if X is neighbour of Y, then Y is also a neighbour of X.

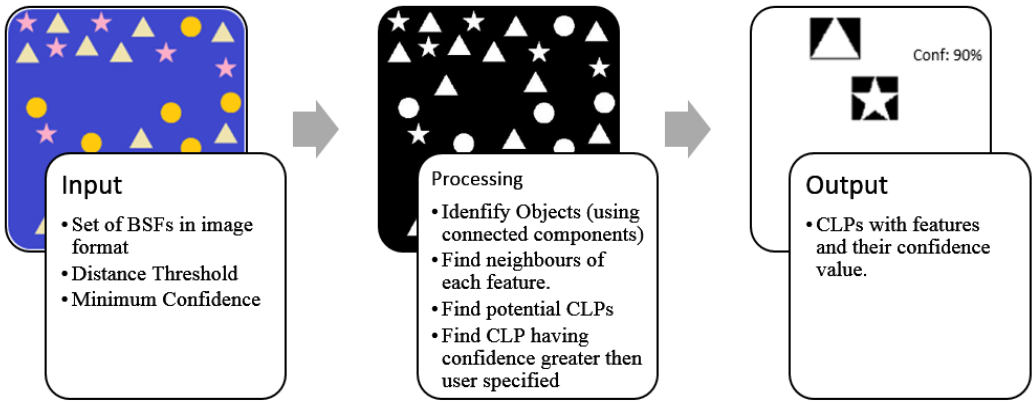

Fig 4: Overview of methodology

3. METHODOLOGY

This section presents a brief description of proposed algorithm. An overview of proposed algorithm is given in figure 4. The algorithm starts by inputting image which represents the occurrence of instances of BSFs, minimum distance threshold to materialize neighbourhood relationship and minimum confidence value to show statistically significant patterns. In the next step, the image is converted to binary form (if it is given in colour form) and all the objects in this image are identified. Then distance between each pair of points is calculated and objects having distance less than or equal to distance threshold are extracted. In next phase, by taking a feature as reference feature (as in reference feature centric model) its neighbours are identified, and it is considered as potential CLP. Next, prevalence and participation index of that CLP is calculated to check the confidence of the pattern; if its confidence is greater than user specified then it is shown in output and algorithm move to next reference feature.

Below a concise algorithm to solve the problem is given

Input:

1) SF = {A set of Boolean spatial features, which can be represented as filled geometric shapes such as star, triangle and circle etc.
2) The neighborhood relationship (i.e. a distance threshold for determining neighborhood relationship)
3) *min_pre* = prevalence threshold; *min_confi* = confidence threshold.

Output:

1) A set of all co-location rules having *confidence* ≥ *min_confi*

Variables:

1) img: $[m \times n]$ integer matrix; original image representing occurrences of BSFs.
2) bwimg: $[m \times n]$ integer matrix; binary image representing occurrences of BSFs.
3) cc: $[1 \times 1]$ structure with fields (*Connectivity*, *ImageSize*, *NumObjects*, *PixelIdxList*)
4) s: $[cc.NumObjects \times 1]$ structure with fields (*Centroid*, *ConvexArea*, *FilledImage*)
5) dist_thresh: Integer; distance threshold to materialize neighborhood relationship
6) min_dist: $[cc.NumObjects \times cc.NumObjects]$ integer matrix; to store distance between each point.
7) adjacent_objects and symb $[1 \times cc.NumObjects]$ matrices

Method:

1) img = read_image(path);
2) bwimg = convert_to_binary(bwimg,luminace_level)
3) cc = find_all_conncected_components(bw)
4) s = find_region_properties(cc, 'PixelIdxList', 'Centroid', 'ConvexArea')
5) min_dist = calculate_ euclidean_distance()
6) for each connected_component
   select components having distance less than dist_thresh
   end
7) for i=1:number_elements(refferenced_object_adjcnt_obj)
   identify_neighbour()
   Find_confidence_mesure()
   end
8) Display_CLP()

Method steps explained

1) As an input this algorithm requires a set of BSFs (the image in which BSFs are marked by distinct geometric point objects), distance threshold value (to materialize the neighborhood relationship). The set of BSFs with their location are converted to image format for further processing in this algorithm.

2) The algorithm works on binary image (a digital image that has only two possible pixel values i.e. 0 for black and 1 representing white pixel) format. It initially convert the image into binary image, if it is provided in other format. It converts the image to binary image using the concept of image thresholding, and this steps outputs an image (say BW) in which; all the pixels that have luminance greater than the specified level (*ll)* are replaced with value 1 (means a white pixel) and all other pixels are replace with value 0 (displaying black pixel). The luminance threshold value is calculated by Otsu's method, which selects the threshold value aiming to abate the intra-class variance of black and white pixels.

3) In the next phase, we found the Connectivity of the connected components (objects) in the binary image

(obtained in step 2). As a result of this step, we determined number of connected components with pixel value locations.

4) For each connected objects, a set of properties like: 'Centroid', 'Convex Area' and 'Pixel Id List' are computed.

a. 'Centroid' is a $1 \times k$ vector that represents the center point of mass of the region (or in other words center of the shape representing the BSF in image). Centroid is a value pair, in which the first value or element represents the x-coordinate (or horizontal co-ordinate) of the center of image, whereas second value or element represents the y-coordinate (or vertical coordinate). All the other elements and values of Centroid vector are in sequence of dimension.

b. 'Convex Area' is a scalar that represents the total number of pixels in identified segmented areas (in other words each shape that represent BSFs).

c. 'Pixel Id List' is a vector that contains linear indexes of the pixel distribution in the region.

5) Compute the distances between centroids of each pair of objects using Euclidean distance. The Euclidean distance is calculated by providing two values, first one is a $m_x \times n$ data matrix (say X), which is manipulated as m_x $(1 \times n)$ row vectors $x_1, x_2, \ldots, x_{mx}$, and second value is $m_y \times n$ data matrix(say Y), which is considered and manipulated as m_y $(1 \times n)$ row vectors $y_1, y_2, \ldots, y_{my}$, and the distance between these vectors is defined by

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)'$$

Distances (in pixels) between each pair of centroids is stored in $m \times n$ data matrix (say X) and $m_y \times n$ data matrix (say Y). The rows in matrix X and Y shows the observations and columns shows the variables.

6) Select each object/feature and calculate its distance from each object in the image.

7) For each reference object identify its neighbors (i.e. objects that have distance less than or equal to distance threshold).

a. Consider it as potential CLP

b. Calculate its prevalence measure, participation index and confidence

8) Check either a valid CLP and show it if it is valid

4. IMPLEMENTATION AND TESTING

This section describes implementation and experimental evaluation with respect to other algorithms of proposed algorithm. The algorithm is implemented in MATLAB R2013b, and it is tested using test data generated manually. We also compare execution time of proposed algorithm with join based and join-less based approaches. All the test were conducted on Intel 2.8 MHz Corei3 machine with 4 GB of memory. The summary of results is presented in figure 5 and 6.

As it can be seen in figure 5 that the algorithm outperforms then both (join and join-less) approaches till number of features are less than 40K. Its performance slightly slows down after number of features exceed 40K approximately, as compared to join-less approaches.

In second experiment, we check the execution time of proposed algorithm by changing the number of features and kept number of objects same (1K) and compared the execution time. Number of feature time slightly increases the execution time of all the algorithms. A details description is given in Figure 6.

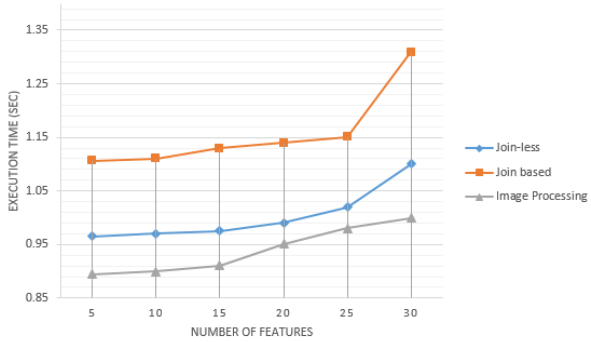

Fig 5: Execution time of different algorithms by varying number of features

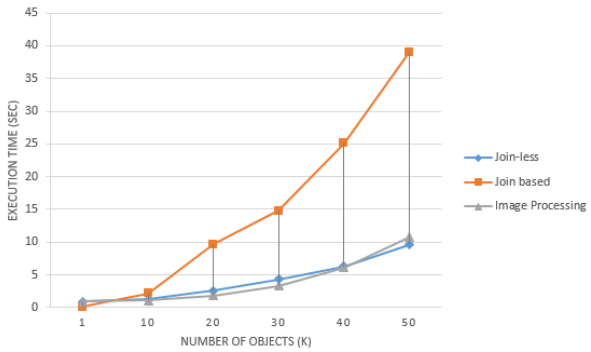

Fig 6: Execution time of different algorithms by varying number of objects

5. CONCLUSION AND FUTURE WORK

This paper proposed, tested and compared an image processing based algorithm to discover co-location patterns. This algorithm selects a BSF as in reference feature centric model and uses the distance threshold value to materialize neighbourhood relationship. It uses the prevalence and confidence measure to show statistical significance of discovered co-location patterns.

As for future work, image processing based algorithms can be extended to mine line string co-location patterns. It can also be extended to mine co-incidence patterns which includes frequently occurring events in same time period.

6. REFERENCES

[1] Zhang, X., et al. Fast mining of spatial collocations. in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004. ACM.

[2] Zala, M., et al., A Survey on Spatial Co-location Patterns Discovery from Spatial Datasets. arXiv preprint arXiv:1402.1327, 2014.

[3] Shekhar, S. and Y. Huang, Discovering spatial co-location patterns: A summary of results, in Advances in Spatial and Temporal Databases. 2001, Springer. p. 236-256.

[4] Adilmagambetov, A., O.R. Zaiane, and A. Osornio-Vargas, Discovering Co-location Patterns in Datasets with Extended Spatial Objects, in Data Warehousing and Knowledge Discovery. 2013, Springer. p. 84-96.

[5] Mohan, P., et al. A spatial neighborhood graph approach to regional co-location pattern discovery: A summary of results. in 19 th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS. 2011.

[6] Koperski, K. and J. Han. Discovery of spatial association rules in geographic information databases. in Advances in spatial databases. 1995. Springer.

[7] Huang, Y., et al. Mining confident co-location rules without a support threshold. in Proceedings of the 2003 ACM symposium on Applied computing. 2003. ACM.

[8] Munro, R., S. Chawla, and P. Sun. Complex spatial relationships. in Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. 2003. IEEE.

[9] Xiong, H., et al. A Framework for Discovering Co-Location Patterns in Data Sets with Extended Spatial Objects. in SDM. 2004. SIAM.

[10] Wang, J., W. Hsu, and M.L. Lee. A framework for mining topological patterns in spatio-temporal databases. in Proceedings of the 14th ACM international conference on Information and knowledge management. 2005. ACM.

[11] Shekhar, S., et al., Trends in spatial data mining. Data mining: Next generation challenges and future directions, 2003: p. 357-380.

[12] Barua, S. and J. Sander. Mining statistically sound co-location patterns at multiple distances. in Proceedings of the 26th International Conference on Scientific and Statistical Database Management. 2014. ACM.

[13] Shekhar, S., et al., Identifying patterns in spatial information: A survey of methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**(3): p. 193-214.

[14] Yoo, J.S., S. Shekhar, and M. Celik. A join-less approach for co-location pattern mining: A summary of results. in Data Mining, Fifth IEEE International Conference on. 2005. IEEE.

[15] Shekhar, S. and Y. Huang. Co-location rules mining: A summary of results. in Proc. Spatio-temporal Symposium on Databases. 2001.