Discovering Co-Location Patterns in Datasets with Extended Spatial Objects

Aibek Adilmagambetov, Osmar R. Zaiane, and Alvaro Osornio-Vargas

Department of Computing Science and Department of Paediatrics
University of Alberta, Edmonton, Canada
{adilmaga,zaiane,osornio}@ualberta.ca

Abstract. Co-location mining is one of the tasks of spatial data mining, which focuses on the detection of the sets of spatial features frequently located in close proximity of each other. Previous work is based on transaction-free apriori-like algorithms. The approach we propose is based on a grid transactionization of geographic space and designed to mine datasets with extended spatial objects. A statistical test is used instead of global thresholds to detect significant co-location patterns.

1 Introduction

Co-location mining aims to discover patterns of spatial features often located close to each other in geographic proximity. An example is a co-location of symbiotic species of plants and animals depending on ecological conditions. The main purpose of co-location mining is to come up with a set of hypotheses based on data features and statistics that can be useful for domain experts to reduce the range of possible patterns that are hidden and need to be checked. Even though this task seems to be similar to association rule mining (ARM), the adaptation of ARM techniques is not trivial due to the fact that features are embedded into a geographic space and there is no clear notion of transactions.

Most of the existing approaches to the co-location mining problem [1–4] 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.

We propose a new framework which combines co-location mining, frequent pattern and association rule mining. A statistical test is used to determine the significance of co-location patterns and rules. A co-location is considered as significant if it has a surprisingly high level of prevalence in comparison with randomized datasets which are built under the null hypothesis that the features are independent from each other. We improve computation with filtering techniques.

The motivating application of this paper is the detection of possible spatial associations of chemicals and cases of childhood cancer. Although some people are genetically predisposed to cancer, most of the cases of cancer are caused by environmental factors, such as air pollutants, radiation, infections, tobacco, and alcohol. However, the 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. A collaborative research effort with the Faculty of Medicine is trying to identify associations between cancer cases and known emissions by industry. Some chemicals are proven to be carcinogens while others are not known to cause cancer in isolation. It is unknown if certain combinations of chemicals can be associated with higher rates of cancer. Moreover, even if potentially problematic combinations are not emitted by the same industry, atmospheric conditions can contribute to the mixture. We deploy our model on the dataset containing information on chemical emission points and amounts of release in Alberta, Canada, and childhood cancer cases with their location when they were first diagnosed. Our data is obtained from the National Pollutant Release Inventory (NPRI), Canada's legislated, publicly accessible inventory of pollutant releases, as well as the health authorities in Alberta for 1254 cancer cases of children younger than 19 between 2002 and 2007. NPRI for the province of Alberta provided 1465 points releasing a variety of chemicals among 47 of interest, some carcinogenic and some not classifiable as to carcinogenicity. In this paper we explain a modeling framework which is used to handle the data as accurately as possible. 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 investigate further the relationship between cancer and specific combinations of chemicals.

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

2 Related Work

2.1 Co-Location Mining

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

Spatial Statistics Approaches use statistical techniques such as cross K-functions with Monte-Carlo simulations [5], mean nearest-neighbor distance, and spatial regression models [6]. The disadvantages of these approaches are the expensive computation time and the difficulty of application to patterns with more than two spatial features.

Spatial Data Mining Approaches could be categorized into several types. Transaction-based approaches work by creating transactions over space and using association rules [7–9]. One of the ways, a reference-centric model, creates

transactions around a reference feature. However, this approach may consider the same instance set several times. Another approach, a window-centric model, divides the space into cells and considers instances in each cell as a transaction which causes a problem of some instance sets being divided by cell boundaries.

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 a map layer is constructed for each spatial feature based on instance clusters or boundaries of clusters [10]. The authors propose two algorithms for cluster association rule mining, vertical-view and horizontal-view approaches. In the former, clusters for layers are formed and layers are segmented into a finite number of cells. Then, a relational table is constructed where the element is equal to one if the corresponding cell satisfies the event in a layer, and zero otherwise. The association rule mining is applied to the table. The second approach uses 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 and $CC\%$ is the clustered confidence.However, these approaches might be sensitive to the choice of clustering methods, and assume that features are explicitly clustered.

Another type of spatial join-based methods - instance-join algorithms - is similar to classical association rule mining. Shekhar and Huang [1] proposed a co-location pattern mining framework which is based on neighborhood relations and the 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 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. A co-location pattern is considered prevalent, or interesting, if for each feature of the pattern at least $PI\%$ instances of that feature form a clique with the instances of all other features of the pattern according to the neighborhood relationship. Similarly to association rule mining, only frequent $(k-1)$-patterns are used for the k-candidate generation.

The approaches mentioned above use thresholds for measures of interestingness, which causes meaningless patterns to be considered as significant with a low threshold, and a high threshold may prune interesting rare patterns.

3 Algorithm

Various approaches to the co-location mining problem have been proposed during the past decade. However, most of them focused on improving the performance of existing frameworks which have several disadvantages. Several studies addressed these issues but only separately, and these issues remain major hurdles for some real-world applications such as our motivating problem of finding co-locations of cancer cases and pollutant emission points.

First, the usage of thresholds for the detection of interesting co-location patterns and rules is the main limitation factor of many co-location mining algorithms. In spatial datasets the features usually have a varying number of instances; they could be extremely rare or be present in abundance. Therefore, one threshold for 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). In this work we use the statistical test which replaces one global threshold. It is proposed for co-location mining by Barua and Sander [11]. The pattern is considered significant, if the probability of seeing the same or greater value of the prevalence measure in N artificial datasets is less than α (the significance level) under the null hypothesis that there is no spatial dependency among features of the pattern. Each candidate pattern is evaluated separately rather than applying one threshold for all of them.

Second, most co-location mining approaches are designed for spatial datasets with point features. However, other types of objects may exist in spatial data such as lines (roads) and polygons (polluted regions). Even though the framework for extended objects [4] deals with lines and polygons, it also uses one threshold for the prevalence measure. If the statistical test is applied to this model, computationally expensive GIS overlay methods should be used for each candidate pattern in order to calculate its prevalence measure in a real and randomized datasets. When the number of patterns and simulation runs in the statistical test are large, this method could become prohibitively expensive.

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

3.1 Algorithm Design

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

We propose a new transaction-based approach that is suitable for extended spatial objects. Previous transaction-based methods have some limitations. A

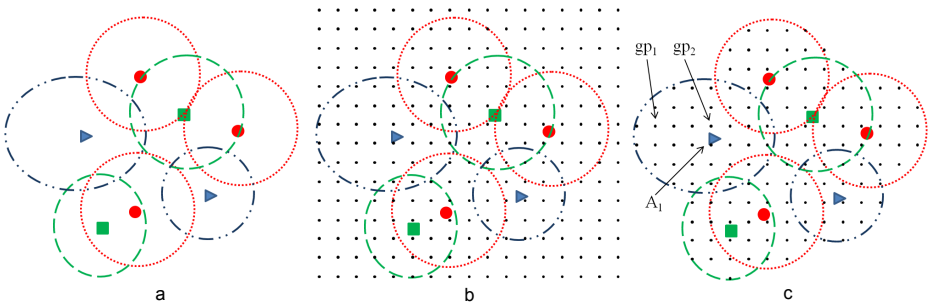

Fig. 1. Transactionization step: (a) an example spatial dataset with point feature instances and their buffers; (b) a grid imposed over the space; (c) grid points which intersect with buffers are used to create transactions.

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 these models we propose a new transactionization method. In order to transform spatial data into transactions, we use a grid which points are imposed over the given map. Fig. 1 (a) displays an example dataset with buffers around spatial point instances, and a grid is laid over it (Fig. 1 (b)). Similarly, buffers can also be created around linear and polygonal spatial objects. In a two-dimensional space, the grid points represent a square regular grid.

Each point of the grid can be seen as a representation of the 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. A probability of a feature being in a transaction is also stored and it may depend on the distance from the spatial object. For example, the grid point gp_2 in Fig. 1 (c) is located closer to the point A_1 than the point gp_1; therefore, $p(A, gp_2) > p(A, gp_1)$. The granularity of the grid should be carefully chosen for each application and it could depend on the average size of the region covered by a spatial object and its buffer. Choosing too great a distance between grid points may negatively affect the accuracy of the results because small feature regions and their overlaps might get a different number of intersecting grid points depending on the grid imposition. The short distance between grid points leads to a great number of derived transactions, and the following computation of pattern significance levels might become prohibitively expensive.

Given a set of transactions T, derived after the transactionization of the spatial dataset, and a set of spatial features F, the prevalence measure value is calculated for all candidate co-location patterns or rules. In some applications experts look for sets of features that are co-located with each other. The expected support $ExpSup(P)$ might be used to define the level of the interestingness of a pattern P. For other applications, researchers intend to analyze a predefined set of rules. For example, for a dataset of disease outbreaks and possible cause

factors a typical co-location rule is of the form $C \rightarrow D$, where C is a subset of cause features and D is a disease feature. 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 $X \cap Y = \emptyset$. Algorithm 1 shows the outline of our model in the case when co-location patterns are mined.

Definition 1 *The probability $p(P,t)$ of the pattern P occurring in a transaction t is the product of the corresponding feature instance probabilities, $p(P,t) = \prod_{f \in P} p(f,t)$.*

Definition 2 *The expected support $ExpSup(P)$ of a pattern P is defined as the sum of expected probabilities of presence of P in each of the transactions t in the database, $ExpSup(P) = \sum_{t \in T} p(P,t)$.*

Definition 3 *The expected confidence $ExpConf(X \rightarrow Y)$ of a rule $X \rightarrow Y$ is defined as $ExpConf(X \rightarrow Y) = ExpSup(X \cup Y)/ExpSup(X)$.*

The next step, the statistical test, helps to estimate the likelihood of seeing the same level of the prevalence measure or greater under a null hypothesis that features of a pattern or rule are spatially independent from each other.

Definition 4 *A pattern P is said to be significant at level α, if the probability p of seeing the observed expected support $ExpSup_{obs}$ or larger in a dataset, complying with a null hypothesis, is not greater than α. (The same for $ExpConf_{obs}$)*

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 $X \rightarrow Y$ in a randomized dataset which is generated under the null hypothesis. In order to estimate the probability p, the expected confidence of the co-location rule in R randomized datasets is calculated. Having the number of simulations R, the value of p is computed as:

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

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 the predefined level of significance α, the null hypothesis is rejected. Therefore, the co-location rule $X \rightarrow Y$ is significant at level α.

3.2 Candidate Filtering Techniques

The calculation of the p-value is repeated for all candidate co-location patterns or rules. The number of candidates grows exponentially with the number of spatial

features in the dataset. In addition, the 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 less transactions contain all its features. Therefore, one 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 candidate patterns and rules that are de facto insignificant.

- First, after the calculation of the prevalence measure for candidate patterns in a real dataset, a subset of patterns may have a prevalence measure value equal to zero. Obviously, these patterns cannot be statistically significant and they can be excluded from the set of candidate patterns (lines 6-7 in Algorithm 1).
- Second, during the calculation of the p-value for the candidate patterns for which the observed prevalence is higher than zero, some of the candidate patterns might show a p-value that exceeds 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 $((5+1)/(99+1) > 0.05)$ and, therefore, can be excluded from the following 89 checks (lines 15-17 in Algorithm 1). With this filter, after the last simulation run the set of candidates contains only significant patterns.

4 Modeling Framework

The modeling framework that is used to handle and analyze the data is an important part of practical research. In theoretical studies it could be simplified in order to generalize the task and define algorithms that could be applied for a wide range of applications. However, the usage of general approaches and algorithms may result in misleading or even wrong results. For example, the neighborhood distance threshold is an important measure of 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, e.g., 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 obstacle. Most domains of research, if not all, have their own nuances that must be taken into account by researchers in order to get most accurate and significant results.

The motivating task of this paper, detecting co-locations of pollutants and cancer cases, has unique difficulties and challenges. The distribution of a pollutant is not uniform and it could depend on several factors: types of pollutants,

Algorithm 1. Mining significant co-location patterns

Input: Spatial dataset S; Level of significance α; Number of simulation runs R.
Output: Set of significant co-location patterns P

```
1: Impose a grid over the real dataset
2: T ← set of derived transactions
3: CP ← set of candidate patterns
4: for each cp ∈ CP do
5:    cp.ExpSup_obs ← ComputeExpSup(T)
6:    if cp.ExpSup_obs = 0 then
7:       CP ← CP/cp
8:    end if
9: end for
10: for i = 1 → R do
11:    Impose a grid over the i-th randomized dataset
12:    T ← set of derived transactions
13:    for each cp ∈ CP do
14:       cp.ExpSup_sim[i] ← ComputeExpSup(T)
15:       if cp.ExpSup_sim[i] ≥ cp.ExpSup_obs then
16:          cp.R_≥ExpSup_obs ← cp.R_≥ExpSup_obs + 1
17:          cp.α ← (cp.R_≥ExpSup_obs + 1) / (R + 1)
18:          if cp.α > α then
19:             CP ← CP/cp
20:          end if
21:       end if
22:    end for
23: end for
24: P ← CP
25: return P
```

amounts of release, climatic conditions (wind, precipitation), topography, etc. Various chemicals have different levels of harmfulness. In addition, the pollutant concentration is directly proportional to the 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 the concentration of chemicals. Certainly, we do not aim to reproduce complicated air pollution distribution models. Instead, our model gives a simple framework that increases the accuracy of results while operating with available data.

4.1 Pollutant Amounts

The dataset on pollutants contains the 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. The range of the

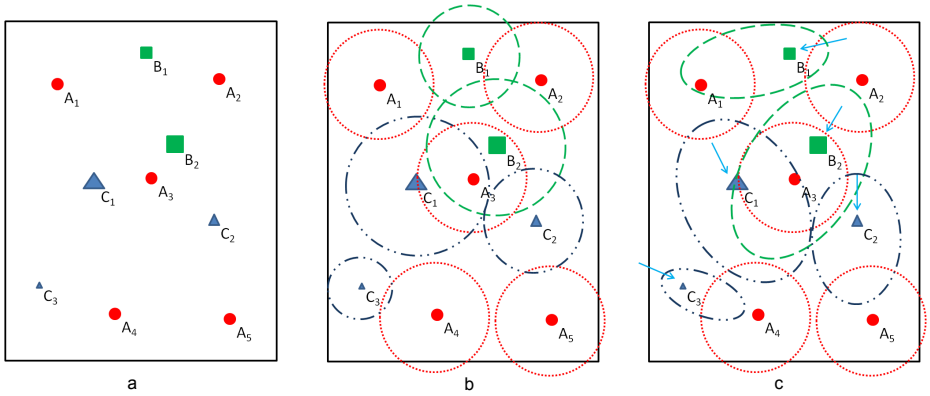

Fig. 2. Modeling framework usage examples: (a) an example spatial dataset (A - cancer, B and C - pollutants); (b) buffer sizes vary depending on the pollutant release amount; (c) buffer shapes change with the wind direction and speed (shown by arrows).

average amount values varies from several kilograms to tens of thousands tons; the maximum average yearly release in the dataset is 80,000 tons. Certainly, one distance threshold for all pollutant emissions is inaccurate, because the more a chemical is released, the farther it distributes from a source point. Fig. 2 (a) displays an example dataset containing cancer points (feature A) and chemical points (features B and C). On Fig. 2 (b) buffer zones around pollutant points are based on the amount released. 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 denote average active living zones.

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

4.2 Wind Speed and Direction

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

Regarding the wind speed and direction, two situations are possible. First, the region where a facility is located is windless throughout the year. In this case, the pollutant is assumed to disperse in a circular region around the source point with the radius of the circle derived from the released amount as discussed in the previous subsection. However, the second situation is more frequent - there is nonzero wind speed with a prevailing wind direction. In this case we presume that the original distribution circle is morphed into a more ellipse-like region. Our calculations of the characteristics of the ellipse are based on the works by Getis and Jackson [12], and Reggente and Lilienthal [13]. The major axis of the ellipse is in the direction of the prevailing wind. Furthermore, the coverage area of the ellipse is supposed to remain constant. 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. Fig. 2 (c) illustrates elliptical buffer regions; their forms are dependent on the wind speed and its frequent direction.

The lengths of the major semi-axis a and minor semi-axis b are derived from the equations: $a = r + \gamma|\boldsymbol{v}|$, $b = \frac{r^2}{a}$, where r - the radius of the original circle, $\boldsymbol{v}$ - the wind speed, and γ - the stretching coefficient.

The larger the value of the stretching coefficient, the longer is the length of the ellipse's major axis. We fixed γ at 0.3, but it could have a different value for each of the pollutants. The calculation of b follows our assumption that the area of the ellipse is equal to the area of the original circle.

In order to get the values of the wind speed and prevailing wind direction, the interpolation of wind fields between weather stations is used. The data of monitoring stations comes from two sources. First, the data on 18 stations is obtained from Environment Canada, which provide 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 the direction (out of possible eight directions) with the highest average occurrence count. Second, the data on 156 stations is derived from AgroClimatic Information Service (ACIS), a provincial government service. The data is daily, between 2005 and 2011. In order to make the data consistent, the average wind speed and the most frequent wind direction are calculated using the same methods as for the federal government website data.The climate normals from two sources are combined and used to make interpolations in ArcGIS tool [14]. 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) the number goes only in one direction. 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°.

Interpolation of wind fields requires a technique that considers nonlinear nature of the wind direction attribute. The transformation is done according to the work by Williams [15]. 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 $X = S \sin\theta$ (northern wind) and $Y = S \cos\theta$ (eastern wind).

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:

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

The calculation of wind direction angle θ' is more complicated. From geometry, the wind direction is calculated as $\theta' = \tan^{-1}(Y'/X')$. However, due to the fact that the inverse tangent is defined only for values between -90° and 90°, each quadrant (the section of the graph which depends on the signs of wind vector components; for example, Quadrant I is bounded by positive X' and Y', Quadrant II - by positive X' and negative Y') requires its own formula [15]. As a result we get interpolated values of wind speed and wind direction for each point of the studied space.

5 Experimental Evaluation

We conducted experiments on a real dataset, containing data on pollutant emissions and childhood cancer cases. The information on pollutants is for the 2002-2007 period and contains the type of 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 1,465 pollutant emission points; several chemicals might be released from the same location. The number of cancer points (addresses where a child lived when cancer was diagnosed) is 1,254. Claiming discovering causality is wrong and controversial and thus we attempt only to find "associations" rather than "causalities". The results are still subject to careful evaluation by domain experts in our multidisciplinary team and the publication of the found associations is not authorised at this point. It suffices to mention, however, that some surprising rules were discovered indicating significant association between groups of chemicals, that were not categorized individually as carcinogens, and childhood cancer, as well as rules with pairs of chemicals such that one was known as carcinogenic but did not associate with cancer in our data except in the presence of another that acted as a catalyzer.

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; its effect is also evaluated. The number of simulations for the statistical test is set to 99, so that with the observed data the denominator in Equation (1) 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.

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.

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 a naive approach all candidates would be checked in each simulation run. With our filter excluding rules with zero-level confidence, 10,125 candidates remain. The usage of the second filtering method (the exclusion of candidates which p-values passed α) considerably reduces the amount of computation. While in the first simulation run the confidence value is computed for 10,125 rules, in the last run only 482 candidates are evaluated.

Effect of the Grid Granularity As already mentioned, the granularity of the grid (the distance between grid points which affects the number of points per unit of space) is crucial for the accuracy of the results. Having too long 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, too short distance between points leads to the greater number of transactions. Decreasing the distance by two increases the transaction set size approximately by four. Therefore, more computation needs to be done. The grid resolution might be set up depending on the average buffer size.

In addition to the grid with 1 km granularity, we conducted two experiments with 2 and 0.5 km grids. As mentioned above, the algorithm finds 482 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 less number of transactions. 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 the results improves.

6 Conclusion

In this paper, we proposed a new solution to the co-location mining problem. The transactionization step allows the conversion of spatial data into a set of transactions. The usage of thresholds like in previous algorithms is replaced by the statistical test. In addition, our approach takes into account uncertainty of data. In order to decrease computation, the filtering techniques are presented. The experiments on real and synthetic datasets showed that our approach finds significant co-location patterns and rules.

References

1. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In: Proc. of the 7th International Symposium on Advances in Spatial and Temporal Databases. (2001) 236–256
2. Huang, Y., Pei, J., Xiong, H.: Mining co-location patterns with rare events from spatial data sets. Geoinformatica **10**(3) (2006) 239–260
3. Yoo, J.S., Shekhar, S.: A joinless approach for mining spatial colocation patterns. IEEE Trans. on Knowl. and Data Eng. **18**(10) (2006) 1323–1337
4. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoo, J.S.: A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proc. of the 2004 SIAM international conference on data mining. (2004)
5. Cressie, N.: Statistics for spatial data. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley (1991)
6. Chou, Y.: Exploring spatial analysis in geographic information systems. OnWord Press (1997)
7. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of the 20th International Conference on Very Large Data Bases. (1994) 487–499
8. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases. In: Proc. of the 4th International Symposium on Advances in Spatial Databases. (1995) 47–66
9. Morimoto, Y.: Mining frequent neighboring class sets in spatial databases. In: Proc. of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. (2001) 353–358
10. Estivill-Castro, V., Lee, I.: Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data. In: Proc. of the 6th International Conference on Geocomputation. (2001)
11. Barua, S., Sander, J.: SSCP: mining statistically significant co-location patterns. In: Proc. of the 12th international conference on Advances in spatial and temporal databases. (2011) 2–20
12. Getis, A., Jackson, P.H.: The expected proportion of a region polluted, by k sources. Geographical Analysis **3**(3) (1971) 256–261
13. Reggente, M., Lilienthal, A.J.: Using local wind information for gas distribution mapping in outdoor environments with a mobile robot. In: Sensors, 2009 IEEE. (2009) 1715 –1720
14. : ArcGIS Desktop: Release 10 (ESRI 2011)
15. Williams, R.G.: Nonlinear surface interpolations: Which way is the wind blowing? In: Proc. of 1999 Esri International User Conference. (1999)