On Discovering Co-Location Patterns in Datasets: A Case Study of Pollutants and Child Cancers

Jundong Li · Aibek Adilmagambetov · Osmar R. Zaïane · Alvaro Osornio-Vargas · Osnat Wine

Received: date / Accepted: date

Abstract We intend to identify relationships between cancer cases and pollutant emissions and attempt to understand whether cancer in children is typically located together with some specific chemical combinations or is independent. Co-location pattern analysis seems to be the appropriate investigation to perform. Co-location mining is one of the tasks of spatial data mining which focuses on the detection of co-location patterns, the sets of spatial features frequently located in close proximity of each other. Most previous works are based on transaction-free apriori-like algorithms which are dependent on user-defined thresholds and are designed for boolean data points. Due to the absence of a clear notion of transactions, it is nontrivial to use association rule mining techniques to tackle the co-location mining problem. The approach we propose is based on a grid “transactionization” of the geographic space and is designed to mine datasets with extended spatial objects. Uncertainty of the feature presence in transactions is taken into account in our model. The statistical test is used instead of global thresholds to detect significant co-location patterns and rules. We evaluate our approach on synthetic and real datasets. This approach can be used by researchers looking for spatial associations between environmental and health factors. In addition, we explain the data modelling framework which is used on real datasets of pollutants (PRTR/NPRI) and childhood cancer cases.

Keywords Co-location mining · Uncertain data mining · Association rule and frequent pattern mining · Air Pollutant and environmental health

Jundong Li · Aibek Adilmagambetov · Osmar R. Zaïane
Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, T6G 2E8
E-mail: {jundong1, adilmaga, zaiane}@ualberta.ca

Alvaro Osornio-Vargas · Osnat Wine
Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada, T6G 1C9
E-mail: {osornio, osnat}@ualberta.ca
1 Introduction

Co-location mining aims to discover patterns of spatial features often located close to each other, i.e., in geographic proximity. An example of a pattern is a co-location of symbiotic species of plants and animals depending on ecological conditions. Figure 1 illustrates a sample spatial dataset with point features. As can be observed, instances of feature “+” are often located close to instances of “◦”. Similarly, objects of feature “⋆” are seen close to instances of “▽”. The main purpose of co-location mining is to come up with a set of hypotheses based on data features and statistics that can be useful for domain experts so they can reduce the range of possible patterns that are hidden in the data sets and need to be checked. A discovery of spatial co-location patterns may lead to useful knowledge in various applications. For instance, one might be interested in animal species that live close to certain types of landmarks such as rivers, meadows, forests, etc. In another example, co-location patterns which involve crime incidents and locations of various businesses can be useful for criminologists. Some of the application domains for co-location mining are biology, urban studies, health sciences, earth and atmospheric sciences, etc. Even though this task seems to be similar to association rule mining (ARM) which is used in knowledge discovery, the use and adaptation of ARM techniques is not trivial due to the fact that features are embedded into a geographic space and there is no clear notion of transactions. ARM consists of discovering rules that express association between items in a database of transactions. The associations are based on a notion of frequency bound by a given threshold like minimum frequency.

A motivating application of this paper is the detection of possible spatial associations of different chemicals and cases of childhood cancer. Cancer,
a multifactorial class of diseases, characterized by uncontrolled growth of abnormal cells, their invasion into other tissues, and metastasis, is one of the leading causes of adults death in both developed and developing world [8,11]. Although some people are genetically predisposed to cancer, most cases of cancer are suspected to be linked to environmental factors such as air pollutants, radiation, various infections, tobacco, and alcohol. However, causes of childhood cancer are difficult to determine partially because of the fact that childrens cancer cases are relatively rare and the levels of exposure to environmental factors are hard to evaluate. A collaborative research effort with the Faculty of Medicine at University of Alberta is trying to identify associations between cancer cases and known emissions by industry. Some chemicals are proven to be carcinogens while others are not known to cause cancer alone. It is unknown if certain combinations of chemicals can be associated with higher rates of cancer. Moreover, even if potentially problematic combinations are not emitted by the same industry, atmospheric conditions can contribute to the mixture. We deploy our model on the dataset containing information on chemical emission points, amounts of released chemicals and childhood cancer cases location when they were first diagnosed in the provinces of Alberta and Manitoba, Canada. Figure 2 displays part of the dataset with rectangles representing pollutant emission points, triangles for cancer cases, and polygons for urban municipalities. We need to build a modeling framework which handles the data as accurately as possible and takes into account various factors which affect distribution of chemicals. While we are not intending to find causalities, the goal of the study is to identify potential interesting spatial associations in order to state hypotheses and further investigate a relationship between childhood cancer and specific combinations of chemicals.

Most of the existing approaches to the co-location mining problem [31,23,36,35] deploy a framework which requires a user-defined minimum prevalence threshold. Without prior knowledge it could be difficult to choose a proper threshold. Furthermore, spatial features often have various frequencies in datasets, and one global threshold might lead to omission of some co-location patterns and rules with rare events or detection of meaningless patterns. Another limitation of most algorithms is that they work with point spatial features and one neighbourhood distance threshold, whereas in reality there are datasets which, in addition to point instances, also have lines and polygons, e.g., a road network map. Furthermore, the information in some datasets is uncertain: the presence of the feature in the region could depend on different factors. For example, a pollutant released from a facility distributes according to the climatic factors in its area, and the probability of detecting chemical in the region close to the emission point is higher than in remote regions.

In this paper, we propose a new framework which combines co-location mining, frequent pattern and association rule mining. A statistical test is used to determine the significance of co-location patterns and rules. A co-location is considered as significant if it has a surprisingly high level of prevalence in comparison with randomized datasets which are built under the null
hypothesis that the features are independent from each other. The uncertainty of information is modeled as a dependence on the distance from the spatial object. We explain filtering techniques that help to decrease computation.

The remainder of the paper is organized as follows. An overview of the related work is given in Section 2. The proposed framework and its outline are described in Section 3. Section 4 describes the challenges and modeling framework used to mine the co-location patterns between pollutants and childhood cancer cases. The experiments are presented in Section 5, followed by conclusions in Section 6.

This work is an extension of our DaWak’13 paper [3].

2 Related Work

This section starts with a brief introduction of spatial data mining and some of its applications and tasks and then gives a general overview of significant studies related to co-location mining, frequent mining and association rule mining with certain and uncertain data.

2.1 Spatial Data Mining

Spatial data mining is the process of extracting interesting and useful patterns in geographic datasets. It is a growing and promising field which has gained close attention of researchers during the last two decades. The technological advances in data storage and widespread use of GPS technologies, remote sensing devices, and location-based services have created large amounts of spatial data. The spatial data processing and analysis is useful in a wide range of applications such as business applications, population analysis, social
On Discovering Co-Location Patterns in Datasets: A Case Study

sciences, environmental sciences, and many others. For example, a businessman may be interested in spatial analysis in order to find the best location for a new store based on population data and current store locations.

In contrast to classical data mining, spatial data mining has some specific features. In classical data mining it is assumed that data objects are independent from each other like different transactions in association rule mining. However, in spatial datasets, objects situated close to each other tend to be more similar and have the same characteristics than objects located farther. Another example is a gradual change of temperature and precipitation levels. This observation is called spatial autocorrelation. Another difficulty in dealing with spatial data is relatively higher complexity of data object types and their relations. There are not only points but also lines and polygons in spatial databases. The relations between objects are implicit like intersection, containment, enclosure, etc.

Various types of methods and approaches are used in spatial data mining. Some of the tasks of spatial data processing and analysis include the following research areas like spatial clustering, spatial characterization, spatial trend detection, spatial classification, outlier detection, prediction of events as well as co-location mining [19,32].

2.2 Co-location Mining

Co-location mining is one of the tasks of spatial data mining that can be divided into two classes of methods: spatial statistics approaches and spatial data mining approaches.

2.2.1 Spatial Statistics Approaches

Spatial Statistics Approaches deploy statistical techniques such as cross K-functions with Monte-Carlo simulations [17], mean nearest-neighbor distance, and spatial regression models [14] to evaluate co-location patterns of two features and find co-locations among them. The disadvantages of these approaches are expensive computation time and difficulty of applying them to patterns of sizes more than two spatial features.

2.2.2 Spatial Data Mining Approaches

Spatial Data Mining Approaches could be categorized into transaction-based methods (works with transactions), and spatial join-based methods (use spatial joins of instance tables or feature layers).

Transaction-based approaches work by creating transactions over the space and using association rule mining algorithms on these transactions [6,26,28]. One of these methods, a reference centric model [29], creates transactions around a reference feature specified by a user. Each set of spatial features that form neighbourhood relationships with an instance of the reference feature is
considered as a transaction. However, not all applications have a clearly defined reference feature. For example, in urban studies features could be schools, fire stations, hospitals, etc., and there is no one specific feature of interest. 

Another approach, a window-centric model \[31\], divides the space into cells and considers instances in each cell as a transaction. The model can consider all possible windows as transactions or use spatially disjoint cells. However, a major drawback of the model is that some instance sets are divided by the boundaries of cells, so some of the spatial relationship information is lost. In addition, maximal cliques (maximal sets of instances which are pair-wisely neighbors) in spatial data are proposed to be used as transactions \[7, 27\], but this approach does not preserve the information on how close or remote are objects in cliques as long as they are considered being neighbours.

Spatial join-based approaches work with spatial data directly. They include cluster-and-overlay methods and instance-join methods. In the cluster-and-overlay approach, clustering is used to mine associations. For example, concentrations of objects in layers are found in order to search for possible causal features \[21\]. In another work \[20\], a map layer is constructed for each spatial feature based on clusters of instances or boundaries of clusters. The authors of \[20\] propose two algorithms for cluster association rule mining, vertical-view and horizontal-view approaches. In the former, clusters for layers (features) are formed and layers are segmented into a finite number of cells. Then, a relational table is constructed where an element is equal to one, if the corresponding cell satisfies an event in a layer, and the element is zero otherwise. An association rule mining algorithm is applied to the table. The second approach evaluates intersections of clustered layers. A clustered spatial association rule is of the form \(X \rightarrow Y (CS\%, CC\%)\), where \(X\) and \(Y\) are the sets of layers, \(CS\%\) is the clustered support (the ratio of the area that satisfies both \(X\) and \(Y\) to the total area of the study region), and \(CC\%\) is the clustered confidence (the percentage of cluster areas of \(X\) that intersect with clusters of \(Y\)). However, these approaches might be highly sensitive to a choice of clustering methods. In addition, they assume that features are clustered, even though spatial features may not form explicit clusters.

Another type of spatial join-based method: the instance-join algorithm, is similar to classical association rule mining. One of the first proposed co-location pattern mining frameworks of this type \[31, 24\] is based on neighbourhood relations and the participation index concept.

Clustering is used in a similar approach \[25\]. For two spatial features \(f_1\) and \(f_2\), if the density of instances of \(f_1\) in proximity of objects of feature \(f_2\) is higher than an overall density of \(f_1\), then feature \(f_1\) is considered to be co-located with feature \(f_2\), i.e. their objects tend to be situated close to each other. This algorithm suffers from the same limitation as the previous approach. It is based on an assumption that spatial instances of a feature are situated close to each other and form clusters which may not be the case in some real-world applications.

The basic concepts of the co-location mining framework are analogous to concepts of association rule mining. As an input, the framework takes a set
of spatial features and a set of instances, where each instance is a vector that contains information on the instance ID, the feature type of the instance, and the location of the instance. As an output, the method returns a set of co-location rules, where a co-location rule is of the form $C_1 \rightarrow C_2 (PI, cp)$, where $C_1$ and $C_2$ are co-location patterns, $PI$ is the prevalence measure (the participation index), and $cp$ is the conditional probability. The participation index $PI(C)$ of a co-location pattern $C$ is defined as:

$$PI(C) = \min_{f_i \in C} \{pr(C, f_i)\}$$ (1)

where $pr\{C, f_i\}$ is the participation ratio of a feature $f_i$ in a co-location $C$ and it is computed as:

$$pr(C, f_i) = \frac{\text{number of distinct instances of } f_i \text{ in instances of } C}{\text{total number of instances of } f_i}$$ (2)

A co-location pattern is considered prevalent, or interesting, if for each feature of the pattern at least $PI\%$ instances of that feature form a clique with the instances of all other features of the pattern according to the neighbourhood relationship. Similar to association rule mining, only frequent $(k-1)$ patterns are used for the $k$ candidate generation process. A co-location rule $C_1 \rightarrow C_2$ is considered prevalent, if its conditional probability is higher than a threshold. The conditional probability $cp(C_1 \rightarrow C_2)$ is defined as:

$$cp(C_1 \rightarrow C_2) = \frac{\#\text{distinct instances of } C_1 \text{ in instances of } C_1 \rightarrow C_2}{\text{total \#instances of } C_1}$$ (3)

In the approach mentioned above, it is assumed that spatial features occur with similar levels of frequency. Therefore, if a dataset contains rare spatial features, co-locations involving these rare events will be pruned by a prevalence threshold because more frequent features dominate rare ones and no pattern with a rare event can become prevalent. For example, a rare disease will not be captured in co-location patterns due to the fact that its causes are more frequent in the database. To solve this limitation, Huang et al. [23] continue their previous work by introducing an algorithm that finds co-location patterns with rare features. Instead of the participation index threshold, the authors propose to use the maximal participation ratio threshold. Briefly, a co-location pattern is considered prevalent if $maxPR\%$ instances of at least one of the features in the pattern are co-located with instances of all other features, where $maxPR$ is the maximal participation ratio:

$$maxPR(C) = \max_{f_i \in C} \{pr(C, f_i)\}$$ (4)

It is not well explained how the algorithm deals with noise features. For example, if some features have only a limited number of instances, it is highly probable that every co-location with these features will be considered prevalent.
Both mentioned methods use computationally expensive instance joins to identify instances of co-location patterns, and their running time grows fast as the number of instances and sizes of candidate patterns increase. Yoo et al. [37] propose a partial-join approach for mining co-location patterns. A study space is partitioned into square cells with the side length equal to a neighbourhood distance threshold. A set of spatial instances in a cell form a clique. Join operations are required to identify neighbourhood relationships divided by boundaries of cells. Even though this approach reduces the computation time, it still requires a large amount of spatial joins.

The joinless algorithm [36] is a follow-up work to the partial-join approach. It further decreases computation time of constructing neighbourhood relationships. The main idea is to find star neighbourhoods instead of calculating pairwise distances between all instances in a dataset. The neighbourhood relationship is materialized in the form of a table where for each instance, all its neighbors are listed. Then, in order to ensure that pattern instances form cliques, an instance-lookup scheme is used to filter co-location instances. In addition, three filtering steps are used to find a set of prevalent co-location patterns. The authors prove that their algorithm finds a complete and correct set of co-location patterns and rules. The experiments on synthetic and real datasets show that the joinless approach has better performance in terms of the running time than the join-based algorithm.

Based on their work, Xiao et al. [34] improve the running time by dividing spatial objects into partitions and detecting neighboring instances in dense regions first. The algorithm finds instances in dense regions and maintains an upper bound on a prevalence measure for a candidate pattern. If the upper bound becomes less than a threshold, the method decides that it is a false candidate and stops identifying its instances in less dense regions.

Several other work extended the basic co-location mining framework. For example, Xiong et al. [35] propose a framework for detecting patterns in datasets with extended objects. Extended objects are objects that are not limited to spatial points but also include lines and polygons. Buffers are created around spatial instances; their sizes might depend on types of features. In the proposed model, candidate patterns are pruned by a coverage ratio threshold. In other words, if an area covered by features of a candidate pattern is greater than a predefined threshold, this pattern is considered prevalent. In order to lessen a usage of geographic information systems (GIS) overlay methods, a coarse-level mining step is used. At this level, minimum buffer bounding boxes of spatial objects are considered by the algorithm instead of true buffer shapes. Then, patterns that have coarse level coverage ratio higher than the threshold are evaluated using actual buffers. Compared to previous models, this approach takes into account shapes of spatial objects and their distribution in space rather than using one neighbourhood distance for varying types of features. Expensive GIS overlays are used in this method and a filtering technique is proposed in order to improve its performance.

The approaches mentioned above use thresholds on interestingness measures, which causes meaningless patterns to be considered as significant with
Table 1 Comparison of association rule mining and co-location mining [23].

| Association rule mining | Co-location mining |
|-------------------------|--------------------|
| Item                    | Spatial feature    |
| Itemset                 | Spatial feature set|
| Frequent pattern        | Co-location pattern|
| Support & Confidence    | Interestingness measures|
| Transactional database  | Spatial database   |

a low threshold, and a high threshold may prune interesting rare patterns. Instead of a threshold based approach, Barua and Sander [9] use the statistical test to mine frequent co-location patterns. The participation index of a pattern in observed data is calculated as in previous studies. Then, for each co-location pattern the authors compute a probability $p$ of seeing the same or greater value of the prevalence measure under a null hypothesis model. A co-location is considered significant if $p \leq \alpha$, where $\alpha$ is a level of significance.

2.3 Frequent Pattern Mining

The co-location mining problem is similar to the canonical data mining problem: association rule mining. The most classical example of association rule mining is discovering sets of goods that are often bought together. The concepts of association rule mining and co-location mining are compared in Table 1.

Considering the similarity between frequent pattern mining and co-location pattern mining, we discuss the classical frequent pattern mining problem in this subsection. The concepts of frequent pattern and association rule mining were first introduced by Agrawal et al. [5]. Various approaches to these problems have been proposed over the past two decades. Apriori [6] is the first and one of the most-known algorithms used for frequent itemset mining. This approach is designed to work on transactional data and consists of a bottom-up candidate generation process where $k$-size candidate itemsets are generated from frequent $(k-1)$-itemsets and tested against the database to obtain frequent $k$-itemsets. This process is repeated until no more candidate patterns can be generated. The correctness of the algorithm is based on the downward closure, or apriori, property, which states that if an itemset is frequent, then all its subsets are also frequent. In other words, an itemset cannot be frequent, if one of its subsets is infrequent.

The ARM problem is defined as follows. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of $m$ items and $T = \{t_1, t_2, ..., t_n\}$ be a set of $n$ transactions where a transaction $t$ is a subset of items in $I$. For an itemset $X \subseteq I$, the support of $X$ is defined as the ratio of transactions in $T$ that contain instances of $X$. An itemset is considered frequent, if its support is higher than a user-specified minimum support threshold. An association rule is a rule of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. The confidence of a rule $X \rightarrow Y$ is the support of $X \cup Y$ divided by the support of $X$. 
2.4 Frequent Pattern Mining with Uncertain Data

The algorithms and approaches mentioned above are constructed to work with data where presence of items in transactions is certain. For example, market basket datasets are certain and precise. However, in some applications, data may be incomplete or may have errors. For instance, sensor reading records might include some erroneous data due to various internal and external factors such as sensor failures or extreme weather conditions. Uncertainty can be expressed in terms of existential probabilities; each item of a transaction is followed by a probability of its existence in this transaction. An example transactional dataset is shown in Table 2.

Most studies use a notion of expected support \([16,15]\) to mine frequent patterns from uncertain databases. The expected support \(E(s(I))\) of an itemset \(I\) is defined as a sum of expected probabilities of the presence of \(I\) in each of the transactions in a database. A probability \(p(I, T)\) of presence of \(I\) in a transaction \(T\) is a product of corresponding probabilities of items in the transaction. An itemset is considered significant if its expected support exceeds a \(\minsup\) threshold.

Several approaches to frequent pattern mining problem with uncertain data have been studied by Aggarwal et al. \([4]\). These approaches are extended from existing classical frequent itemset mining methods and can be divided into two categories: candidate generate-and-test algorithms (extension of Apriori algorithm) and pattern growth algorithms (extensions of FP-growth and H-Mine \([29]\)). According to this study, while FP-growth is efficient and scalable in the deterministic case, its extension to the uncertain case behaves differently due to challenges associated with uncertain data. UH-Mine, an extension of H-Mine, is reported to provide best trade-offs in terms of running time and memory usage.

Bernecker et al. \([10]\) proposed PFIM (Probabilistic Frequent Itemset Mining) model which is based on the possible world paradigm. Instead of the expected support, PFIM uses the frequentness probability as a significance measure. By using a dynamic computation method, the algorithm is reported
to run in $O(|T|\text{minsup})$, where $|T|$ is the number of transactions and \text{minsup} is a user-defined threshold. Without it, the approach runs in exponential time. However, the algorithm requires the \text{minsup} threshold to be defined, and it is nontrivial to apply the statistical test to the frequentness probability.

3 Proposed Algorithm

Various approaches to the co-location mining problem have been proposed during the past decade. However, most of them focused on improving the performance of existing frameworks which have some limitations. Several studies addressed these issues but only separately, and they still cannot assist these algorithms from being used for some real-world applications such as our motivating question of exploring whether co-locations of cancer cases and sets of released chemicals exist.

First, the usage of prevalence measure threshold for the detection of interesting co-location patterns and rules is a main limitation factor of many co-location mining algorithms. In spatial datasets, features usually have a varying number of instances; they could be extremely rare or be present in abundance. Therefore, one threshold for the participation index (or any other significance measure) cannot capture all meaningful patterns, while other patterns could be reported as significant even if their relation is caused by autocorrelation or other factors. In addition, most current algorithms use a candidate generation process which forms $(k + 1)$-size candidates only from significant $k$-size patterns. However, a set of features could be interesting even if some of its subsets are not significant. For example, two chemicals may not be correlated with disease separately, but could do it when in combination. In this paper, we use the statistical test to replace the prevalence measure threshold. It is first proposed for co-location pattern mining by Barua and Sander [9]. The pattern is significant, if the probability of seeing the same or greater value of the prevalence measure in $N$ artificial datasets is less than $\alpha$ (the significance level) under the null hypothesis that there is no spatial dependency among features of the pattern. Each candidate pattern is evaluated separately rather than applying one threshold for all of them.

Second, most co-location mining approaches use one distance threshold to identify neighbourhood relationships among spatial objects. However, in some applications it might oversimplify the real situation. For instance, in zoological research, various species have different habitat ranges: birds (especially, birds of prey) might interact with other species on greater distances, while subterranean animals are limited in their movements. Therefore, a usage of one distance threshold might lead to wrong results. Furthermore, most current co-location mining frameworks are designed to work with point data; however, other types of objects may exist in spatial datasets such as lines (roads, communications) and polygons (polluted regions, areas which had no precipitation for some period or were exposed to other climatic factors). Even though the framework for extended objects [35] deals with lines and
polygons, it also uses a threshold for a prevalence measure. Furthermore, this framework cannot deal with uncertainty in datasets as is explained in the following paragraph.

Third, in some applications, information in datasets is uncertain; data may be incomplete or may have errors. For example, distribution of a chemical released from a chimney in a polluted region is not uniform. Areas closer to an emission point are generally exposed to higher pollutions than places far away from the release point. Another example is climatic data collected by sensors which might have errors in their readings. Uncertainty can be expressed in terms of existential probabilities; each item of the transaction is followed by the probability of its existence in this transaction. Uncertainty in datasets has been researched for the frequent itemset mining problem. However, to the best of our knowledge there is no such work done for spatial data.

In this paper, we propose a new framework that addresses aforementioned limitations. It uses grid-based “transactionization” (creating transactions from a dataset). The statistical test is performed on the derived set of transactions to get significant co-location rules or patterns.

3.1 Algorithm Design

The objective of this work is to detect significant patterns in a given spatial dataset that have a prevalence measure value higher than the expected one. The spatial dataset may contain points, lines or polygons. A buffer is first built around each spatial object, and it defines the area affected by that object; for example, the buffer one around an emission point shows the area polluted by a released chemical. The buffer size might be one for all objects or it might be different for each of the spatial instances depending on various factors, which may vary for different applications. In addition, the likelihood of the presence of the corresponding feature in the region covered by the object and its buffer is not uniform and may depend on factors such as distance from the object.

We propose a new transaction-based approach that is suitable for extended spatial objects and uncertain data. Recall that previous transaction-based methods have some limitations. A window-centric model cuts off neighbourhood relations of instances located close to each other but in different partitions. A reference-centric model may get duplicate counts of spatial instances. In addition, it is nontrivial to generalize this approach to applications with no reference feature. Instead of the previous models we propose a new grid-based transactionization method. In order to transactionize spatial data we use a grid which points are imposed over a given map. Figure 3(a) illustrates an example dataset with buffers around spatial point instances, and a grid is laid over it in Figure 3(b). Similarly, buffers can also be created around linear and polygonal spatial objects. In two-dimensional space grid points represent a square regular grid. Due to the spheroid shape of the Earth, a grid used for real-world applications becomes irregular. However,
with a careful choice of a grid granularity this fact should not considerably affect the accuracy of results.

Each point of the grid can be seen as a representation of a respective part of the space. A grid point may intersect with one or several spatial objects and their buffers. A transaction is defined as a set of features corresponding to these objects. Let us assume that a sensor capable of detecting various features is placed at each grid point. A set of features detected by each sensor can be seen as a transaction. However, sensor readings are not fully reliable; they are uncertain and can be affected by extreme environmental conditions, sensors’ hardware, durability, and other factors. For example, it is possible that a sensor detects a pollutant only if a certain amount of it is present in the sensor’s environment. In addition, a likelihood of a presence of a feature in a region covered by an object and its buffer is not uniform. Alternatively, since we work with estimates and not with sensors and sensor collected data, we can use in our model the notion of concentration of features. While the fading concentration is not a probability, it can be used to show the feasibility of our model using uncertainty. Intuitively, a feature is more likely to be detected in buffer parts which are closer to a feature point than in parts that are farther away from it. Furthermore, spatial datasets can be noisy and contain errors; locations of instances and their presence can be uncertain. In order to take into account these uncertainties, a probability of a feature being present in a transaction is also stored. One of the ways to model uncertainty when transforming a spatial dataset into a set of transactions is to use a distance from a spatial object to a grid point (our method of estimating a feature presence probability is explained in the following section).

For example, a grid point $gp_2$ in Figure 3(c) is located closer to a point $A_1$ than a point $gp_1$; we can assume that $p(A, gp_2) > p(A, gp_1)$. An example transaction set is already shown in Table 2. When a grid point intersects with several instances of the same feature or their buffers, the highest existential probability is taken as a probability of detecting this feature at the given grid point. Algorithm 1 displays the pseudocode of the transactionization step. Moreover, the granularity of the grid, or a distance between points of the
Algorithm 1 GetTransactions(S): Transactionization step.

1: $T = \emptyset$: set of transactions
2: $G$: set of grid points
3: Build buffer zones around spatial objects of $S$
4: Impose a grid $G$ over the dataset $S$
5: for all point $g \in G$ do
6:   $t = \text{get a set of features which instances contain } g \text{ with corresponding existential probabilities}$
7:   $T = T \cup t$
8: end for
9: return $T$

grid, should be carefully chosen for each project or application and it may depend on an average size of a region covered by a spatial object and its buffer. A great distance between grid points may negatively affect accuracy of results because small feature regions and their overlaps might get a different number of intersecting grid points depending on a grid imposition. On the other hand, when a distance between grid points is too short, the number of derived transactions increases, and the following computation of pattern significance levels might become prohibitively expensive, especially when the number of candidates is large.

Given a set of transactions $T$, derived after the transactionization of a spatial dataset, and a set of spatial features $F$, a prevalence measure value is calculated for all candidate co-location patterns or rules. In some applications experts look for sets of features that are often co-located with each other, but not necessarily in a cause-effect realationship. In this case, which is analogous to the frequent pattern mining, the expected support $\text{ExpSup}(P)$ might be used to define a level of interestingness of a pattern $P$. In frequent pattern mining with certain data the support of a pattern is counted deterministically as the number of transactions containing all features of the pattern. However, in the case of uncertain data, transactions are probabilistic and, therefore, the support is counted in expected value. The expected support of a pattern $P$ is defined as:

**Definition 1** The expected support $\text{ExpSup}(P)$ of a pattern $P$ is defined as the sum of probabilities of presence of $P$ in each of the transactions $t$ in the uncertain database:

$$\text{ExpSup}(P) = \sum_{t \in T} p(P, t). \quad (6)$$

The probability of a presence of a pattern $P$ in a transaction $t$ can be computed as follows:

**Definition 2** The probability $p(P, t)$ of the pattern $P$ occurring in a transaction $t$ is the product of corresponding feature instance probabilities:

$$p(P, t) = \prod_{f \in P} p(f, t). \quad (7)$$
For some other applications, researchers intend to discover a predefined set of rules. For example, for a dataset of disease outbreaks and possible cause factors, a typical co-location rule is of the form $C \rightarrow D$, where $C$ is a set of cause features and $D$ is a disease feature. For this scenario, the expected confidence $ExpConf(X \rightarrow Y)$ can be used as a prevalence measure of a co-location rule $(X \rightarrow Y)$, where $X \subseteq F$, $Y \subseteq F$, and $X \cap Y = \emptyset$. And the definition of $ExpConf(X \rightarrow Y)$ is defined as follows:

**Definition 3** The expected confidence $ExpConf(X \rightarrow Y)$ of a rule $X \rightarrow Y$ is defined as:

$$
ExpConf(X \rightarrow Y) = \frac{ExpSup(X \cup Y)}{ExpSup(X)}.
$$

In the previous two steps, we build buffers around each instances and then transactionization each instances over the map, and a prevalence measure value is calculated for all candidate co-location patterns or rules. Now our goal is to discovery a set of significant patterns and rules. As discussed above, the usage of one threshold on a prevalence measure may result in the discovery of wrong patterns and the omission of interesting ones. Moreover, only co-location patterns or rules with surprising levels of prevalence measure should be considered as significant. In other words, it is unlikely that instances of features in a significant pattern are located close to each other only by chance according to a predefined significance level threshold.

Thus, in order to avoid the chance that instance of features are co-located together, we need to do some randomization tests, the instances of features are considered to be independent in each randomized test. So in this step, we utilize the statistical test, to help us to estimate the likelihood of seeing the same level of the prevalence measure or greater under a null hypothesis that features of a pattern or rule are spatially independent from each other, i.e. the randomization test. The definition of significance is as follows:

**Definition 4** A co-location pattern $P$ is considered significant at level $\alpha$, if the probability $p$ of detecting the observed expected or larger support $ExpSup_{obs}(P)$ in a dataset complying with a null hypothesis is not greater than $\alpha$.

The same logic can be applied to a case when significant co-location rules are mined:

**Definition 5** A co-location rule $R$ is considered significant at level $\alpha$, if the probability $p$ of detecting the observed or higher expected confidence $ExpConf_{obs}(R)$ in a dataset complying with a null hypothesis is not greater than $\alpha$.

Let us suppose that the expected confidence $ExpConf$ is used as a prevalence measure. Let $ExpConf_{obs}(X \rightarrow Y)$ denotes the expected confidence of a co-location rule $X \rightarrow Y$ in a real dataset, and $ExpConf_{rand}(X \rightarrow Y)$ denotes the expected confidence of rule $X \rightarrow Y$ in a randomized dataset which is generated under the null hypothesis. The expected confidence of the
co-location rule in each of \( R \) randomized datasets is calculated in order to estimate the probability \( p \). Having the number of simulations \( R \), the value of \( p \) is computed as:

\[
p = \frac{R_{\geq \text{ExpConf}_{\text{obs}}} + 1}{R + 1},
\]

where \( R_{\geq \text{ExpConf}_{\text{obs}}} \) is the number of simulations in which \( \text{ExpConf}_{\text{rand}}(X \rightarrow Y) \geq \text{ExpConf}_{\text{obs}}(X \rightarrow Y) \). The observed dataset is added to both numerator and denominator.

If the \( p \)-value is less or equal to a predefined level of significance \( \alpha \), the null hypothesis is rejected. Therefore, it is unlikely that the features of the rule are spatially independent; they are not situated close to each other only by chance. The co-location rule \( X \rightarrow Y \) is considered significant at level \( \alpha \).

The above explanation can also illustrate a process of detecting co-location patterns. The difference is that instead of the expected confidence, the expected support \( \text{ExpSup} \) is used as a prevalence measure.

In order to estimate probability \( p \), a set of randomized datasets is generated under the null hypothesis. Each randomized dataset has the same number of instances of each feature as in the original dataset. In addition, distribution of instances of each feature in a randomized dataset should be similar to its distribution in the original data. For instance, disease cases should be placed within populated areas. Obviously, random placement of disease cases all over the study region can lead to invalid results, especially in the case when most of the region is unpopulated. Another example can be found in biology. Some animal species may have various requirements to their habitats such as a location close to water reservoirs or presence of certain types of vegetation. This observation needs to be taken into account in a randomized dataset generation process. In the next section, we will discuss how the randomized dataset affect the discovery of co-location rules.

Algorithm 2 shows the pseudocode of our approach in a case when co-location patterns are mined in a spatial dataset and the expected support is used as a prevalence measure.

3.2 Candidate Filtering Techniques

The calculation of the \( p \)-value is repeated for all candidate co-location patterns or rules. The number of candidates grows exponentially with the number of spatial features in the dataset. In addition, accuracy of the \( p \)-value depends on the number of simulation runs. Therefore, the more randomized datasets are checked, the more accurate are the results. These two factors may lead to an enormous amount of computation. However, the support of a co-location decreases as the size of a candidate pattern or rule increases, because fewer transactions contain all its features. Therefore, researchers might put a threshold on the support or the maximal size of a candidate in order to analyze only patterns and rules that are backed by a meaningful number of transactions. In addition, we use the following filtering techniques to
Algorithm 2 Mining significant co-location patterns.

Require: Spatial dataset $D$.
Level of significance $\alpha$.
Number of simulation runs $R$.
Set of randomized spatial datasets $RD[1..R]$.

Ensure: Set of significant co-location patterns $P$

1: $T$: set of transactions
2: $CP$: set of candidate patterns
3: $T = GetTransactions(D)$
4: for all $cp \in CP$ do
5:   $cp.ExpSup_{obs} = ComputeExpSup(cp, T)$
6:   if $cp.ExpSup_{obs} = 0$ then
7:     $CP = CP \setminus cp$
8:   end if
9: end for
10: for $i = 1 \rightarrow R$ do
11:   $T = GetTransactions(RD_i)$
12:   for all $cp \in CP$ do
13:     $cp.ExpSup_{sim}[i] = ComputeExpSup(cp, T)$
14:     if $cp.ExpSup_{sim}[i] \geq cp.ExpSup_{obs}$ then
15:       $cp.R = cp.R + 1$
16:     end if
17: end for
18: end for
19: $P = CP$
20: return $P$

exclude from the analysis candidate patterns and rules that are definitely not significant.

1. After the calculation of a prevalence measure for candidate patterns in a real spatial dataset, some of the patterns may have a prevalence measure value equal to zero. It means that combinations of features of these patterns do not exist in the dataset. Obviously, these patterns cannot be statistically significant and they can be excluded from the set of candidate patterns (lines 6-8 in Algorithm 2). In some applications a low-value threshold on a prevalence measure can be used in order to get significant patterns and rules with a certain level of interestingness. In this case, candidate patterns and rules with prevalence value lower than this threshold can also be pruned and excluded from further computations.

2. During the calculation of the $p$-value for candidate patterns for which an observed prevalence is higher than zero, some of the candidate patterns might show a $p$-value that has already exceeded the level $\alpha$. For example, let us assume that the number of simulation runs is 99 and $\alpha = 0.05$ (which is most commonly used $p$-value to claim significance). If after ten simulation runs, the prevalence measure of a pattern $P$ is greater than the observed prevalence in 5 randomized datasets, pattern $P$ already surpassed
the threshold \((5 + 1)/(99 + 1) > 0.05\). Therefore, it definitely cannot be significant and can be excluded from the following 89 checks (lines 16-19 in Algorithm 2). Thus, the computation time is greatly reduced. With this filter, after the last simulation run the set of candidates contains only significant patterns or rules.

3.3 Advantages of the Proposed Algorithm

By combining techniques of co-location mining and frequent pattern mining, we address the limitations of previous models. Our framework has the following advantages:

- Our framework does not need thresholds on prevalence measures. The statistical test replaces a usage of one global threshold for a prevalence measure of candidate co-location patterns or rules. Only meaningful patterns are reported as significant. These patterns have the prevalence measure higher than an expected value under a null hypothesis that features of a pattern are independent from each other. Sometimes researchers do not need patterns or rules with very low support values even if they are significant. In this case a threshold on support can be used. However, it should have a relatively low value in comparison with other approaches, so it does not exclude meaningful patterns or rules.

- While a neighbourhood distance threshold used in many co-location algorithms is set to one value for all spatial features, our model can deal with varying buffer sizes. A buffer size may depend on types of features or on attributes of individual spatial objects. So, the algorithm can be used for applications where features differ from each other with regard to the effect to the environment around them, e.g., plant and animal species. Moreover, the model can be further extended to take into account not only point instances but also other types of spatial objects such as lines and polygons which are present in many spatial datasets.

- In most previous algorithms, two or more objects are considered to have a neighbourhood relationship, if they are located at a distance not farther than a distance threshold. However, these approaches do not take into account spatial information and context: how close or far the objects are situated from each other. Figure 4 illustrates an example of two pairs of neighboring spatial objects with corresponding buffer zones. Both pairs, \(A_1 - B_1\) and \(A_2 - B_2\), are neighbors and treated similarly by most co-location approaches, even though \(A_1\) and \(B_1\) are closer to each other than \(A_2\) to \(B_2\). Being located at the closer distance, the instances of the former pair are more likely to be related than the instances of the latter pair. By using buffer zones around spatial instances and transactions that are created from grid points, our algorithm ensures that the spatial location of objects is not ignored. The pair \(A_1 - B_1\) gets more transactions (shown as black dots in Figure 4) than the second pair of objects. Therefore, the real situation is represented more accurately. Consider another example. Let
Fig. 4 Neighboring objects $A_1 - B_1$ and $A_2 - B_2$. In the transactionization step, the intersection of $A_1$ and $B_1$ receives more transactions (black dots) than the pair $A_2$ and $B_2$.

Fig. 5 Intersection of neighboring objects.

spatial points $A_1$, $B_1$ and $C_1$ be pairwisely neighbors (Figure 5(a)). They are considered to form a clique by previous algorithms. However, as it can be seen in Figure 5(b) with certain buffer sizes it is possible that an actual intersection area of three buffers is relatively small. Furthermore, a scenario exists when there is no intersection of the three objects at all, although they form pairwise neighbourhood relationships, as it is illustrated in Figure 5(c). Our buffer-based framework is able to distinguish these cases. A varying number of transactions is derived from intersecting regions of multiple objects depending on distances between them and their buffer sizes.

– Similarly to classical frequent pattern mining applications where data can be certain (deterministic) or uncertain (probabilistic), spatial datasets can also exhibit uncertainty of feature existence in space. In other words, a probability of detecting an existence of a feature in a region closer to an observation point is higher than in regions situated farther from it. By
taking into account uncertainty and including it in our framework, we believe that our model increases accuracy of results.

4 Modeling Framework

A modeling framework that is used to handle and analyze data is an important part of any practical research. In theoretical studies it could be simplified in order to generalize a task and define algorithms that could be applied for a wide range of applications and domains. However, the usage of general approaches and algorithms may result in misleading or even wrong results. For example, a neighbourhood distance threshold is an important measure of an interaction and relationship between features. Obviously, one distance threshold cannot capture accurately all links among features. In biology, various animal species have different home ranges, areas where they search for food. Rodents may require little space, while birds forage on wider regions. Another example is derived from urban studies. Two points of interest, for example, a shopping mall and a grocery store, could be situated on a distance exceeding a threshold, but if they are connected by a high quality road, they are more likely to be co-located than other two points positioned seemingly close to each other but separated by some obstacles. Most domains of research, if not all, have their own nuances that must be taken into account by researchers in order to get most accurate and significant results.

The motivating task of this paper, detecting co-locations of pollutant emission points and childhood cancer cases, has unique difficulties and challenges. A distribution of a pollutant in a region is not uniform and it could depend on several factors: types of pollutants, amounts of release, weather conditions (wind, precipitation), topography, etc. Various chemicals have different levels of harmfulness and toxicity. In addition, a pollutant concentration might be inversely proportional to the distance from the emitting point. These are only a few examples. We show how we tackled some of these problems such as pollutant amounts, wind speed and direction, and uncertainty of presence of chemicals. Certainly, we do not aim to reproduce complicated air pollution distribution models which require considering many other variables and parameters. Instead, our model gives a simple framework that attempts to simulate real world conditions while operating with available data.

4.1 Pollutant Amounts

The dataset on pollutants contains the data on estimated yearly releases of chemicals by industry, according to Canada's National Pollutant Release Inventory [12]. For our research we use the average amount of released chemicals by a facility on a given year. The range of the average amount values varies from several kilograms to tens of thousands tons; the minimum
and maximum average yearly release in the dataset is 0.001 grams and 80,000 tons, respectively. Fig. 6 (a) displays an example dataset containing cancer points (feature A) and chemical points (features B and C). On Fig. 6 (b) buffer zones around pollutant points are based on the amount of a release at that location. For example, instance $C_1$ has a larger zone affected by this source point than instance $C_3$ which has smaller amount of emission. Buffer zones of cancer points aren’t changed.

For simplicity, we decided to take the maximal distance as the natural logarithm function of the release amount. This function gives a smooth curve which does not grow as fast as linear or root functions that give large numbers for heavier releases. Even though this technique oversimplifies the real world conditions of pollutant dispersion, it helps to make the results more precise. Other functions can be used to calculate the maximal distribution distance and they can depend on a type of a pollutant (a heavier chemical settles faster and on a shorter distance from a chimney) or a height of a chimney. An additional point that could be considered in future work is that the area very close to a chimney does not get polluted, and the higher is the chimney, the bigger is that region.

4.2 Wind Speed and Direction

The weather conditions and topographical features may affect the distribution of chemicals in the air. The examples of these factors are prevailing winds, precipitation, relative humidity, mountains, hills, etc. At the first step in this part of the modeling framework we include the wind speed and the prevailing wind direction on source points as variables of the model.

Regarding the wind speed and direction, two situations are possible. First, a region where a facility is located is windless throughout the year. In this case, a pollutant is assumed to disperse in a circular region around the source.
point with a radius of a circle derived from a released amount as discussed in the previous subsection. However, a second situation is more frequent; there is nonzero wind speed with the prevailing wind direction. In this case we presume that the original distribution circle is morphed into a more ellipse-like region. Figure 6(c) illustrates elliptical buffer regions; their forms are dependent on the wind speed and its frequent direction. Our calculations of the characteristics of an ellipse are based on the works by Getis and Jackson [22], and Reggente and Lilienthal [30]. The major axis of the ellipse is in the direction of the prevailing wind. We assume that the area polluted by a chemical when wind is present is the same as when there is no wind. Therefore, the coverage area of the ellipse is kept equal to the area of the original circle. The source point can be placed on the major axis of the ellipse between the center and upwind focus; in our model we locate it in the middle of the segment between these two points. Figure 7 shows an example of buffer transformation. The original circle buffer zone around the emission point $P$ is changed to an ellipse. As Figure 7 shows that the center of the buffer (i.e. the source of the pollutant) becomes a foci of the ellipse.

Obviously, wind with a higher speed distributes chemicals to greater distances. Therefore, we need to include the wind speed value in the computations. The lengths of the major semi-axis $a$ and minor semi-axis $b$ are dependent on the wind speed and derived from the equations:

$$a = r + \gamma |v|,$$

(10)

$$b = \frac{r^2}{a},$$

(11)

where $r$ is the radius of the original circle, $v$ is the wind speed, and $\gamma$ is the stretching coefficient.

The larger is a value of the stretching coefficient, the longer is the ellipse’s major axis. In this work it is fixed at 0.3, but it could be changed and have a different value for each of pollutants. The calculation of the length of the semi-minor axis $b$ follows our assumption that the area of the ellipse is equal to the area of the original circle.

We improve our model by using elliptical buffer zones, which depend on the average wind speed and most frequent wind direction, instead of circular buffers. However, this is only a simplified model. Other factors which affect
chemical distribution in air might be taken into account in future research to more accurately simulate real processes.

4.3 Wind Stations and Data Interpolation

In order to get values of the wind speed and prevailing wind direction, an interpolation of wind fields between weather stations is used. The data of monitoring stations in Alberta comes from two sources. First, the data from 18 stations is obtained from Environment Canada [18] which provides climate normals that are based on climate stations with at least 15 years of data between 1971 and 2000. The most frequent wind direction is a direction (out of possible eight directions) with the highest average occurrence count. Second, the data from 156 stations is derived from AgroClimatic Information Service (ACIS) [2]. The locations of stations in Alberta are displayed in Figure 8. The data provided by ACIS is daily from 2005 to 2011. In order to make the data consistent, the average wind speed and the most frequent wind direction are calculated using a method similar to the one used by Environment Canada [18]. The average wind speed is simply the average value of this parameter for all available days. The wind direction is rounded to eight points of the compass. A direction with the highest count of daily observations is assigned as the most prevailing wind direction.

Unlike Alberta, the data of monitoring stations in Manitoba comes from only one source, 20 stations from Environment Canada [18], it also contains climate normals information from the years of 1971 to 2000. The locations of stations in Manitoba are displayed in Figure 9.

The climate normals from two sources in Alberta and one source in Manitoba are used to make interpolations in the ArcGIS tool [1]. However, ArcGIS is restricted to linear surface interpolations and the wind direction is a nonlinear attribute. In linear systems (e.g., the number of sunny days or days with precipitation) there is only one unique path when moving from one number to another, for example, if we want to move the temperature from 37°C to 40°C, the only path is to linearly increase the first degree. On the other hand, nonlinear systems may have several paths. For example, there are clockwise and counter-clockwise directions to move from 90° to 270°: through 0° or 180°. These directions go from one point to the second but both are unique. Therefore, linear interpolations lead to wrong results when deployed directly to non-linear systems.

Interpolation of wind fields requires a technique that considers non-linear nature of the wind direction attribute. A transformation is done according to the work by Williams [33]. The wind speed and wind direction from each monitoring station is represented as a vector with the magnitude $S$ (wind speed) and direction $\theta$ (wind direction). The vector is divided into axial components $X$ (northern wind) and $Y$ (eastern wind):

$$X = S \sin \theta,$$

(12)
Based on these two components, two ArcGIS surface interpolations are created. The type of interpolation used is spline. As a result we get two grids: for northern $X'$ and eastern wind $Y'$. The magnitude of the vector, the wind

$$Y = S \cos \theta.$$
speed $S'$, is computed as:

$$S' = \sqrt{X'^2 + Y'^2}. \quad (14)$$

The calculation of wind direction angle $\theta'$ is more complicated. From geometry, the wind direction is calculated as $\theta' = \tan^{-1}(Y'/X')$. However, the inverse tangent is defined only for values between $-90^\circ$ and $90^\circ$ and it is only half of our domain. Therefore, each of the four quadrants of our domain (the quadrants are shown in Figure 10) requires its own formula [33]:

Quad I: $\theta' = \tan^{-1}(X'/Y')$, \quad (15)
Quad II: $\theta' = \tan^{-1}(Y'/X') + 90^\circ$, \quad (16)
Quad III: $\theta' = \tan^{-1}(X'/Y') + 180^\circ$, \quad (17)
Quad IV: $\theta' = \tan^{-1}(Y'/X') + 270^\circ$ \quad (18)

As a result we get interpolated values of wind speed and wind direction for each point of the studied space.

4.4 Data Uncertainty

The dispersion of a pollutant in a distribution region is not uniform, and intuitively the concentration of the pollutant near a chimney is higher than at a border of the dispersion region. Furthermore, pollutants are subject to decay and deposition processes. In other words, it is more likely that people living near an emitting facility are exposed to higher levels of pollutants than people who live kilometers away from the facility. Therefore, presence of a chemical
in a given point is uncertain and a probability of detecting it depends on a
distance from the point to the emission source. This dependency is inversely
proportional. For example, in Figure 3(c) the probability of detecting A at the
point \( g_{p1} \) is lower than at the point \( g_{p2} \).

Various functions can be used to determine the dependency of the pollutant
presence probability in a given point on the distance to the emitting facility.
For instance, with a categorical function (Figure 11(a)), we assign probabilities
according to distance ranges, e.g., 1.0 for 0-2 km from the facility, 0.75 for 2-4
km, 0.50 for 4-6 km, etc. Another example is a linear function (Figure 11(b))
which can be represented as \( 1 - x'/x \), where \( x' \) is the distance from a given point
to the facility and \( x \) is the maximal distance where pollutant distributes. In
this work we use a third example, the curve function (Figure 11(c)), which
is derived from the cosine function, \( p = \frac{\cos \pi x}{2} + 0.5 \). With this function
the probability decreases slowly with the increasing distance. Then, it starts
decreasing more linearly, and at the end slows down again. We believe that the
curve function models the real-life pollutant behavior more accurately than
the other two methods.

These three examples are only some of possible curves that can be used
to model pollutant distribution within buffer zones. However, other functions
could be used in order to improve the accuracy of the results. They could depend on the types of chemicals. For example, denser chemicals may settle out in a region closer to the emitting facility, while only small amounts would reach places at medium and far distances.

For datasets which contain other types of spatial objects, i.e. lines and polygons. Grid points intersecting a line or located inside a polygon are assigned a feature presence probability of one. For example, point $gp_3$ in Figure 12. Uncertainty for grid points positioned in buffer zones depends on the shortest distance from the point to the line or polygon. Points $gp_1$ and $gp_2$ in Figure 12 are located in buffer zones of line $L$ and polygon $P$ respectively. Existential probabilities at these points are computed using shortest distances to respective spatial objects.

5 Experimental Evaluation

5.1 Real Data

We conduct experiments on two real datasets which contain data on pollutant emission points and childhood cancer cases in the province of Alberta and Manitoba, Canada. The sources of the databases are the National Pollutant Release Inventory (NPRI, the data is publicly available) [12], and the provincial cancer registries. The National Pollutant Release Inventory (NPRI), established in 1992, is the national Pollutant Release and Transfer Register (PTRT) of Canada. The information on pollutants is taken for the period between 2002 and 2007 and contains the type of a chemical, location of release, and average amount of release per year. In order to get reliable results, the chemicals that had been emitted from less than three facilities are excluded from the dataset. There are 47 different chemicals and 1,442 pollutant emission points in Alberta; 26 different chemicals and 545 pollutant emission points in Manitoba, several chemicals might be released from the same location. The number of cancer points (the centroids of postal code regions where children lived when cancer was first diagnosed) are 1254 and 520 in Alberta and Manitoba, respectively.
Table 3  Detailed pollutants and cancers information in Alberta and Manitoba.

| Dataset | #Pollutants | #Cancers | #Cancers (urban) | #Cancers (rural) |
|---------|-------------|----------|------------------|-----------------|
| Alberta | 1,455       | 1,254    | 1,120            | 134             |
| Manitoba| 545         | 520      | 400              | 120             |

Environmental pollutants are suspected to be one of the causes of cancer in children. However, there are other factors that could lead to this disease (genetic susceptibility, parental exposure to chemicals or radiation, parental medical conditions, etc.). Considering this fact, we attempt to find “correlations” rather than “causalities”. The results are currently under careful evaluation by domain experts in our interdisciplinary team. It suffices to mention, however, that some surprising rules were discovered indicating significant association between groups of chemicals, not categorized individually as carcinogens, and children cancers. Additionally, we identified rules co-locating cancer cases with a pair of chemicals, one known as carcinogen and another with no-carcinogenic properties. Since the carcinogen alone did not correlate with cancer, the occurrence of those pairs of chemicals suggests a “catalyzing” effect by the non-carcinogen.

We are interested in co-location rules of the form $Pol \rightarrow Cancer$, where $Pol$ is a set of pollutant features and $Cancer$ is a cancer feature. The expected confidence is used as a prevalence measure. The distance between points in a grid is 1 km; the change in the grid granularity is also evaluated. The number of simulations (randomized datasets) for the statistical test is set to 99, so that with the observed data the denominator in Equation (9) is 100. The level of significance $\alpha$ is set to 0.05. The size of an antecedent of candidate rules is up to three. Larger candidates have low support values due to the fact that the average number of features in a transaction in the experiment is up to 3.

The randomized datasets that are used in the statistical test are generated as follows. Pollutant emitting facilities are not random and usually located close to regions with high population density, while they are not present in other places (e.g., in protected areas). Due to this observation, we do not randomize pollutant points all over the region, but instead keep locations of facilities and randomize pollutants within these positions. Among 1,254 cancer points in Alberta, 1,134 were located within “urban” municipalities (cities, towns, villages, etc.) and the rest were diagnosed in “rural” areas. For the cancer cases in Manitoba, 400 out of 520 were located in urban areas, while the rest, 120 in rural areas. In order to have the randomized cancer occurrence rate close to the real-world rate, we keep the number of cancer feature instances positioned in “urban” /“rural” regions the same as in the real dataset. The number of random cancer cases placed within each “urban” municipality is directly proportional to the number of children counted in the 2006 census [13]. The rest of the cancer cases which are located in rural regions in the real datasets, are randomly placed on the map of Alberta and Manitoba but not in urban areas. The detailed information of pollutants and cancers in Alberta and Manitoba is displayed in Table 3.
5.1.1 Comparison with Certain Data Method

We compare the results of our uncertain data method (UM) with the results of a method using certain deterministic data (CM) where existential probabilities are not stored as a part of a transactional database. As an interestingness measure in the CM we use the confidence $\text{Conf}(\text{Pol} \rightarrow \text{Cancer})$, which is a fraction of transactions containing all features in Pol that also include the cancer feature.

$$
\text{Conf}(\text{Pol} \rightarrow \text{Cancer}) = \frac{\text{Sup}(\text{Pol} \cup \text{Cancer})}{\text{Sup}(\text{Pol})}.
$$

(19)

In Alberta, the number of significant co-location rules detected by both UM or CM together is 496; from these 204 rules are found by both methods, 278 rules are identified only by UM, and 14 rules by CM. In Manitoba, the number of detected significant co-location rules by UM and CM is 362 and 263, respectively, and these two methods have 232 common rules. From these two datasets, we can find that UM method covers most of significant co-location rules detected by CM method, moreover, UM captures more statistical significant rules that can not be mined by CM otherwise. The difference in results could be explained by the fact that our approach deals with probabilities of feature presence in transactions rather than with deterministic values. It considers not only the presence of a feature in a transaction but also distances from grid points to pollutant features and cancer cases. The grid points that are situated closer to spatial instances are given more weight than points located relatively farther.

Some of the co-location rules discovered by the uncertain method have a low level of $\text{ExpSup}(\text{Pol})$ or $\text{ExpSup}(\text{Pol} \cup \text{Cancer})$. For example, 348 out of 482 rules in the Alberta dataset and 193 out of 362 rules in the Manitoba have $\text{ExpSup}(\text{Pol} \cup \text{Cancer})$ less than 1. It means either a low number of transactions or a relatively long distance from grid points. Although with a low $p$-value ($\leq 0.05$) they have an expected confidence level higher than in most randomized datasets, domain experts might not be interested in these co-location rules. In that case, a threshold on the expected support might be introduced to the model for detection of significant co-location rules. This threshold should not be set too high, so that the algorithm does not miss some of the interesting co-location rules or patterns with rare features.

5.1.2 Effect of Randomization

As mentioned above, in the randomized datasets, we take an intuitive reasonable strategy by randomizing the pollutant emitting facilities and cancer cases. We randomize the pollutant emitting facilities in the areas with high density population, and randomize the cancer cases according to the population distribution in “urban” and “rural” areas, moreover, the number of cancer cases in “urban”/“rural” regions are the same as the real dataset.
We compare this randomized strategy with other two randomized methods: randomizing the pollutants only and randomizing cancer cases only.

First in the Alberta dataset, it is discussed above that when we randomize both pollutants and cancers, we get 488 significant co-location rules by UM. Then we try to fix the cancer cases as the real dataset and only randomize the pollutant emitting facilities among the high density population regions. Up to 710 rules are reported under this setting. However, when we fix the pollutant facilities and only randomize the cancer cancers, only 127 rules are detected.

Then in the Manitoba dataset, we found 362 significant rules with both pollutant facilities and cancer cases randomized. 243 and 314 significant rules are reported when we only randomize the pollutant facilities and only randomize cancer cases, respectively.

All of three randomize strategies ensure that in the randomized datasets, the pollutant emitting facilities and the cancer cases are independent from each other. In our experiment, we take the first randomization strategy, by randomizing both the pollutant facilities and cancer cases which is more reasonable. In the future, we need our domain experts to evaluate the detected co-location rules to help us choose the best randomization method.

5.1.3 Effect of the Grid Granularity

As already mentioned, a granularity of the grid (a distance between grid points which affects the number of points per unit of space) is crucial for the result accuracy. A great distance between grid points may lead to omission of some regions of the space especially when the average buffer distance is short. On the other hand, when the distance between points is too short, more transactions are derived by the algorithm. Decreasing the distance by a factor of two increases the transaction set size approximately by four times. Therefore, more computation needs to be done during the statistical test step. The grid resolution might be set up depending on the average buffer size.

For the Alberta dataset, in addition to the grid with a distance of 1 km between its points, we conduct two experiments with 2 and 0.5 km grids. As mentioned above, the algorithm reports 482 significant co-location rules with 1 km grid. With 2 km granularity 547 rules are detected from which 335 are present in both 1 and 2 km result sets, and 212 are unique for 2 km grid. The difference means that 2 km distance between grid points is too long for our dataset, where the average buffer size is 7.3 km, and its accuracy is comparatively low due to the smaller number of transactions which is not sufficient to capture intersections of instance buffers accurately. The 0.5 km granularity grid reports 472 co-location rules as significant. From these, 426 are found with both 1 and 0.5 km grids, and 46 rules are identified only by 0.5 km grid. As we can see, the difference between 0.5 km and 1 km result sets is smaller than that between 1 km and 2 km grids. As the distance between points in a grid decreases, the accuracy of the results improves.

We also conduct three sets of experiments with different grid granularity, 0.5 km, 1 km and 2 km on the dataset of Manitoba. With 1 km grid granularity,
we detect 362 rules significant co-location rules. When the grid distance is increased to 2 km distance, 280 rules are reported and a large portion of them (271) rules also appear in 1 km result set. 364 significant co-location rules are found with 0.5 km grid granularity. Among these 364 rules, 356 rules are found in both 1 km and 0.5 km grid granularity. As can be observed, the difference between 1 km and 0.5 km grid distance is very small, and 1 km grid distance can cover most of significant rules detected by both 2km and 0.5km grid distance. Therefore, in our experiment, we choose 1km grid distance without loss of accuracy and efficiency. The industry emitted chemicals in Alberta and Manitoba are not the same, indeed they overlap very lightly. Therefore, we did not compare the set of significant co-location rules discovered in both datasets.

5.1.4 Effect of Filtering Technique

In the Alberta dataset, the number of candidate co-location rules in the experiment (i.e. not yet determined as significant) is 17,343 (co-locations with the antecedent size up to three). With the naïve approach all candidates are checked in each of the 99 simulation runs which results in a large amount of computation. After the exclusion of rules with zero-level confidence, 10,125 candidates remain in these two datasets which also form a big set. Figure 13(a) show that the usage of the second filtering method (the exclusion of candidates which p-value) passed α during the evaluation of randomized datasets) considerably reduces the amount of computation. In the first simulation run the confidence value is computed for 10,125 rules, while 3,098 candidate rules are checked in 20th simulation, and only 488 candidates are evaluated in the last run.

While in the Manitoba dataset, the total number of the candidate co-location rules with up to three pollutants is 2,951. Only 665 rules are left if we exclude the zero-level confidence co-location rules. It means that the naïve method can already help us prune a large amount of unnecessary rules. Figure 13(b) shows the usage of the second filtering method, in the first simulation run the 665 rules are checked, the candidate rules reduce to 408 in the 10th simulation which reduce around \( \frac{1}{3} \) candidate rules, and in the last simulation only 362 candidates are checked.

5.2 Synthetic Data

We conduct experiments on synthetic datasets to demonstrate that our framework can discover a correct set of co-location rules. In addition, we show that our transaction-based method takes into account spatial context and information.
5.2.1 Discovery of Co-Location Rules

In order to evaluate our algorithm on the synthetic data we generate a dataset and attempt to emulate the real-world information. Similarly to the real dataset, it contains point features that appear in the antecedent part of co-location rules (features $C$), and a disease feature $D$ as a consequent. The study region is a 100×100 unit square. The buffer size is 1 unit. We simulate 7 chemicals by using $C = \{C_1, ... , C_7\}$, and $D$ is simulated as the cancer. The features $C_1$ and $C_2$ have 20 instances each and they are associated with each other. The features $C_3$ and $C_4$ have 30 points each; 20 of them are associated with each other, while remaining 10 instances are placed randomly. These two pairs represent co-located chemicals. The disease feature $D$ is positively associated with sets $C_1 \cup C_2$, $C_3 \cup C_4$, and with 30 out of 40 instances of feature $C_5$. It has no association with the feature $C_6$ (30 instances), and negatively correlated with $C_7$ (30 points), so that no pair of instances $D$ and $C_7$ are neighbors. In addition there are 30 disease cases spread randomly. We look for co-location rules of the form $C \rightarrow D$. In 99 randomized datasets all eight features are distributed randomly with no association (neither positive nor negative) with each other.

The significant co-location rules with $p$-value $\leq 0.05$ are shown in Table IV. As expected, rules 5, 6, and 11 are reported as significant because they have strong correlation of $C$ features and feature $D$. Rules 1-4 are also detected because either total ($C_1 - C_2$) or partial of ($C_3 - C_4$) have associations with $D$. Rules with features $C_6$ and $C_7$ are not reported because of their zero and negative association with the disease feature. The remaining co-location rules (7-10, 12-14) are detected due to their random correlation with features associated with feature $D$. However, they all have very low $ExpSup(C + D)$ values and can be pruned if a threshold on $ExpSup$ is used as discussed in the experiments with the real data.

The experiment on synthetic data shows that our approach finds co-location rules in which features in the antecedent part are co-located with
Table 4: Co-location rules detected in synthetic data. *ExpSup* is the value of the expected support of patterns of the form \( C + D \), where \( C \) is the set of cause features and \( D \) is the disease feature.

| N | Co-location Rule | *ExpSup* | *ExpConf* |
|---|------------------|----------|-----------|
| 1 | \( C_1 \rightarrow D \) | 763.1 | 0.41 |
| 2 | \( C_2 \rightarrow D \) | 765.8 | 0.42 |
| 3 | \( C_3 \rightarrow D \) | 717.1 | 0.26 |
| 4 | \( C_4 \rightarrow D \) | 807.8 | 0.30 |
| 5 | \( C_5 \rightarrow D \) | 1,256.6 | 0.34 |
| 6 | \( C_1 + C_2 \rightarrow D \) | 432.8 | 0.50 |
| 7 | \( C_1 + C_4 \rightarrow D \) | \( 1.0 \cdot 10^{-3} \) | 0.82 |
| 8 | \( C_1 + C_5 \rightarrow D \) | 10.6 | 0.49 |
| 9 | \( C_2 + C_4 \rightarrow D \) | 0.4 | 0.49 |
| 10 | \( C_2 + C_5 \rightarrow D \) | 14.4 | 0.44 |
| 11 | \( C_3 + C_4 \rightarrow D \) | 390.5 | 0.53 |
| 12 | \( C_5 + C_6 \rightarrow D \) | 2.8 | 0.08 |
| 13 | \( C_1 + C_2 + C_4 \rightarrow D \) | \( 4.8 \cdot 10^{-4} \) | 0.83 |
| 14 | \( C_1 + C_2 + C_5 \rightarrow D \) | 7.9 | 0.51 |

the feature in the consequent part. A threshold with a relatively low value can help to exclude rules with noise features.

5.3 Distance between Features

In this experiment we evaluate the effect of an average distance between features on the expected support. Recall that one of the advantages of our algorithm is that it takes into account a distance between spatial objects, so two objects located close to each other are represented in more transactions than a pair of objects situated farther (Figure 4). Let us consider two scenarios: 1) objects which belong to two distinct features are located on average very close to each other, and 2) they are situated on the farthest possible distance so they are still considered to have neighbourhood relationships. Most previous approaches assign the same prevalence measure value in both cases as long as a neighbourhood relationship is kept. Obviously, it is not correct; the prevalence measure should be higher in the first situation. On the other hand, with our approach in the first case the features are included in more transactions with higher existential probabilities. This leads to a higher prevalence measure than in the second case.

For this experiment we create synthetic datasets with two spatial features \( f_1 \) and \( f_2 \). The study region is a 100×100 unit square. The buffer size is 1 unit. In each dataset, features have 30 instances each. We randomly place the instances of feature \( f_1 \) in the study region. One instance of feature \( f_2 \) is placed on a varying distance \( d \) from an instance of \( f_1 \). The distance \( d \) between instances of two features is taken randomly from ten ranges \( [0.0, 0.2), [0.2, 0.4), \ldots, [0.8, 1.0) \).
Table 5 The average expected support for ranges of an average distance between two spatial features.

| N  | Range      | Average ExpSup |
|----|------------|----------------|
| 1  | [0.0, 0.2) | 1,558.9        |
| 2  | [0.2, 0.4) | 1,355.3        |
| 3  | [0.4, 0.6) | 1,017.1        |
| 4  | [0.6, 0.8) | 649.9          |
| 5  | [0.8, 1.0) | 353.3          |
| 6  | [1.0, 1.2) | 155.9          |
| 7  | [1.2, 1.4) | 52.8           |
| 8  | [1.4, 1.6) | 16.6           |
| 9  | [1.6, 1.8) | 8.3            |
| 10 | [1.8, 2.0) | 5.7            |

0.4), ..., [1.8, 2.0]) (given in units). The first range [0.0, 0.2) is for the scenario when features are located very close to each other on average. The last range [1.8, 2.0) simulates a situation when an intersection of each pair of instance buffers is very small. The expected support of pattern \((f_1 \cup f_2)\) is calculated and averaged over 100 synthetic datasets for each of ten ranges.

The results are presented in Table 5. As can be observed, the expected support rapidly decreases with the increase in the average distance between instances of features \(f_1\) and \(f_2\). Expectedly, the range [0.0, 0.2) gets the highest value of the expected support, and the range [1.8, 2.0) has the lowest prevalence value. While a pattern with these features would be considered having the same prevalence measure value in all ten synthetic datasets by most previous algorithms, our transaction-based approach takes into account the actual spatial information and a relative proximity or remoteness of features from each other.

6 Conclusion

Co-location pattern and rule mining is one of the tasks of spatial data mining. Discovery of co-location patterns and rules can be useful in many projects and applications and may lead to the discovery of new knowledge in various domains. In this paper we propose a new co-location mining framework which combines classical co-location mining, and uncertain frequent pattern and association rule mining. The approach was motivated by a real-world application of detecting possible associations of pollutant emission points and childhood cancer cases. We take into account some of the limitations that can prevent previously proposed approaches from being used in some real-world applications and domains. Our novel transactionization method allows conversion of spatial data into a set of transactions by imposing a regular grid over a given map. Each grid point can be seen as a representation of a study region. Features of objects and their buffers that contain a grid point
form a transaction. In addition, our approach takes into account uncertainty of data by storing feature existence probabilities in transactions in order to simulate the real scenarios. A probability of feature presence in a transaction depends on a distance from the feature instance to the respective grid point. A usage of user-defined thresholds on prevalence measures like in previous algorithms is replaced by the statistical test which helps to identify significant co-location patterns and rules that are unlikely to occur only by chance. In order to decrease computation, the filtering techniques are presented which prune candidate patterns and rules that appear to be definitely not significant.

The experiments on two real and synthetic datasets show that our approach finds significant co-location patterns and rules. We deploy three different randomization strategies to find significant co-location rules. The effect of grid granularity is also evaluated. A dependence of a prevalence measure value on an average distance between feature instances is shown. The usage of transactions preserves spatial context and information such as relative locations of instance objects and distances between them. The consideration of feature presence probabilities helps to distinguish various cases when feature instances are situated at different distances from grid transaction points. We also demonstrate that the difference in the results obtained by our uncertain data model and certain data method can be explained and justified.

The motivating application of this paper has its unique challenges. We examine several factors which affect dispersion of pollutants in air. In order to more accurately model chemical distribution we used buffer zones differing in their sizes which depend on released amounts. Circular buffers transformed into elliptical figures with the consideration of wind speed and its direction at locations of emitting facilities. Finally, we model uncertainty of a pollutant presence at transaction points. In addition to pollution, other factors can also cause cancer in children. In this paper we do not intend to find true causalities but attempted to identify possible associations of pollutants and childhood cancer. The results that are derived by our algorithm can be useful for domain experts and help in further analysis of pollutant-cancer relationships. These algorithms we propose can also be used to discover co-locations for other diseases and other multiple factors.

References

1. ArcGIS Desktop: Release 10 (ESRI 2011)
2. (ACIS), A.I.S.: Live alberta weather station data. http://www.agric.gov.ab.ca/app116/stationview.jsp
3. Adilmagambetov, A.; Zaiane, O.R.; Osornio-Vargas, A.: Discovering co-location patterns in datasets with extended spatial objects. In: DaWaK’13, pp. 84–96. Springer (2013)
4. Aggarwal, C.C., Li, Y., Wang, J., Wang, J.: Frequent pattern mining with uncertain data. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 29–38 (2009)
5. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec. 22, 207–216 (1993)
6. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of the 20th International Conference on Very Large Databases, pp. 487–499 (1994)
7. Al-Naymat, G.: Enumeration of maximal clique for mining spatial co-location patterns. In: Proc. of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 126 –133 (2008)
8. Armstrong, B., Doll, R.: Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. International Journal of Cancer 15(4), 617–631 (1975)
9. Barua, S., Sander, J.: SSCP: mining statistically significant co-location patterns. In: Proc. of the 12th international conference on Advances in Spatial and Temporal Databases, pp. 2–20 (2011)
10. Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Zuefle, A.: Probabilistic frequent itemset mining in uncertain databases. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 119–128 (2009)
11. Boffetta, P., Nyberg, F.: Contribution of environmental factors to cancer risk. British medical bulletin 68(1), 71–94 (2003)
12. Canada, E.: National Pollutant Release Inventory. Tracking Pollution in Canada. http://www.ec.gc.ca/inrp-engC/.
13. Canada, S.: 2006 Census. http://www12.statcan.gc.ca/census-recensement/2006/index-eng.cfm
14. Chou, Y.H.: Exploring spatial analysis in geographic information systems (1997)
15. Chui, C.K., Kao, B.: A decremental approach for mining frequent itemsets from uncertain data. In: Advances in Knowledge Discovery and Data Mining, pp. 64–75 (2008)
16. Chui, C.K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. In: Advances in Knowledge Discovery and Data Mining, pp. 47–58 (2007)
17. Cressie, N.A.: Statistics for Spatial data (1991)
18. Data, E.C.N.C., Information: Canadian climate normals or averages 1971-2000. http://climate.weatheroffice.gc.ca/climate_normals/index_e.html
19. Ester, M., Kriegel, H.P., Sander, J.: Algorithms and applications for spatial data mining. In: Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, pp. 160–187 (2001)
20. Estivill-Castro, V., Lee, I.: Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In: Proc. of the 6th International Conference on Geocomputation, pp. 24–26 (2001)
21. Estivill-Castro, V., Murray, A.T.: Discovering associations in spatial data?an efficient medoid based approach. In: Research and Development in Knowledge Discovery and Data Mining, pp. 110–121 (1998)
22. Getis, A., Jackson, P.H.: The expected proportion of a region polluted, by k sources. Geographical Analysis 3(3), 256–261 (1971)
23. Huang, Y., Pei, J., Xiong, H.: Mining co-location patterns with rare events from spatial data sets. Geoinformatica 10(3) (2006)
24. Huang, Y., Shekhar, S., Xiong, H.: Discovering colocation patterns from spatial data sets: A general approach. IEEE Trans. on Knowl. and Data Eng. 16(12), 1472–1485 (2004)
25. Huang, Y., Zhang, P.: On the relationships between clustering and spatial co-location pattern mining. In: Proc. of the 18th IEEE International Conference on Tools with Artificial Intelligence, pp. 513–522 (2006)
26. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In: Proc. of the 4th International Symposium on Advances in Spatial Databases, pp. 47–66 (1995)
27. Kwan Kim, S., Kim, Y., Kim, U.: Maximal cliques generating algorithm for spatial co-location pattern mining. In: Secure and Trust Computing, Data Management and Applications, pp. 241–250 (2011)
28. Morimoto, Y.: Mining frequent neighboring class sets in spatial databases. In: Proc. of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 353–358 (2001)
29. Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., Yang, D.: H-mine: Hyper-structure mining of frequent patterns in large databases. In: Proc. of 2001 IEEE International Conference on Data Mining, pp. 441–448 (2001)
30. Reggente, M., Lilienthal, A.J.: Using local wind information for gas distribution mapping in outdoor environments with a mobile robot. In: Sensors, 2009 IEEE, pp. 1715–1720 (2009)
31. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In: Proc. of the 7th International Symposium on Advances in Spatial and Temporal Databases, pp. 236–256 (2001)
32. Shekhar, S., Zhang, P., Huang, Y.: Spatial Data Mining (2010)
33. Williams, R.G.: Nonlinear surface interpolations: Which way is the wind blowing? In: Proc. of 1999 Esri International User Conference (1999)
34. Xiao, X., Xie, X., Luo, Q., Ma, W.Y.: Density based co-location pattern discovery. In: Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1–10 (2008)
35. 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: Proc. of 2004 SAIM International Conference on Data Mining, pp. 78–89 (2004)
36. Yoo, J.S., Shekhar, S.: A joinless approach for mining spatial colocation patterns. IEEE Trans. on Knowl. and Data Eng. 18(10), 1323–1337 (2006)
37. Yoo, J.S., Shekhar, S., Smith, J., Kumquat, J.P.: A partial join approach for mining co-location patterns. In: Proc. of the 12th ACM International Workshop on Geographic Information Systems, pp. 241–249 (2004)