University of Alberta

DISCOVERING CO-LOCATION PATTERNS AND RULES IN UNCERTAIN SPATIAL DATASETS

by

Aibek Adilmagambetov

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2012
Edmonton, Alberta

Abstract

Co-location mining, which focuses on the detection of co-location patterns, is one of the tasks of spatial data mining. A co-location pattern is a set of spatial features frequently located in close proximity of each other. Most previous works are based on transaction-free apriori-like algorithms which use user-defined thresholds and are designed for point objects. 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 geographic space and can be extended for spatial extended objects. Uncertainty of a 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 real and synthetic data. In addition, we explain the data modeling framework which is used on a real dataset of pollutants and childhood cancer cases.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

A motivating application of this thesis is the detection of possible spatial associations of different chemicals and cases of childhood cancer. Cancer, a class of diseases characterized by uncontrolled growth of abnormal cells, their invasion into other tissues, and metastasis, is one of the leading causes of death in both developed and developing world. Although some people are genetically predisposed to a high risk of developing cancer, most cases of this disease are caused, at least partially, by 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 children's cancer cases are rare and the levels of exposure to environmental factors are hard to evaluate.

We are interested in a discovery of co-location rules in a dataset which contains information on chemical emission points and amounts of release, and childhood cancer cases in the province of Alberta, Canada. Figure 1.1 displays part of the dataset with rectangles representing pollutant emission points, triangles - cancer cases, and polygons - 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.

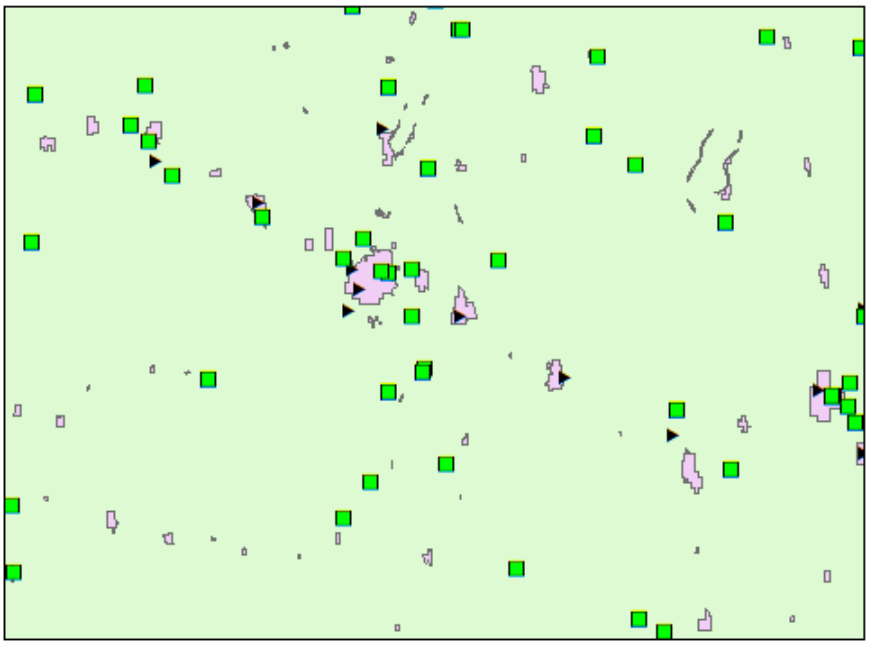


Figure 1.1: Part of the dataset: rectangles - pollutant points, triangles - cancer cases, polygons - urban municipalities.

The goal of co-location mining is to discover patterns of spatial features that are often located close to each other in geographic proximity. An example of a co-location pattern is a combination of symbiotic species of plants and animals depending on ecological conditions. Figure 1.2 illustrates a sample spatial dataset with point features. As can be observed, instances of feature "+" are often located close to instances of "$\circ$". Similarly, objects of feature "$\star$" are seen close to instances of "$\triangledown$".

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 interesting for domain experts so they can reduce a range of possible patterns that are hidden in datasets 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.

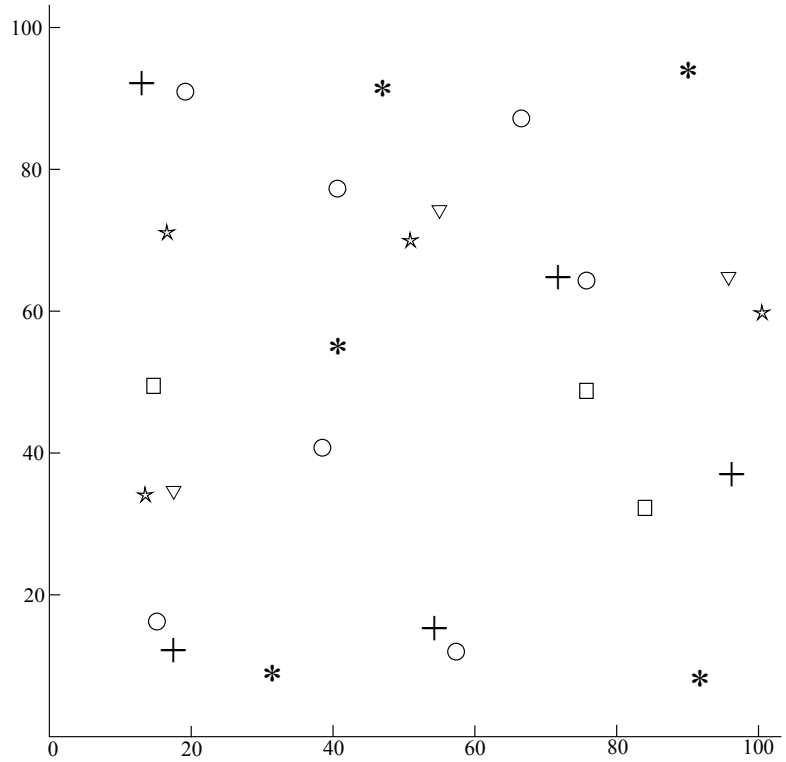


Figure 1.2: Sample dataset with point spatial features. Instances of feature sets $\{+, \circ\}$ and $\{\star, \triangledown\}$ are often located close to each other.

Most of the existing approaches to the co-location mining problem [29, 21, 36, 33] 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 neighborhood 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, information in some datasets is uncertain; a presence of a feature in a region could depend on different factors. For example, a pollutant released from a facility distributes according to climatic factors in its area, and a probability of detecting chemical in the region close to the emission point is higher than in remote regions.

1.2 Thesis Statements

In this thesis we address the challenges related to the co-location pattern and rule mining problem. We study the feasibility of resolving these challenges by claiming the following statements:

1. A spatial dataset can be transactionized that is transformed into a set of transactions which better preserves spatial information than previous approaches and may be used to calculate prevalence measure values of candidate co-location patterns and rules.

2. It is possible to model uncertainty of feature presence in transactions created from spatial datasets and use the expected support and expected confidence as prevalence measures in the field of co-location mining.

3. The statistical test may be used in co-location mining to replace thresholds when identifying significant co-location patterns and rules.

1.3 Thesis Contributions

This thesis makes the following contributions:

1. We propose a new framework which combines co-location mining with frequent pattern and association rule mining. Instead of user-defined thresholds on prevalence measures, the statistical test is used to determine the significance of co-location patterns and rules. A co-location is considered significant, if it has a surprisingly high level of the prevalence measure in a real dataset in comparison to randomized, or simulated, datasets. The statistical test ensures that discovery of co-location patterns and rules is not affected by the level of the prevalence threshold. Randomized datasets are built under the null hypothesis that the features of a co-location are independent from each other. When creating simulated datasets for our real application, we attempt to preserve spatial distribution of features. Same numbers of cancer cases are placed within urban and rural areas as it is in a real dataset. The number of

cases within each urban municipality is proportional to its child population. Due to the fact that most emitting facilities are located within certain regions, we do not randomize pollutant points all over the study region. Instead we keep locations of facilities and randomize chemicals within them.

2. We present a new method of transforming a spatial dataset into a set of transactions. Initially, buffers are created around spatial objects; their size can depend on internal and external factors. A grid, which consists of points placed regularly in a study region, is used to create transactions. Each grid point is a representation of a respective parcel of the study area. A transaction is a set of features which instances or their buffers intersect with the grid point. In addition, uncertainty of feature presence at a given point is introduced to the model. It is modeled as a dependence on a distance from the spatial object to the grid point.

3. A usage of the statistical test requires many computations to be done when calculating prevalence measure values of candidate patterns and rules. To address this challenge, we introduce two filtering techniques that help to decrease the computation cost. These filters prune candidate patterns and rules that appear to be clearly not significant during the statistical test. Therefore, there is no need to compute their interestingness measures; less calculation is performed.

4. Finally, instead of using circular buffers in experiments with the real dataset containing information on pollutant emission points and childhood cancer cases, we transform buffers into ellipses by taking into account several factors that affect pollutant distribution: pollutant release amounts, wind speed and direction data, and uncertainty of chemical presence. We believe that the consideration of these factors while creating buffer zones around emitting facilities improves the accuracy of the results.

1.4 Thesis Outline

The remainder of the thesis is organized as follows:

- Chapter 2 gives an overview of the most important studies related to co-location mining, frequent pattern and association rule mining with certain and uncertain data. After a short description of main spatial data mining tasks, we explain current approaches and frameworks used to tackle one of them, a co-location pattern and rule mining problem. A classification of co-location mining algorithms is given along with an explanation of pattern interestingness (prevalence) measures and algorithm designs. Then, we briefly describe frequent pattern and association rule mining. We also explain frequent pattern mining techniques used to identify patterns within uncertain data.

- In Chapter 3, we first explain limitations of a current framework used in most co-location mining approaches. We address the problem of mining co-location patterns and rules in datasets which along with point features may contain extended objects such as polygons and lines. In addition, uncertainty of data is also taken into account. We give an outline of our algorithm, and explain the transactionization step and interestingness measures used in our model. In addition, we describe filtering techniques that reduce computations by excluding some candidate patterns and rules.

- Chapter 4 describes challenges of mining co-location rules in a real-world application of identifying spatial associations between pollutants and childhood cancer cases. We explain our modeling framework which helps to model the real world more accurately. The factors explained in this chapter include pollutant release amounts, an average wind speed and prevailing wind direction at emission points, and uncertainty of a presence of pollutants within buffer zones.

- In Chapter 5, we report the results of the experiments conducted on real and synthetic data. A real dataset contains data on chemicals and childhood cancer cases. First, we evaluate the effect of filtering techniques on the number

of checked candidate patterns. Second, we compare our uncertain data model with a model used to mine certain data. Finally, we study the effect of the granularity of a grid used in a transactionization step. In addition to the real dataset, we evaluate our approach on synthetic data to show that our algorithm finds a correct set co-location rules placed in a synthetic dataset, and better preserves and deals with spatial context and information.

- Finally, Chapter 6 summarizes the conclusions of the thesis. In addition, we explore some of the unresolved challenges and discuss the directions for future research in the field of co-location pattern and rule mining.

Chapter 2

Related Work

This section starts with a short introduction to spatial data mining and some of its applications and tasks. Then it gives an overview of significant studies related to co-location mining, frequent pattern 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 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 [16, 30].

- **Spatial clustering** - the task of grouping data objects into clusters such that the members of a cluster are more similar to each other than to the objects of other clusters. An example of clustering is creation of thematic maps based on climatic, biologic or ecological information.

- **Spatial characterization** - the process of compactly describing a selected subset of a database. For example, one might be interested to find out what are unique characteristics of city regions with high crime rate.

- **Spatial trend detection** - the detection of regular changes of some non-spatial attributes that depend on a distance from a given start point. For instance, a change in precipitation level in some regions can be considered as a trend.

- **Spatial classification** - the task of assigning labels (classes) to data objects based on their and their neighbors attribute values. The classification of eco-regions from satellite images is an example of this task.

- **Outlier detection** - the detection of objects that are inconsistent with the remainder of a dataset. An expensive house in a poor neighborhood is a spatial outlier. Detection of spatial outliers might be useful in many applications and may help to discover new knowledge in spatial datasets.

- **Prediction of events** - the problem of predicting occurrence of events based on spatial and non-spatial data. For instance, prediction of watering place lo-

cations looking on precipitation levels, water reservoirs, occurrence of predators, etc., uses predictive models.

- **Co-location mining** - the discovery of sets of features than are often located close to each other in geographic space. The examples of co-location rules and patterns include symbiotic animal and plant species, climatic events, and other geographic patterns.

2.2 Co-Location Mining

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

2.2.1 Spatial Statistics Approaches

Spatial statistics methods deploy statistical techniques such as cross K-functions with Monte-Carlo simulations [14], mean nearest-neighbor distance, and spatial regression models [11] 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 use the similarity of co-location mining and 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 2.1.

In order to reduce the computation time only frequent $(k-1)$-size patterns are used to generate k-patterns (patterns that consists of k items). It is possible due to the apriori, or downward closure, principle: the subsets of a frequent pattern must also be frequent. However, there is a significant difference between association rule mining and co-location mining problem. In association rule mining transactions

Table 2.1: Comparison of association rule mining and co-location mining [21].

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

are independent from each other. In contrast, in co-location mining problem spatial objects are embedded into geographic space and it is not easy to define explicit transactions.

Spatial data mining approaches could be categorized into transaction-based which work with transactions and spatial join-based methods which use spatial joins of instance tables or feature layers.

Transaction-Based Approaches

Transaction-based approaches work by creating transactions over space and using association rule mining algorithms [5, 24, 26]. One of these methods, a reference-centric model [24], creates transactions around a reference feature specified by a user. Each set of spatial features that form neighborhood 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 [29], 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. The model has a major drawback 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 pairwisely neighbors) in spatial data are proposed to be used as transactions [6, 25], but this approach does not preserve the information on how close or remote are objects in cliques as long as they are considered being neighbors.

Spatial Join-Based Approaches

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 [18]. In another work [17], a map layer is constructed for each spatial feature based on clusters of instances or boundaries of clusters. The authors 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.

Clustering is used in a similar approach [23]. For two spatial features f_1 and f_2, if a 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, 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 a case in some real-world applications.

Another type of spatial join-based methods - instance-join algorithms - is similar to classical association rule mining. One of the first proposed co-location pattern mining frameworks of this type [29, 22] is based on neighborhood relations and participation index concept.

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:

$$PI(C) = min_{f_i \in C}\{pr(C, f_i)\}, \tag{2.1}$$

where $pr(C, f_i)$ is the participation ratio of a featuref_i in a co-location C and it is computed as:

$$pr(C, f_i) = \frac{number\ of\ distinct\ instances\ of\ f_i\ in\ instances\ of\ C}{total\ number\ of\ instances\ of\ f_i}. \tag{2.2}$$

A co-location pattern is considered prevalent, or interesting, if its PI exceeds a user-defined threshold. In other words, for each feature of the prevalent pattern at least $PI\%$ instances of that feature form a clique with the instances of all other features of the pattern according to the neighborhood relationship. Similarly to association rule mining, this framework is based on the apriori principle. Therefore, only significant, or frequent, $(k-1)$-size patterns are used for k-size 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:

$$cp(C_1 \rightarrow C_2) = \frac{number\ of\ distinct\ instances\ of\ C_1\ in\ instances\ of\ C_1 \rightarrow C_2}{total\ number\ of\ instances\ of\ C_1}. \tag{2.3}$$

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. Huang et al. [21] 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:

$$maxPR(C) = max_{f_i \in C}\{pr(C, f_i)\}. \tag{2.4}$$

It is not well explained how the algorithm deals with noise features. For example, if some features have only 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 neighborhood distance threshold. A set of spatial instances in a cell form a clique. Join operations are required to identify neighborhood relationships divided by boundaries of cells. Even though this approach reduces the computation time, it still requires 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 neighborhood relationships. The main idea is to find star neighborhoods instead of calculating pairwise distances between all instances in a dataset. The neighborhood 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 joinless approach has better performance in terms of the running time than the join-based algorithm.

In their work, Xiao et al. [32] 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 works extended the basic co-location mining framework [29]. For example, Zhang et al. [39] proposed an approach that extends the notion of co-location patterns and detects star, clique, and generic co-location patterns. Some of the works focused on identifying maximal [34] and closed co-location patterns [35]. A co-location pattern P is said to be maximal, if it is prevalent but no super event set of P is prevalent. A co-location pattern P is considered closed, if there is no super set $P' \supset P$ such that $PI(P') = PI(P)$. Mining maximal and closed patterns might be useful in situations when researchers are interested in frequent patterns of the maximal size.

Xiong et al. [33] introduced 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 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 neighborhood 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 a low threshold, and a high threshold may prune interesting rare patterns. Instead of a threshold-based approach, Barua and Sander [7] 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 α is a level of significance.

2.3 Frequent Pattern Mining

The concepts of frequent pattern and association rule mining were first introduced by Agrawal et al. [4]. Various approaches to these problems have been proposed over past two decades. Apriori [5] 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 association rule mining 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.

$$conf(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}. \tag{2.5}$$

The FP-growth algorithm, proposed by Han et al. [20], does not need a candidate generation process, therefore helping to reduce the computation time. The method uses a divide-and-conquer approach. After a first scan of a database a list of frequent items is formed where items are sorted by frequency in descending order. Infrequent items are removed. By using this list, the transactional database is converted into the FP-tree (frequent pattern tree). For each item i in the tree, the algorithm composes a sub-database (a set of transactions which have i as prefix). Then, this process is repeated recursively for each sub-database. Frequent patterns are obtained by concatenating a suffix pattern with frequent patterns from the FP-tree.

Data in transactional databases can be stored in several ways. The most common is a horizontal data layout where each transaction contains its ID and a list of items that are present in that transaction. The Apriori and FP-growth algorithms are designed to work with this type of transactional databases. The second widely used format is a vertical data layout. In databases which use this data storage format, a dataset is represented as a set of items. Each item is stored with a set of transaction IDs in which this item is present.

Zaki [38] proposed Eclat (Equivalence CLASS Transformation), an algorithm which employs datasets in vertical data format. A first scan of a dataset creates transaction sets for each item. At each level of the approach $(k+1)$-size itemsets are generated from frequent k-size itemsets by intersecting their transaction sets. The support is easily calculated because all information is already present for each itemset, so there is no need to scan the database. The process is repeated until no more candidate itemsets are left.

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 in certain. For example, market basket

Table 2.2: Example of a transactional dataset.

ID	Transaction
1	A(0.7); B(1.0); C(0.2)
2	A(0.9); D(0.5); E(0.4); F(0.8)
3	B(0.3); D(1.0); G(0.7)
4	A(0.1); B(0.6); C(0.7); E(0.2); G(0.4)
5	C(0.5); D(0.2); E(0.8)
6	B(0.6); C(0.3); E(1.0); F(0.4)

datasets are certain and precise. However, in some applications and domains 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 2.2.

Most studies use a notion of expected support [13, 12] 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 presence of I in each of 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. [3]. 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 [27]). According to this study, while FP-growth is efficient and scalable in the deterministic case, its extension to 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. [8] 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 $O(|T|minsup)$, where $|T|$ is the number of transactions and $minsup$ is a user-defined threshold. Without it the approach runs in exponential time. However, the algorithm requires the $minsup$ threshold to be defined, and it is nontrivial to apply the statistical test to the frequentness probability.

Chapter 3

Algorithm

The goal of this thesis is to design an algorithm that takes into account the limitations of previously proposed approaches to the co-location mining problem which are listed in the beginning of this chapter. Then, we present our novel algorithm which combines co-location mining and frequent pattern mining problems and uses the statistical test to identify significant co-location patterns and rules. The main steps of the algorithm are explained in detail. Furthermore, we explain two filtering techniques that reduce the amount of computations by pruning definitely insignificant candidate patterns or rules, thus decreasing the number of candidates to be checked.

3.1 Limitations of Previous Approaches

Various approaches to the co-location mining problem have been proposed during the past decade. Most of them focused on extending and improving the performance of existing frameworks. However, these frameworks have several limitations. Several studies addressed these issues but only separately, and they still can prevent these algorithms from being used for some real-world applications such as our motivating problem of finding co-locations of cancer cases and sets of released chemicals.

- **Prevalence measure thresholds**. A usage of thresholds for 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 cause it when they are combined.

- **Neighborhood distance threshold**. Most co-location mining approaches use one distance threshold to identify neighborhood 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 [33] deals with lines and polygons, it also uses a threshold for a prevalence measure. Furthermore, this framework cannot deal with uncertainty in datasets which is explained in the following paragraph.

- **Data uncertainty**. 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. Uncer-

tainty 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.

3.2 Algorithm Design

The objective of this work is to detect significant co-location patterns or rules in a given spatial dataset that have a prevalence measure value higher than an expected one. We propose a new framework that addresses the limitations mentioned in the previous section. A new grid-based transactionization method is deployed to transform a spatial dataset into a set of transactions. Transactions contain probabilities of feature existence and are used to compute a prevalence measure of patterns and rules. In this work, instead of having one threshold on a prevalence measure, we use the statistical test. It is proposed for co-location mining by Barua and Sander [7]. A pattern is considered significant, if a probability of seeing the same or greater value of a prevalence measure in R artificial datasets is less than α (the significance level) under a null hypothesis that there is no spatial dependency among features of the pattern. Each candidate pattern is evaluated separately rather than applying one threshold to all of them.

The design of the algorithm is presented in Figure 3.1. The algorithm includes three main parts.

1. The **initialization** step, in which buffer zones are built around instances of spatial features. Each buffer represents a region where a particular instance has an impact on other objects.

2. The **transactionization** step. A spatial dataset is transformed into a set of transactions deploying a grid-based method. The derived set of transactions is used to calculate prevalence, or interestingness, measure values of candidate co-location patterns or rules.

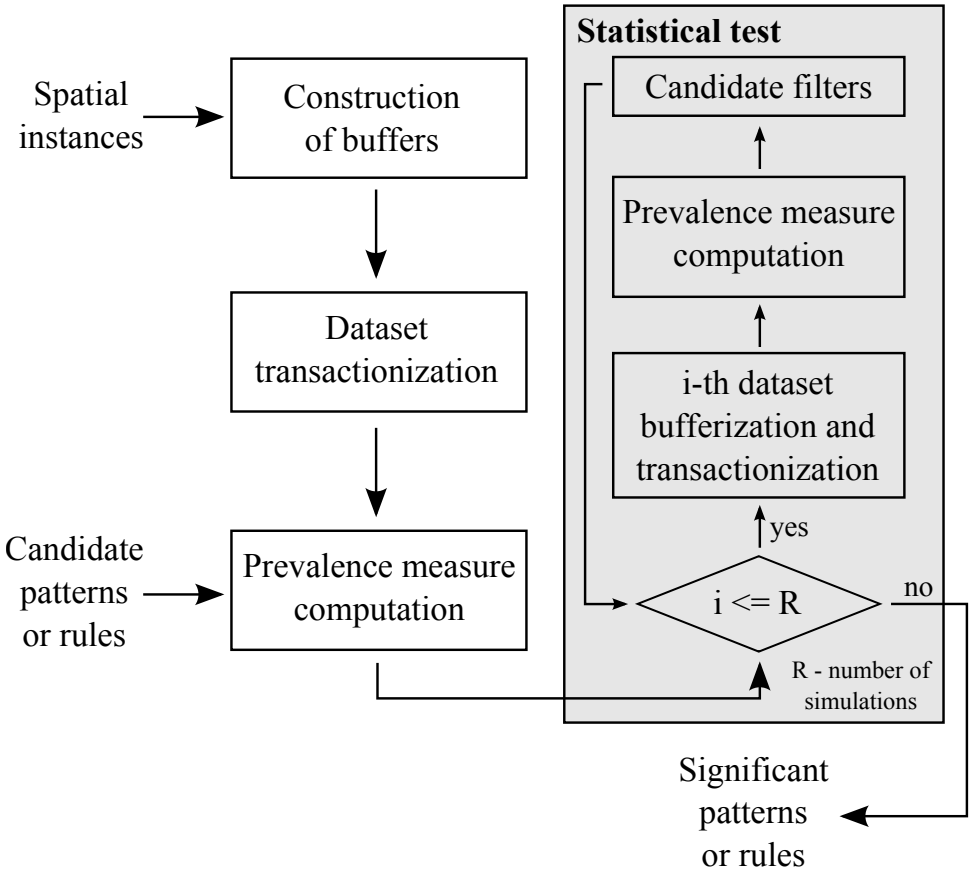


Figure 3.1: The algorithm design.

3. The **statistical test** is performed in order to get significant co-location rules or patterns. It includes both previous parts as subroutines when checking a set of randomized datasets. In addition, filters are used to prune candidates that are definitely not significant.

3.2.1 Initialization

In the initialization step, a buffer is built around each spatial object. It defines an area affected by that object; for example, a buffer zone around an emission point shows the area polluted by a released chemical. Buffers can be constructed using GIS systems. A buffer size, a distance from an object or its sides to buffer boundaries, can be chosen depending on various factors that may vary for different applications and projects. Possible cases for a buffer size choice include the following scenarios:

- One global buffer size for all spatial features and their instances. This approach might be used when there is no clear knowledge on an impact of various features on a region around them.

- Varying buffer size which depends on a type of a feature. Furthermore, a buffer size can be given as an attribute of an object instance. For example, emission points with varying amounts of released pollutants are assigned different buffer sizes.

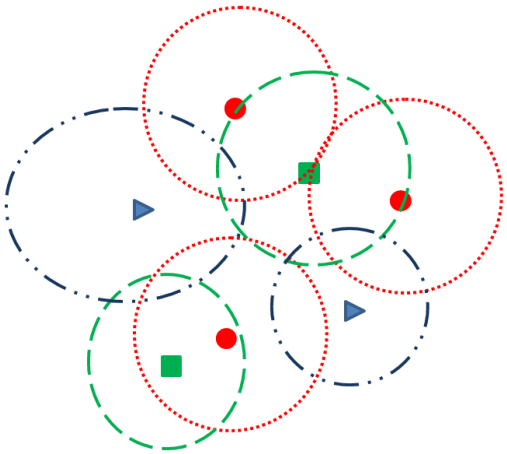
- A shape of a buffer may also change and it may depend on external factors. For instance, wind affects dispersion of pollutants and changes original circular buffers into other shapes depending on wind speed and its direction. In another example, topography features like high mountains or deep rivers impact distribution of animal species.

Figure 3.2(a) displays an example spatial dataset with buffers of various sizes that are formed around spatial point objects.

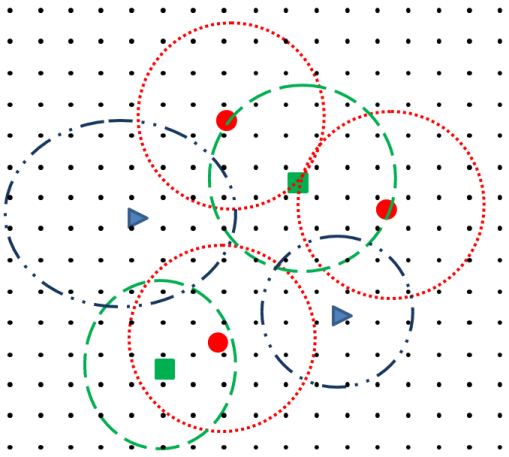
3.2.2 Transactionization

Recall that previous transaction-based methods have some limitations. A window-centric model cuts off neighborhood 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.

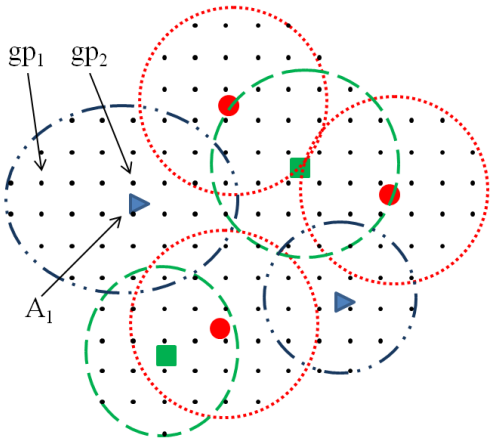
Instead of previous models we propose a new grid-based transactionization method. In order to transaction spatial data we use a grid which points are imposed over a given map. Figure 3.2(a) illustrates an example dataset with buffers around spatial point instances, and a grid is laid over it in Figure 3.2(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 shouldn't considerably affect accuracy of results.



(a) A sample spatial dataset with point feature instances and their buffers.



(b) A grid imposed over the space.



(c) Grid points which intersect with buffers are used to create transactions.

Figure 3.2: Transactionization step.

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 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 uni-

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 each** point $g \in G$ **do**
6: t = get a set of features which instances contain g with corresponding existential probabilities
7: $T = T \cup t$
8: **end for**
9: **return** T

form. Alternatively, since we do not have 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 chapter). For example, a grid point gp_2 in Figure 3.2(c) is located closer to a point A_1 than a point gp_1; we can assume that $p(A, gp_2) > p(A, gp_1)$. 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.

A granularity of the grid, or a distance between points of the 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.

3.2.3 Prevalence Measures

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. Various interestingness measures are proposed to define a significance of frequent patterns and association rules in certain and uncertain data. In this thesis we deploy two of them - the expected support and the expected confidence.

In some applications experts look for sets of features that are often co-located with each other and they do not need to discover cause-effect relationships. In this case, which is analogous to the frequent pattern mining problem, the expected support $ExpSup(P)$ might be used to define a level of interestingness of a pattern P, which is a subset of F (a set of features in a spatial dataset).

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. In the case of uncertain data, transactions are probabilistic and, therefore, the support is counted in expected value. The expected support was first defined by Chui et al. [13] for frequent pattern mining and is based on the possible worlds model. Briefly, for every feature f and every transaction t two possible worlds exist: one world where f is present in t and the second where f does not exist in t. The probability of the former being the true world is $p(f, t)$ and the probability of the latter world is $1 - p(f, t)$. In the case of several features, the probability of a single world is computed as the product of the probabilities of all features. Then, the expected support of a pattern is obtained by summing the support of the pattern in each of possible worlds multiplied by the probability of the world under an assumption that probabilities of features in the pattern are determined through independent observations. The number of possible worlds is 2^m where m is the number of feature instances in all transaction of an uncertain dataset. However, this compli-

cated formula can be reduced and the expected support of a pattern P is calculated as follows:

Definition 1 *The expected support $ExpSup(P)$ 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:*

$$ExpSup(P) = \sum_{t \in T} p(P, t). \tag{3.1}$$

The probability $p(P, t)$ of a presence of a pattern P in a transaction t depends on probabilities of features in P being present in t and 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:*

$$p(P, t) = \prod_{f \in P} p(f, t). \tag{3.2}$$

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.

For some applications, researchers intend to analyze co-location rules. For example, for a dataset of disease outbreaks and possible cause factors a typical rule is of the form $C \rightarrow D$, where C is a subset of cause features and D is a disease feature. This task is similar to the association rule mining problem. For these projects, 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 $X \cap Y = \emptyset$.

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

$$ExpConf(X \rightarrow Y) = \frac{ExpSup(X \cup Y)}{ExpSup(X)}. \tag{3.3}$$

In addition to the expected support and expected confidence, other interestingness measures could be applied to co-location mining with uncertain data. The measures used in frequent pattern and association rule mining include lift, conviction, cosine, etc. They can be used in a variety of domains depending on requirements of applications and projects.

Algorithm 2 Mining significant co-location patterns.

Input: Spatial dataset D.
Level of significance α.
Number of simulation runs R.
Set of randomized spatial datasets $RD[1..R]$.
Output: Set of significant co-location patterns P

T: set of transactions
CP: set of candidate patterns
$T = GetTransactions(D)$
for each $cp \in CP$ **do**
  $cp.ExpSup_{obs} = ComputeExpSup(cp, T)$
  if $cp.ExpSup_{obs} = 0$ **then**
    $CP = CP \backslash cp$
  end if
end for
for $i = 1 \rightarrow R$ **do**
  $T = GetTransactions(RD_i)$
  for each $cp \in CP$ **do**
    $cp.ExpSup_{sim}[i] = ComputeExpSup(cp, T)$
    if $cp.ExpSup_{sim}[i] \geq cp.ExpSup_{obs}$ **then**
      $cp.R_{\geq ExpSup_{obs}} = cp.R_{\geq ExpSup_{obs}} + 1$
      $cp.\alpha = \frac{cp.R_{\geq ExpSup_{obs}} + 1}{R+1}$
      if $cp.\alpha > \alpha$ **then**
        $CP = CP \backslash cp$
      end if
    end if
  end for
end for
$P = CP$
return P

3.2.4 Statistical Test

In the previous steps, a prevalence measure value is calculated for all candidate co-location patterns or rules. Now, the goal is to identify a set of significant patterns or rules. As discussed above, a usage of a threshold on a prevalence measure may result in discovery of wrong patterns and omission of interesting ones. Instead, only co-location patterns or rules that have surprising levels of a prevalence measure should be reported 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.

In our algorithm, we use statistical hypothesis testing to estimate significance of patterns and rules. A null hypothesis is that features of a pattern or rule are spatially independent from each other. If a likelihood, or probability p, of seeing the same level of the prevalence measure or greater under the null hypothesis is lower than the significance level α, the features are spatially co-located and the pattern or rule is considered significant. The following statement defines a significant co-location pattern.

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

The same logic can be applied to a case when co-location rules are mined. Therefore, a significant co-location rule is defined as follows:

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

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.

Let us suppose that the expected confidence $ExpConf$ is used as a prevalence measure. Let $ExpConf_{obs}(X \rightarrow Y)$ denote the expected confidence of a co-location rule $X \rightarrow Y$ in a real dataset, and $ExpConf_{rand}(X \rightarrow Y)$ - 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:

$$p = \frac{R_{\geq ExpConf_{obs}} + 1}{R + 1}, \tag{3.4}$$

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

If the p-value is less or equal to a predefined level of significance α, 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 α.

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 $ExpSup$ is used as a prevalence measure.

3.3 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 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 the p-value that have already exceeded the level α. For example, let us assume that the number of simulation runs is 99 and $\alpha = 0.05$. 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.

3.4 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

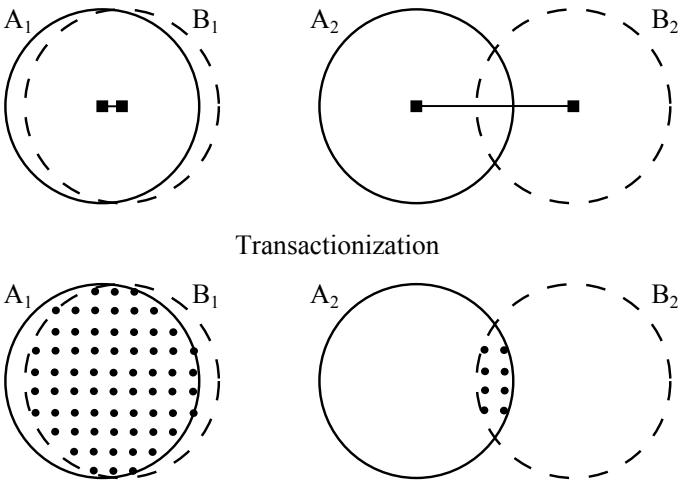
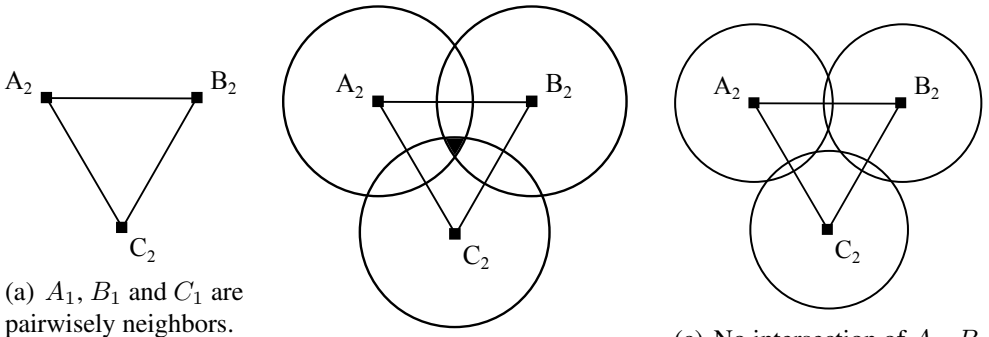


Figure 3.3: Neighboring objects $A_1 - B_1$ and $A_2 - B_2$. In the transactionization step, the intersection of A_1 and B_1 receives more transactions (black dots) than the pair A_2 and B_2.

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 neighborhood 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 in an 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 neighborhood 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 3.3 illustrates an example of two pairs of neighboring



(a) A_1, B_1 and C_1 are pairwisely neighbors.

(b) An intersection of A_1, B_1 and C_1.

(c) No intersection of A_1, B_1 and C_1.

Figure 3.4: Intersection of neighboring objects.

spatial objects with corresponding buffer zones. Both pairs, A_1-B_1 and $A2-B2$, 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 3.3) than the second pair of objects. Therefore, the real situation is represented more accurately. Consider another example. Let spatial points A_1, B_1 and C_1 be pairwisely neighbors (Figure3.4(a)). They are considered to form a clique by previous algorithms. However, as it can be seen in Figure3.4(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 three objects at all, although they form pairwise neighborhood relationships, as it is illustrated in Figure3.4(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.

Chapter 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, a usage of general approaches and algorithms may result in misleading or even wrong results. For example, a neighborhood 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.

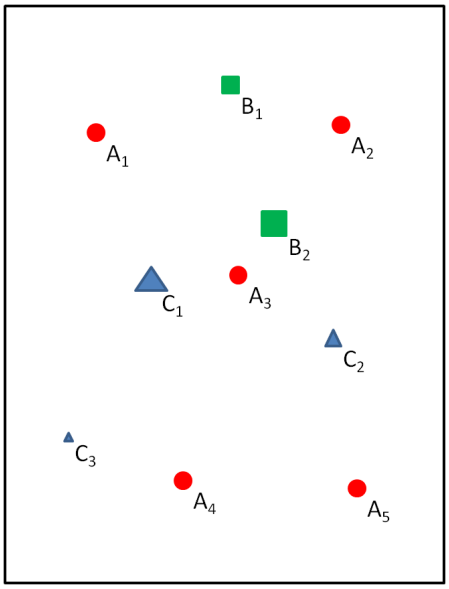
The motivating task of this thesis, detecting co-locations of pollutant emission points and childhood cancer cases, has unique difficulties and challenges. A distribution of a pollutant in a region is not uniform and it could depend on several factors: types of pollutants, amounts of release, climatic conditions (wind, precipitation), topography, etc. Various chemicals have different levels of harmfulness and toxicity. In addition, a pollutant concentration is inversely proportional to a distance

from an emitting point. These are only several 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 many variables and parameters. Instead, our model gives a simple framework that attempts to simulate real world conditions while operating with available data.

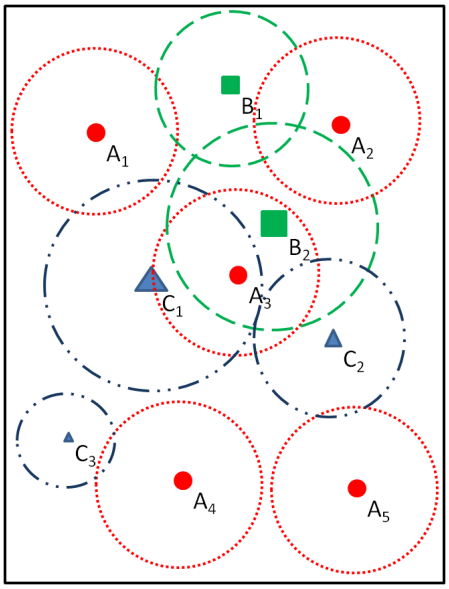
4.1 Pollutant Amounts

The dataset on pollutants contains data on yearly releases of chemicals. For our research we take an average amount of release for a year on given facilities and chemicals, which is further normalized by Toxic Equivalency Potentials (TEPs) when they are available. TEP shows the relative risk associated with one kilogram of a chemical in comparison with the risk caused by one kilogram of benzene. Chemicals with high TEPs are extremely toxic. A range of average amount values varies from several kilograms to tens of thousands tons; the maximal average yearly release in the dataset is 80,000 tons. Certainly, one distance threshold for all pollutant emissions is inaccurate, because the more amount of a chemical is released, the farther it distributes from a source point. Figure 4.1(a) displays an example dataset containing cancer points (feature A) and chemical points (features B and C). In Figure 4.1(b), the buffer zones around pollutant points are based on an 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 are not changed.

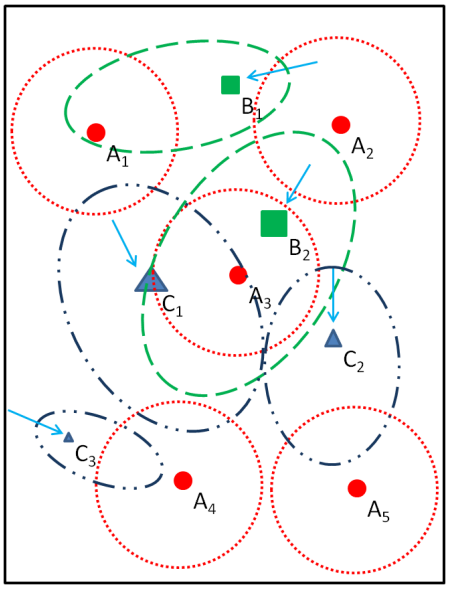
As a function of a dependency of a buffer size on a release amount we use the natural logarithm function. This function gives a smooth curve which does not grow as fast as linear or root functions that give large numbers for heavier releases. Even though this technique oversimplifies the real-world conditions of pollutant dispersion, it helps to make results more precise than when using one buffer size for different amounts of chemicals. Other functions can be used to calculate the maximal distribution distance and they can depend on a type of a pollutant (a heav-



(a) A sample spatial dataset (A - cancer, B and C - pollutants).



(b) Buffer sizes vary depending on the pollutant release amount.



(c) Buffer shapes change with the wind direction and speed (shown by arrows).

Figure 4.1: Modeling framework usage examples.

ier 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 an area very close to a chimney does not get polluted, and the higher the chimney, the bigger the region.

4.2 Wind Speed and Direction

The climatic conditions and topographical features may affect distribution of chemicals in 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 at 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 4.1(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 [19], and Reggente and Lilienthal [28]. 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 4.2 shows an example of buffer transformation. The original circle buffer zone around the emission point P is changed to an ellipse.

Obviously, wind with a higher speed distributes chemicals to greater distances. Therefore, we need to include the wind speed value in 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:

$$a = r + \gamma|\vec{v}|, \tag{4.1}$$

$$b = \frac{r^2}{a}, \tag{4.2}$$

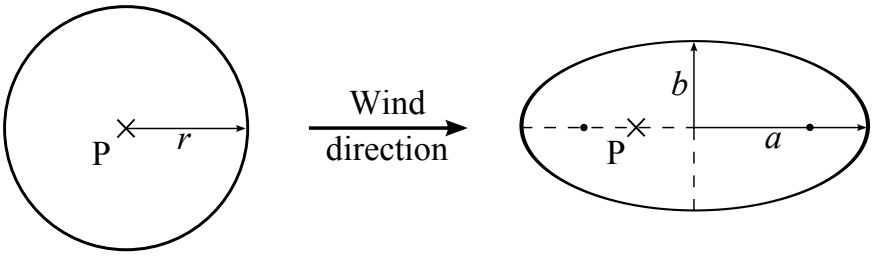


Figure 4.2: A buffer circle around emission point P is morphed into an ellipse.

where r - the radius of the original circle, $\vec{v}$ - the wind speed, and γ - 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.

4.2.1 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 [15] 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 are displayed in Figure 4.3.

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

(a) 18 stations of Environment Canada. (b) 156 stations of ACIS.

Figure 4.3: The monitoring stations in Alberta.

calculated using a method similar to the one used by Environment Canada [15]. 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.

The climate normals from two sources are combined and used to make interpolations in 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 path when moving from one number to another. 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 [31]. The wind speed and wind direction from each monitoring station is represented as a vector with the magnitude S (wind speed) and direction θ (wind direction). The vector is divided into axial components X (northern wind) and Y (eastern wind):

$$X = S \sin \theta, \tag{4.3}$$

$$Y = S \cos \theta. \tag{4.4}$$

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 speed S', is computed as:

$$S' = \sqrt{X'^2 + Y'^2}. \tag{4.5}$$

The calculation of wind direction angle θ' 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° and 90° and it is only half of our domain. Therefore, each of the four quadrants of our domain (the quadrants are shown in Figure 4.4) requires its own formula [31]:

$$Quad\ I : \theta' = \tan^{-1}(X'/Y'), \tag{4.6}$$

$$Quad\ II : \theta' = \tan^{-1}(Y'/X') + 90°, \tag{4.7}$$

$$Quad\ III : \theta' = \tan^{-1}(X'/Y') + 180°, \tag{4.8}$$

$$Quad\ IV : \theta' = \tan^{-1}(Y'/X') + 270°. \tag{4.9}$$

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

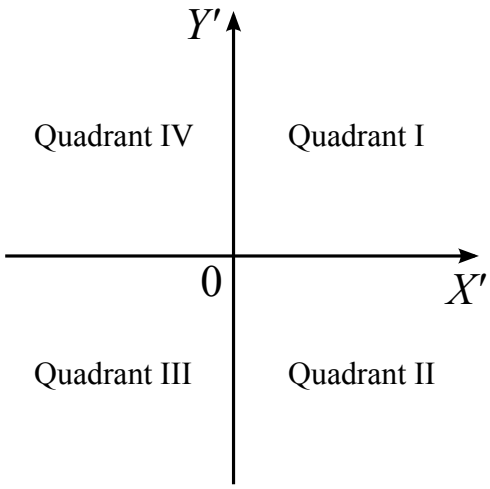


Figure 4.4: Four quadrants defined by the signs of values of X' (northern wind) and Y' (eastern wind).

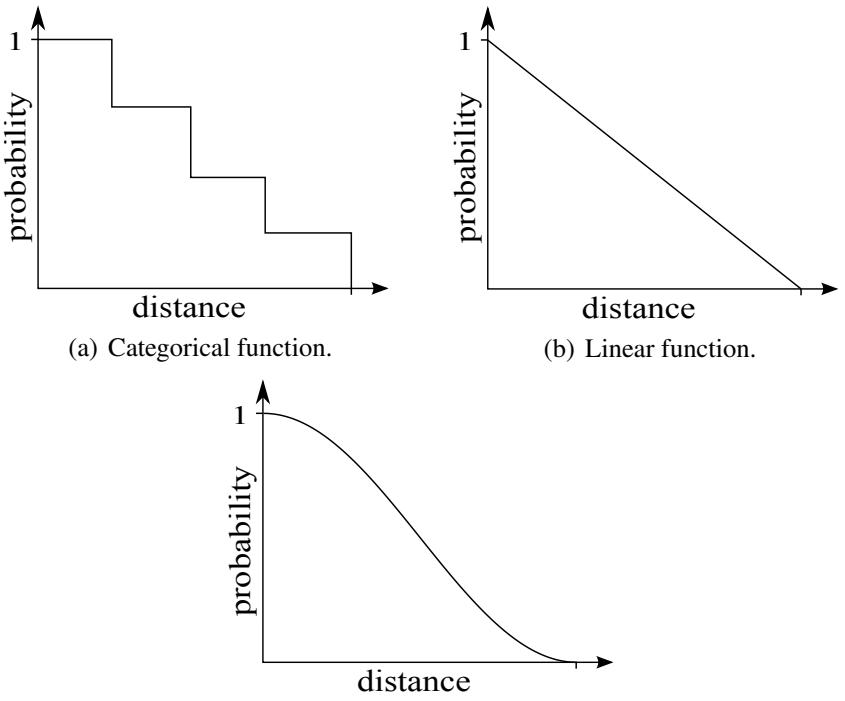
4.3 Data Uncertainty

A dispersion of a pollutant in a distribution region is not uniform. Intuitively, A concentration of the pollutant near a chimney is higher than at a border of the dispersion region. Furthermore, pollutants are a 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 3.2(c) the probability of detecting A at the point gp_1 is lower than at the point gp_2.

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.

- When using a categorical function (Figure 4.5(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 4.5(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.



(a) Categorical function.

(b) Linear function.

(c) Curve function ($\frac{\cos \pi x}{2} + 0.5$).

Figure 4.5: Examples of functions that can be used to represent the dependency of the pollutant presence probability on the distance to the source point.

- In this work we use a third example, the curve function (Figure 4.5(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 declining 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 results. They could depend on types of chemicals. For example, most of a heavy chemical may settle out in a region closer to the emitting facility, while only small amounts reach places at medium and far distances.

Some datasets in addition to point features may contain other types of spatial objects, i.e. lines and polygons. For these datasets, uncertainty may be modeled as follows. Grid points intersecting a line or located inside a polygon are assigned

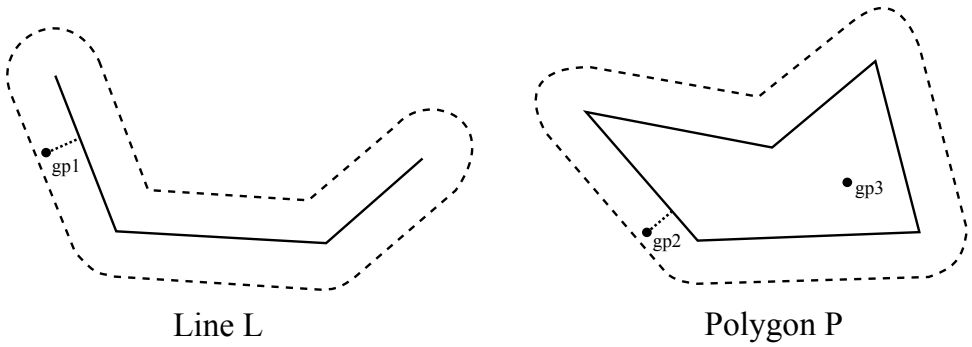


Figure 4.6: Defining a distance to an object in datasets with polygons and lines.

a feature presence probability of one. For example, point gp_3 in Figure 4.6. 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 4.6 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.

Chapter 5

Experimental Evaluation

In this chapter, we present the results of experiments conducted on a real dataset containing pollutant emission points and childhood cancer cases to evaluate the proposed algorithm. In addition, the algorithm is applied and evaluated on synthetic data.

5.1 Real Data

We conduct experiments on a real dataset which contains data on pollutant emission points and childhood cancer cases in the province of Alberta, Canada. The sources of the databases are the National Pollutant Release Inventory (NPRI, the data is publicly available) [9], and the provincial and national childhood cancer registries. 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; 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) is 1,254. The model that is used to define buffer zones and uncertainty in the data is explained in the previous chapter.

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 subject for careful evaluation by domain experts in our multidisciplinary team.

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 (3.4) is 100. The level of significance α 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 1.95.

5.1.1 Randomized Datasets

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. Out of 1,254 cancer points, 1,134 are located within dense "urban" municipalities (cities, towns, villages, etc.) and the rest are diagnosed 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 [10]. The rest 120 cases, which are located in rural regions in the real dataset, are randomly placed on the map of Alberta but not in urban areas.

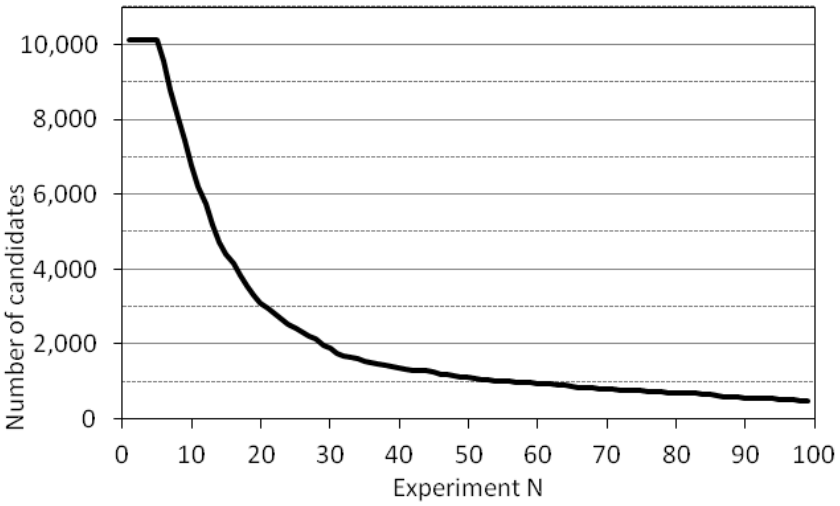


Figure 5.1: The number of candidate rules evaluated in each simulation run with the filtering technique.

5.1.2 Effect of Filtering Techniques

The number of candidate co-location rules in the experiment is 17,343 (co-locations with the antecedent size up to three). With the naive approach all candidates are checked in each of 99 simulation runs which results in a large amount of computation. In order to prune insignificant candidates, we propose two filtering techniques. After the exclusion of rules with zero-level confidence 10,125 candidates remain which also form a big set. Figure 5.1 shows that the usage of the second filtering method (the exclusion of candidates which p-values 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 20-th simulation, and only 488 candidates are evaluated in the last run.

5.1.3 Comparison with the Certain Data Method

In this experiment 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 $Conf(Pol \rightarrow Cancer)$, which is a

fraction of transactions containing all features in Pol that also include the cancer feature.

$$Conf(Pol \rightarrow Cancer) = \frac{Sup(Pol \cup Cancer)}{Sup(Pol)}. \tag{5.1}$$

The number of significant co-location rules detected by either UM or CM is 496. From these 204 rules are found by both methods, 278 rules are identified only by UM, and 14 - by CM. 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 $ExpSup(Pol)$ or $ExpSup(Pol \cup Cancer)$. For example, 348 out of 482 rules have $ExpSup(Pol \cup Cancer)$ 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 (≤ 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.

In order to show that the uncertain data method deals with spatial information better than the certain data method we conduct an experiment with co-location patterns instead of rules. Recall that the expected confidence is computed as a division of $Sup(Pol \cup Cancer)$ by $Sup(Pol)$. Therefore, a pattern may be reported by either UM or CM not only because of varying distances from grid points to instance objects in a real and randomized datasets. The support values ($Sup(Pol \cup Cancer)$ and $Sup(Pol)$) could also vary considerably. It is possible that $ExpSup_{obs}(Pol \cup Cancer) \leq ExpSup_{rand}(Pol \cup Cancer)$ and $ExpConf_{obs}(Pol \cup Cancer) > ExpConf_{rand}(Pol \cup Cancer)$ at the same time, and vice versa. When mining

co-location patterns, $ExpSup_{obs}$ of a significant pattern is higher than $ExpSup_{rand}$ in at least 95% of simulations, and it is feasible to find out what are the causes of the difference in the results by two methods.

Seven patterns are detected only by the certain data method, and fourteen - only by UM. In order to analyze these results, the average probability $p(P,t)_{obs}$ of seeing all features of a co-location pattern P (pollutants and cancer) in transactions of the real dataset is compared with the average probability $p(P,t)_{rand}$ in randomized datasets. For the patterns that are detected only by CM, we only take the randomized datasets which caused the difference in p-values in CM and UM, i.e., in these datasets $Sup_{obs} > Sup_{rand}$ and $ExpSup_{obs} \leq ExpSup_{rand}$, and vice versa for the patterns discovered only by UM.

The results are shown in Table 5.1; for simplicity codes are given instead of pollutant names. As can be seen, all seven co-location patterns discovered by CM have $p_{obs} < avg\ p_{rand}$. It means that in simulated datasets emission points and cancer cases are located closer to grid points than in the real dataset, which causes $ExpSup_{rand}$ in simulated data to be higher than $ExpSup_{obs}$. This is the reason why these patterns are not detected by the uncertain method. Moreover, fourteen patterns that are found only by UM have $p_{obs} > avg\ p_{rand}$, which means that in randomized datasets spatial objects are located on average farther from grid points than in the real dataset. The features in these patterns are more likely to be associated with each other. Looking at the difference in the results of CM and UM, we can conclude that the uncertain data method in addition to neighborhood relationships between spatial objects also takes into account spatial information and relative locations of objects and grid points.

5.1.4 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 accuracy of results. 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

Table 5.1: Co-location patterns detected by either CM or UM.

Method	Co-location Pattern	Avg $p(P,t)_{obs}$	Avg $p(P,t)_{rand}$
CM	DQ, GZ, Cancer	17.06%	**22.95%**
CM	GF, GZ, Cancer	18.51%	**22.37%**
CM	GF, HG, Cancer	16.80%	**20.26%**
CM	BJ, CG, DP, Cancer	3.10%	**8.30%**
CM	BJ, DV, GF, Cancer	10.17%	**14.65%**
CM	BJ, GF, HG, Cancer	13.66%	**16.59%**
CM	DV, GF, GZ, Cancer	12.80%	**16.70%**
UM	FF, GB, Cancer	**14.78%**	1.32%
UM	BB, FF, GB, Cancer	**14.64%**	1.26%
UM	BF, FF, GB, Cancer	**4.99%**	1.16%
UM	DJ, EK, GB, Cancer	**10.08%**	7.04%
UM	DV, FF, GB, Cancer	**14.66%**	1.26%
UM	DZ, FF, GA, Cancer	**10.44%**	1.99%
UM	EB, FB, FF, Cancer	**8.66%**	1.42%
UM	EN, FB, FF, Cancer	**4.23%**	1.19%
UM	EN, FF, GA, Cancer	**3.18%**	1.03%
UM	FB, FF, FY, Cancer	**10.50%**	1.61%
UM	FB, FF, GD, Cancer	**14.24%**	1.64%
UM	FF, GB, GZ, Cancer	**14.74%**	1.30%
UM	FF, GB, HD, Cancer	**14.69%**	1.17%
UM	FF, GB, HG, Cancer	**14.69%**	1.20%

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.

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 granularity grid reported 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 grid. As we can see, the difference between 0.5 and 1 km result sets is smaller than between 1 km and 2 km grids. As the distance between points in a grid decreases, the accuracy of results improves.

5.2 Synthetic Data

We conduct experiments on synthetic datasets to demonstrate that our framework can discover 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 ("cause" features C), and a disease feature D. The study region is a 100x100 unit square. The buffer size is 1 unit. 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 ≤ 0.05 are shown in Table 5.2.

Table 5.2: 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

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 the respective features ($C_1 - C_4$) have associations with D either total ($C_1 - C_2$) or partial ($C_3 - C_4$). 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 the feature in the consequent part. A threshold with a relatively low value can help to exclude rules with noise features.

Table 5.3: 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

5.2.2 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 3.3). 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 neighborhood relationships. Most previous approaches assign the same prevalence measure value in both cases as long as a neighborhood 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 100x100 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), ..., [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 5.3. 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, out transaction-based approach takes into account the actual spatial information and a relative proximity or remoteness of features from each other.

Chapter 6

Conclusion

6.1 Summary

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 a new knowledge in various domains. In this thesis we have proposed a new solution to the co-location mining problem. The approach was motivated by a real-world application of detecting possible associations of pollutant emission points and childhood cancer cases in the province of Alberta, Canada. We explained

A short introduction to spatial data mining and some of its tasks and algorithms was presented in this work. We reviewed various existing approaches to the co-location mining problem, discussed their designs and parameters. We also briefly explained some of the approaches to frequent pattern and association rule mining problems with certain and uncertain data.

We addressed our thesis statements by proposing a new framework which combines classical co-location mining, and uncertain frequent pattern and association rule mining. We took into account some of the limitations that can prevent previously proposed approaches from being used in some real-world applications and domains. A 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 ac-

count uncertainty of data by storing feature existence probabilities in transactions. 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 real and synthetic datasets showed that our approach finds significant co-location patterns and rules. The effect of a grid granularity is evaluated. A dependence of a prevalence measure value on an average distance between feature instances was 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 demonstrated 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 thesis has its unique challenges. We examined 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 modeled uncertainty of a pollutant presence at transaction points. In addition to pollution, other factors can also cause cancer in children. In this thesis we did 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.

Projects and applications from various domains have their unique characteristics and requirements. A basic framework should be deployed in each application with a careful consideration of its specific challenges. One interestingness measure

can be suitable for some applications and lead to wrong findings in others. Each project often needs an individual approach with a careful choice of parameters and processing steps. For instance, stages of data preprocessing and cleaning are important parts of analysis in many applications and, therefore, require a close attention of researchers.

6.2 Contributions

The main contributions of this thesis are as follows:

1. A new framework, which is based on the statistical test, was proposed to tackle the co-location mining problem. We showed that it can find significant co-location patterns and rules by calculating and comparing their prevalence measures (the expected support and expected confidence) in a real and randomized datasets. Our evaluation showed that it can find a correct set of co-location rules.

2. A novel method of creating a set of transactions from a spatial dataset was devised. An uncertainty of a feature presence in transactions was included in our model. The approach based on this transactionization and uncertainty modeling method deals with spatial information better than most previous approaches.

3. A problem of a high computation cost was addressed by using two pruning techniques. They considerably decrease computation time by pruning insignificant candidates and reducing the number of candidate co-location patterns and rules that need to be checked.

4. Some of the factors that affect distribution of pollutants in air were taken into account in the modeling framework. These factors are pollutant release amounts, wind data (average speed and prevailing direction) and uncertainty of a pollutant presence at transaction points.

6.3 Directions for Future Research

In this thesis we proposed a novel algorithm to tackle the problem of co-location mining in spatial data. However, there are still many open challenges to be resolved. The research in the following directions may further improve accuracy of results and widen applicability of co-location mining approaches.

As we discussed in Chapter 3, our approach can be extended to datasets containing not only point instances but also linear and polygonal objects such as roads, recreation zones and parks in cities. The inclusion of these types of spatial objects makes analysis of more spatial datasets possible which can further expand a range of applications of co-location mining. However, addition of complex objects would require more computations in a process of creating buffers and transactioning spatial data. For example, GIS buffer, overlay, and intersection tools are computationally expensive. Therefore, more research is needed to efficiently use spatial libraries and GIS technologies.

We explained and proposed to use two measures of prevalence - expected support and expected confidence - that are analogous to the respective notions of support and confidence in frequent pattern and association rule mining. In addition to these parameters, other interestingness measures such as lift, cosine, leverage, all-confidence, conviction, etc., were defined in previous research for various purposes and tasks of deterministic frequent pattern mining. Each of them is suitable in different scenarios and cases. When adapted to uncertain data mining, these measures can help to adjust the co-location mining task to unique requirements of various projects.

By deploying the statistical test to analyze a set of candidate patterns, we ensure that all significant co-location patterns and rules are discovered by the algorithm. However, with an increase in the number of possible candidates and the number of simulation runs, computations can become prohibitively expensive. The usage of a finer grid with shorter distances between its points increases the number of transactions and also requires more processing time. We use two filtering techniques in order to prune candidates that are definitely not significant. Further research to

define more pruning methods can substantially decrease the number of unneeded checks of false candidates.

In Chapter 4, we explained the modeling framework which was used in the motivating application of this thesis. We believe that by considering pollutant release amounts, average wind speeds and prevailing wind directions, and introducing uncertainty to the model, we increased accuracy of the results in comparison with a case when none of these factors are taken into account. We understand that this model simplifies the real-world situation and more work should be done to further enhance the modeling framework. For instance, different pollutant dispersion and decay models and functions could be deployed.

Bibliography

[1] ArcGIS Desktop: Release 10, ESRI 2011.

[2] AgroClimatic Information Service (ACIS). Live alberta weather station data. `http://www.agric.gov.ab.ca/app116/stationview.jsp`.

[3] Charu C. Aggarwal, Yan Li, Jianyong Wang, and Jing Wang. Frequent pattern mining with uncertain data. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 29–38, New York, NY, USA, 2009. ACM.

[4] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22(2):207–216, June 1993.

[5] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[6] G. Al-Naymat. Enumeration of maximal clique for mining spatial co-location patterns. In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pages 126 –133, 31 2008-april 4 2008.

[7] Sajib Barua and Jörg Sander. SSCP: mining statistically significant co-location patterns. In *Proc. of the 12th international conference on Advances in spatial and temporal databases*, SSTD'11, pages 2–20, Berlin, Heidelberg, 2011. Springer-Verlag.

[8] Thomas Bernecker, Hans-Peter Kriegel, Matthias Renz, Florian Verhein, and Andreas Zuefle. Probabilistic frequent itemset mining in uncertain databases. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 119–128, New York, NY, USA, 2009. ACM.

[9] Environment Canada. National Pollutant Release Inventory. Tracking Pollution in Canada. `http://www.ec.gc.ca/inrp-npri/`.

[10] Statistics Canada. 2006 Census. `http://www12.statcan.gc.ca/census-recensement/2006/index-eng.cfm`.

[11] Yue-Hong Chou. *Exploring spatial analysis in geographic information systems*. OnWord Press, 1997.

[12] Chun-Kit Chui and Ben Kao. A decremental approach for mining frequent itemsets from uncertain data. In Takashi Washio, Einoshin Suzuki, Kai Ting, and Akihiro Inokuchi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 5012 of *Lecture Notes in Computer Science*, pages 64–75. Springer Berlin / Heidelberg, 2008.

[13] Chun-Kit Chui, Ben Kao, and Edward Hung. Mining frequent itemsets from uncertain data. In Zhi-Hua Zhou, Hang Li, and Qiang Yang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 4426 of *Lecture Notes in Computer Science*, pages 47–58. Springer Berlin / Heidelberg, 2007.

[14] Noel A. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1991.

[15] Environment Canada. National Climate Data and Information Archive. Canadian climate normals or averages 1971-2000. `http://climate.weatheroffice.gc.ca/climate_normals/index_e.html`.

[16] Martin Ester, Hans-Peter Kriegel, and Jörg Sander. Algorithms and applications for spatial data mining. In Harvey J. Miller and Jiarwei Han, editors, *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*, pages 160–187. Taylor and Francis, 2001.

[17] Vladimir Estivill-Castro and Ickjai Lee. Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In *Proc. of the 6th International Conference on Geocomputation*, 2001.

[18] Vladimir Estivill-Castrol and Alan Murray. Discovering associations in spatial data an efficient medoid based approach. In Xindong Wu, Ramamohanarao Kotagiri, and Kevin Korb, editors, *Research and Development in Knowledge Discovery and Data Mining*, volume 1394 of *Lecture Notes in Computer Science*, pages 110–121. Springer Berlin / Heidelberg, 1998.

[19] Arthur Getis and P. H. Jackson. The expected proportion of a region polluted, by k sources. *Geographical Analysis*, 3(3):256–261, 1971.

[20] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. ACM.

[21] Yan Huang, Jian Pei, and Hui Xiong. Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, 10(3):239–260, September 2006.

[22] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering colocation patterns from spatial data sets: A general approach. *IEEE Trans. on Knowl. and Data Eng.*, 16(12):1472–1485, December 2004.

[23] Yan Huang and Pusheng Zhang. On the relationships between clustering and spatial co-location pattern mining. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 513–522, Washington, DC, USA, 2006. IEEE Computer Society.

[24] Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In *Proc. of the 4th International Symposium on Advances in Spatial Databases*, SSD '95, pages 47–66, London, UK, UK, 1995. Springer-Verlag.

[25] Seung Kwan Kim, Younghee Kim, and Ungmo Kim. Maximal cliques generating algorithm for spatial co-location pattern mining. In James J. Park, Javier Lopez, Sang-Soo Yeo, Taeshik Shon, and David Taniar, editors, *Secure and Trust Computing, Data Management and Applications*, volume 186 of *Communications in Computer and Information Science*, pages 241–250. Springer Berlin Heidelberg, 2011.

[26] Yasuhiko Morimoto. Mining frequent neighboring class sets in spatial databases. In *Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 353–358, New York, NY, USA, 2001. ACM.

[27] Jian Pei, Jiawei Han, Hongjun Lu, Shojiro Nishio, Shiwei Tang, and Dongqing Yang. H-mine: Hyper-structure mining of frequent patterns in large databases. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 441 –448, 2001.

[28] Matteo Reggente and Achim J. Lilienthal. Using local wind information for gas distribution mapping in outdoor environments with a mobile robot. In *Sensors, 2009 IEEE*, pages 1715 –1720, Oct. 2009.

[29] Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns: A summary of results. In *Proc. of the 7th International Symposium on Advances in Spatial and Temporal Databases*, SSTD '01, pages 236–256, London, UK, UK, 2001. Springer-Verlag.

[30] Shashi Shekhar, Pusheng Zhang, and Yan Huang. Spatial data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 837–854. 2010.

[31] Rochelle Griffin Williams. Nonlinear surface interpolations: Which way is the wind blowing? In *Proc. of 1999 Esri International User Conference*, 1999.

[32] Xiangye Xiao, Xing Xie, Qiong Luo, and Wei-Ying Ma. Density based co-location pattern discovery. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, GIS '08, pages 29:1–29:10, New York, NY, USA, 2008. ACM.

[33] Hui Xiong, Shashi Shekhar, Yan Huang, Vipin Kumar, Xiaobin Ma, and Jin Soung Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *Proc. of the 2004 SIAM international conference on data mining*, 2004.

[34] Jin Yoo and Mark Bow. Mining maximal co-located event sets. In Joshua Huang, Longbing Cao, and Jaideep Srivastava, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6634 of *Lecture Notes in Computer Science*, pages 351–362. Springer Berlin / Heidelberg, 2011.

[35] Jin Soung Yoo and M. Bow. Mining top-k closed co-location patterns. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on*, pages 100 –105, 29 2011-july 1 2011.

[36] Jin Soung Yoo and Shashi Shekhar. A joinless approach for mining spatial colocation patterns. *IEEE Trans. on Knowl. and Data Eng.*, 18(10):1323–1337, October 2006.

[37] Jin Soung Yoo, Shashi Shekhar, John Smith, and Julius P. Kumquat. A partial join approach for mining co-location patterns. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, GIS '04, pages 241–249, New York, NY, USA, 2004. ACM.

[38] Mohammed J. Zaki. Scalable algorithms for association mining. *IEEE Trans. on Knowl. and Data Eng.*, 12(3):372–390, May 2000.

[39] Xin Zhang, Nikos Mamoulis, David W. Cheung, and Yutao Shou. Fast mining of spatial collocations. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 384–393, New York, NY, USA, 2004. ACM.