Discovering Spatial Contrast and Common Sets with Statistically Significant Co-location Patterns

Mohomed Shazan Mohomed Jabbar
Department of Computing Science
University of Alberta
Edmonton, Canada
mohomedj@ualberta.ca

Osmar R. Zaïane
Department of Computing Science
University of Alberta
Edmonton, Canada
zaiane@ualberta.ca

Alvaro Osornio-Vargas
Department of Paediatrics
University of Alberta
Edmonton, Canada
osornio@ualberta.ca

ABSTRACT

Co-location pattern mining is a spatial data mining technique which can be used to find associations among spatial features. Our work is motivated by an application in environmental health where the goal is to investigate whether the maternal exposure during pregnancy to air pollutants could be potentially associated with adverse birth outcomes. Discovering such relationships can be defined as finding spatial associations (i.e. co-location patterns) between adverse birth outcomes and air pollutant emissions. In particular, our application problem requires to find specific co-location patterns which are common to many spatial groups and co-location patterns which can discriminate one spatial group from the others. Traditional co-location pattern mining methods are not capable of finding such specific patterns. Hence, to achieve the spatial group comparison task, we introduce two new spatial patterns: spatial contrast sets and spatial common sets, and techniques to efficiently mine them based on co-location pattern mining. Traditional co-location pattern mining methods rely on frequency based thresholds which discard rare patterns and find exaggerated noisy patterns which may not be equally prevalent in unseen data. Addressing these limitations, we propose to use statistical significance tests instead of frequency to quantify the strength of a pattern. Towards this end, we propose to apply Fisher's exact test to efficiently find statistically significant co-location rules and use them to discover spatial contrast and common sets. Our experiments reveal that the Fisher's test based method could indeed help in finding co-location patterns with a better statistical significance leading to find valid spatial contrast and common sets. With the proposed methods we discovered that air pollutants such as heavy metals, NO_2 and PM are significantly associated with adverse birth outcomes conforming to the existing domain knowledge thus validating our approach. We also evaluated our methods with synthetic datasets which confirmed that our methods indeed extract the patterns we seek to find.

SAC 2017, April 03-07, 2017, Marrakech, Morocco

DOI: http://dx.doi.org/10.1145/3019612.3019665

CCS Concepts

•**Information systems** → **Geographic information systems; Data mining;** Association rules;

Keywords

Co-location patterns; spatial contrast sets; spatial common sets

1. INTRODUCTION

Co-location pattern mining is an important class of spatial data mining algorithms which aims to discover relationships and associations among various spatial features. More specifically a co-location pattern can be defined as a "set of spatial features which are often located together in spatial proximity". There is a wide range of applications of co-location pattern mining varying from business to science. Our current work is motivated by a challenging research question in environmental health: "Do air pollutant emissions play any role in adverse birth outcomes?". We work with the datasets collected by the Canadian Neonatal Network[1] (CNN) to discover such potential relationships between industrial air pollutant emissions and adverse birth cases in 21 Canadian cities. There are many studies suggesting that associations between air pollutants and Adverse Birth Outcomes (ABOs) exist [7]. Discovering such associations turns out to be a co-location pattern mining problem where the goal is to find co-location patterns based on the overlap of air pollutant emission regions and maternal mobility regions during pregnancy. However, when dealing with rich datasets which contain data from multiple spatial regions, one has to look beyond the traditional co-location patterns to discover specific patterns which can uniquely characterize a specific spatial region and contrast it from others and patterns which are common in many spatial regions. For instance, when given adverse birth occurrences from multiple cities in a country like Canada, a valid research question leading to such a mining task would be: "Is there any specific combination of industrial air pollutants more associated to low birth weight in Toronto area than any other city in Canada?" To answer such questions, in contrast to classical co-location patterns, more specific discriminative co-location patterns which can contrast a particular spatial group from the others or co-location patterns

[1] http://www.canadianneonatalnetwork.org/portal/

which are commonly significant in many spatial groups could be of great use. Furthermore, the statistical significance of spatial patterns is important when addressing challenges in environmental health to improve the trust in results and to find patterns with rare occurrences. Considering these, there are three major challenges when dealing with spatial pattern mining problems as above to find specific co-location patterns which are either discriminative or common: 1) Efficiently finding statistically significant co-location patterns irrespective of their prevalence, 2) Finding co-location patterns which can contrast specific spatial groups (i.e. spatial contrast sets), and 3) Finding co-location patterns which are common to many spatial groups (i.e. spatial common sets). In our work, we attempt to address these three challenges.

Traditional co-location pattern mining techniques are not capable of finding rarely occurring but statistically significant patterns owing to the fact that they heavily rely on global prevalence thresholds [16]. Addressing these limitations in existing techniques, few transaction based approaches piggybacking on association rule mining techniques have been proposed recently to find statistically significant co-location patterns [1, 14]. On the other hand, although there exists a class of techniques called contrast set mining to discover discriminative association patterns to characterize a particular group and contrast it from the other groups in non-spatial datasets [11], a similar variant which can discover significant contrast sets to differentiate spatial groups does not exist. Furthermore, no significant work has been done to find co-location patterns which are common to many spatial groups as well. Addressing these limitations and gaps in existing work we propose to use Fisher's exact test to efficiently find statistically significant co-location patterns and further analyze those patterns to discover discriminative co-location patterns which can uniquely characterize a given spatial group and to discover common co-location patterns which are commonly significant in many spatial groups. In this work we introduce two novel classes of co-location patterns called *Spatial Contrast Sets* and *Spatial Common Sets* and propose two new algorithms to discover them. We successfully applied these proposed methods to find interesting associations between air pollutants and ABOs in Canada which suggested that it can be applied to address the above challenges in other similar application problems as well.

The remainder of the paper is organized as follows. In Section 2 we introduce several foundational concepts and their formal definitions needed in developing our methods. We give an overview of some of the important previous works in Section 3. In Section 4 we discuss the dataset and the methodology of designing and developing our co-location pattern mining framework. We discuss our experimental results in Section 5 and we conclude in Section 6.

2. PRELIMINARIES

Co-location rule mining and spatial contrast or common set mining have strong foundations in the association rule mining problem domain. Hence, we formulate our core framework around association rule mining techniques. In association rule analysis, we deal with a transaction database D such that each sample transaction E in D can be defined as a vector of size m. Let $A = \{A_1, A_2, ..., A_m\}$ be a set of feature-value pairs (i.e. $A_1 = (f_1, v_{f1})$ where $f_1 \in F$ is a feature and v_{f1} is its corresponding value) called *items*. Then a transaction E can be defined as a vector consisting of feature-value pairs $\{A_i, A_j, ..., A_m\} \subset A$. Given these values, an **association rule** can be defined as in Definition 1.

Definition 1. An *association rule* is an implication of the form $X \implies Y$ where $X \subset A$, $Y \subset A$ and $X \cap Y = \emptyset$.

Confidence c in $X \implies Y$ is the percentage of data instances in D containing X also containing Y (i.e. $P(Y|X)$). Support s for $X \implies Y$ is the percentage of data instances in D containing $X \cup Y$. Traditional algorithms discover strong association rules by verifying that their s and c exceed some user defined thresholds. Classification association rules (CAR) are a special case of general association rules [2]. Given a set of class labels $C = \{c_1, c_2, ..., c_q\}$ where each instance E in D is associated with a class label c_i and $|C| = q$, a CAR can be defined as an association rule of the form $X \implies c_i$. In such a rule $X \subset A$ and $c_i \in C$.

Given a spatial database S, using a transactionization technique if it can be transformed into a transaction database D^s, item set A^s represents a set of spatial feature-value pairs and E^s represents the data instances in D^s. Given these, based on the definition of the association rules, a co-location rule can be defined as in Definition 2.

Definition 2. A *co-location rule* is an implication of the form $X \implies Y$ where $X \subset A^s$, $Y \subset A^s$ and $X \cap Y = \emptyset$.

Contrast sets are another class of associative patterns which are used to characterize a particular class and contrast it from the others. It can be defined as in Definition 3.

Definition 3. Contrast sets are conjunctions of attribute-value pairs, $X \subset A$, defined on mutually exclusive classes from C such that no $A_i \in X$ occurs more than once.

Contrast sets can be discovered using class association rules. Originally, if set X in class association rule $X \implies c_i$ meets STUCCO deviation conditions [4] as in Equation 1 and 2, then X is considered as a contrast set for class c_i which can distinguish c_i from the other classes. The condition in Equation 1 imposes that the support of a contrast set is significantly different across various groups. The second condition in Equation 2 imposes that the difference of support of a contrast set across different groups is sufficiently large.

$$\exists_{i,j} P(X|c_i) \neq P(X|c_j) \tag{1}$$

$$\max_{i,j} |support(X, c_i) - support(X, c_j)| \geq min\_dev \tag{2}$$

3. RELATED WORK

Traditional co-location rule mining techniques are based on the neighborhood relations and participation indices [15]. In such methods, co-location patterns take the form $C_1 \implies C_2(PI, cp)$, where C_1 and C_2 are spatial feature sets, PI is the participation index or the prevalence measure for the given rule and cp is the conditional probability. The given rule is considered prevalent or interesting only when at least PI% of the instances of each of the features in the rule form a clique with the instances of every other feature in the same rule according to a defined neighbourhood relation. To find rare patterns, some of the previous works have introduced a new measure called max participation ratio $maxPR$% where, if $maxPR$% instances of at least one of the features in the given pattern form a neighbourhood relation with instances of all the other features in the same

pattern, then that co-location pattern is considered prevalent [10]. Most of these techniques depend on user-defined thresholds for interestingness measure and detect a large number of noisy patterns when the threshold is low, and lose rare patterns if the threshold is high.

Due to the limitations indicated above, as an alternative, it is suggested to use statistical significance tests to evaluate a rule [3]. In this approach [3] also a participation index is computed for each pattern in the observed dataset and if the probability of seeing an equal or a greater index under the null hypothesis model is lower than a given level of significance, the pattern is considered as "statistically significant". Another transactionization based co-location mining approach adapts a similar method to find statistically significant rules and is also capable of handling extended spatial objects [1]. However, due to the anti-monotonicity property of the statistical significance, above methods which are based on empirical p-value and random data generation does not scale well with large feature sets. In fact some of the above methods, which generate all the possible patterns and compute the empirical p-value, had to limit the rule size to four in order to avoid the exponential growth of the computational complexity [14]. Addressing this issue a recent technique [14] transforms the spatial dataset into a transaction dataset and applies an efficient statistically significant association rule discovery algorithm called StatApriori [8] to find statistically significant co-location patterns. However, more recently, a better algorithm which is more robust than the StatApriori has been proposed to find statistically significant association rules using Fisher's exact test [9]. In our work we exploit this new approach to robustly find statistically significant co-location patterns.

Contrast sets were first introduced through the STUCCO [4] algorithm as a way to contrast a specific group from the others. As explained in Section 2, STUCCO uses two conditions to find such strong contrasting patterns. Most of the existing contrast set mining techniques like STUCCO depend on two threshold values called support and confidence, and prone to the limitations imposed by them. Hence, as an alternative, it is proposed to use statistically significant association rules to mine contrast sets [11]. However, contrast set mining problem in spatial datasets remains an underexplored area of research. Similarly in existing work no formal definitions of spatial common sets or techniques to discover them have been proposed. This necessitates forming new definitions and developing novel techniques to find contrast and common sets in spatial data mining domain.

4. METHODS

In this section, we first discuss the spatial data for our motivation problem and the preprocessing steps we performed on them. Then we present the outlines of the analytical methods we used in our analysis including two new algorithms proposed to mine spatial contrast and common sets. Those are as follows:

1. Fisher's test based co-location pattern mining method
2. DiSConS: A method to Discover Spatial Contrast Sets
3. DiSComS: A method to Discover Spatial Common Sets

4.1 Data and Preprocessing

In this work, our motivating research question is: "what are the relationships between air pollutants released by industries and adverse birth outcome in Canada?" To address this question, we primarily worked on the datasets collected by the Canadian Neonatal Network about babies admitted to Neonatal Intensive Care Units (NICUs) across 21 cities in Canada during the period of 2006-2010. We compiled the original CNN dataset and obtained 32,836 adverse birth outcome cases with geolocations. We regrouped this dataset to 19 Census Metropolitan Areas (CMAs) in Canada. In this dataset there are three main ABOs of interest: 1) Preterm birth (PTB), 2) Low birth weight at term (LBW), and 3) Small for Gestational Age (SGA). To obtain the air pollutant information of the above CMAs of interest, we used the datasets from the National Pollutant Release Inventory (NPRI) [6] of Canada. More specifically we chose industrial facilities within the 100 km radius of each of the CMA polygons. We only considered the air pollutant emissions from each of the industrial facilities within the time period of 2005-2010. This dataset contains data on estimated yearly releases of 127 chemicals. Finally, to model the air pollutant dispersion and to extend chemical release points to regions we used wind speed and direction data from Environment Canada. We obtained this data from 47 National Air Pollutant Surveillance stations.

In our application problem, we deal with two types of point spatial data objects: 1) ABO cases, and 2) Chemical emission points. We extend these two types of points objects to represent the maternal mobility range of ABO cases and the dispersion region of the air pollutants emitted more accurately. For ABO cases, we define a circular buffer region with a fixed radius (e.g. 5 km) originating from the maternal geolocation to represent the maternal mobility range during the pregnancy. On the other hand, the distribution of a particular pollutant in a given region is not uniform. It could depend on the type of the pollutant, the amount of release, the weather conditions (the wind, precipitation) in the region, topography, etc. We considered some of these factors such as pollutant release amount, toxicity, wind speed and direction when defining the buffer zones of chemical emission points. However, we do not intend to reinvent a comprehensive air pollution distribution model which requires considering many other variables. Instead, we attempt to capture some important real world attributes with available data to improve the overall accuracy of our findings. Firstly, we use the yearly amount of average chemicals released by a facility in a given location to determine their buffer sizes. Based on the previous works [13] we defined the radius of these buffers as the natural logarithm function of the amount of chemicals released at the given location. Then we morph this circular buffer region into an elliptical buffer region based on the average wind speed and direction in that location to more realistically model the chemical dispersion. In this model we used, it is assumed that although the affected region can be different, the area affected by the pollutant is the same [13]. We obtained the average wind speed and direction at chemical emission points from the Environment Canada dataset. Given that information, the lengths of the major semi-axis a and minor semi-axis b of the new elliptical buffer region can be computed using the following equations [13].

$$a = r + \gamma|\vec{v}|, \tag{3}$$

$$b = \frac{r^2}{a}, \tag{4}$$

where r is the radius of the original circle, $\vec{v}$ is the wind speed, and γ is the stretching coefficient. In our experiments, we have used 0.3 as the stretching coefficient. An example scenario for extending the point objects is given in Figure 1.

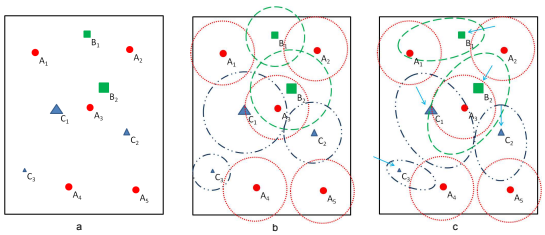

Figure 1: Extending spatial objects: (a) An example spatial dataset (A - ABO, B and C - Pollutants); (b) Buffer sizes of pollutants vary depending on the amount of release; (c) Buffer shapes of pollutant emission points change with the wind direction and speed (as indicated by arrows) [13]

4.2 Co-location Rule Mining Algorithm

The Fisher's test based co-location pattern mining approach we use consists of two major steps: 1) Transactionizing the spatial dataset; 2) Mining for statistically significant association rules with Fisher's test.

4.2.1 *Transactionization*

Transactionization helps to transform a spatial data set to a set of transaction data. This immensely helps to use existing association rule mining techniques on them for the purpose of finding co-location patterns. However, due to the limitations in previous transactionization approaches, such as window-centric and reference-centric models, we adapted a recently proposed grid based transactionization method [1, 14]. This method is outlined in Algorithm 1. Given a spatial dataset S, Algorithm 1 initially generates a set of grid points by overlaying a grid with a suitable granularity level (e.g. 0.5, 1 or 2 km) over the geographic space covering the instances in S. Each point in this grid can be seen as a representation of a specific part of the corresponding geographic space. Once the grid points are obtained, Algorithm 1 defines buffer zones around spatial objects in S as we discussed in Section 4.1. In the next step of the algorithm the constructed grid is imposed over the dataset S. 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. Hence each grid point can be considered as a potential candidate to obtain a transaction. The granularity of the grid should be chosen carefully for each application, and it may depend on an average size of a region covered by a spatial object and its buffer. Once this transactionization is performed an association rule mining technique can be applied on the resulting transaction dataset to find spatial association rules.

4.2.2 *Co-location Rule Mining*

Association rules can be viewed as dependency rules and the statistical significance of the dependency might not be related to the frequency at all. Hence to address the limitations in traditional support-confidence based methods, it has been proposed to adapt an association rule mining approach based on statistical significance tests. Given a rule $X \rightarrow A$, such tests are designed to test the dependency between X and A. The null hypothesis in such a test will be "X and A are independent of each other". The statistical significance of the dependency between X and A is tested by computing the p-value, the probability that the observed or a stronger dependency would have occurred by chance. If this p-value is smaller than a given level of significance α the null hypothesis can be rejected and it can be accepted that the dependency between X and A is statistically significant. In our work, we use Fisher's exact test to measure this statistical significance of rules. Fisher's exact test is commonly used with categorical data where the data objects can be classified in two different ways. In such datasets the test can be used to compute the significance of the association between two kinds of classification. In our case, when given a rule $X \rightarrow A$ these classification schemes are: 1) Data objects having X or not; and 2) Data objects having A or not. Fisher's p-value (i.e. p_F) for such datasets can be computed using the following cumulative hypergeometric distribution.

Algorithm 1 *GetTransactions(S)*: Transactionization step.

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

$$p_F(X \rightarrow A) = \sum_{i=0}^{J} \frac{\binom{m(X)}{m(XA)+i}\binom{m(\neg X)}{m(\neg X \neg A)+i}}{\binom{n}{m(A)+i}} \tag{5}$$

where $J = min\{m(X\neg A), m(\neg XA)\}$, n is the number of total transactions, and $m(.)$ computes the frequency of transactions containing the given items [9]. Another important task in statistically significant rule discovery is to identify redundant rules. A rule, $X \rightarrow A$ can be identified as redundant if there exists a rule, $Y \rightarrow A$ where $Y \subset X$ and M($Y \rightarrow A$) is equally good or better than M($X \rightarrow A$). Here the M is a goodness measure (e.g. Fisher's p-value). We use the Kingfisher algorithm [9] which implements an efficient branch and bound search mechanism on an enumeration tree to detect such non-redundant and statistically significant association rules based on Fisher's exact test. When used with transactionized spatial data it successfully detects non redundant and statistically significant co-location rules. We constrained the Kingfisher algorithm to only to produce co-location rules of the form $X \rightarrow A$ where $A \in \{SGA, LBW, PTB\}$ and X is a set of chemicals.

4.3 Compare and Contrast Spatial Groups

Some of the statistically significant co-location rules we detected using the above approach for various spatial regions could be used to uniquely characterize and contrast a particular spatial group from the others. On the other hand some co-location rules can be useful to represent patterns

which are consistently statistically significant in many spatial groups or regions. In this context, spatial groups can be defined as mutually exclusive groups represented by a specific class and associated with a specific geolocation. PTB cases in Vancouver, LBW cases in Edmonton, and SGA cases in Hamilton can be considered as some of the spatial groups from our motivating application. The first type of rules is useful to discover associations between air pollutants and ABOs, which are specific to a particular spatial group leading to take necessary actions to handle the condition locally. On the other hand, the second type of rules is useful to recognize co-location patterns between industrial air pollutants and ABOs that are common in many spatial regions leading to take necessary actions and create policies to affect many spatial regions or groups. Towards this goal we further analyze the co-location rules we detected previously to discover following two novel classes of patterns: 1) Spatial contrast sets to identify unique patterns which can characterize or contrast a particular spatial group; and 2) Spatial common sets to identify patterns which can commonly be seen across many spatial regions/groups.

4.3.1 Spatial Contrast Sets

As we explained previously, contrast sets can characterize a particular group of data instances and can be used to contrast them from the data belonging to other groups. When dealing with spatial data mining problems identifying contrast sets for groups in specific spatial regions could be of great use to understand which unique variables that are associated with a particular outcome or class in a given spatial region can contrast the same outcome occurring in other regions. We propose a novel type of contrast sets called *Spatial Contrast Sets* to achieve this goal. A formal definition for spatial contrast sets is given in Definition 4.

Definition 4. A *spatial contrast set* is a conjunction of spatial attribute-value pairs (i.e. $A_i = V_{i_j}, ..., A_k = V_{k_l}$ where $A_i \in A$, $A_k \in A$ and in the case of binary variables $V_{i_j} \in \{0, 1\}$ and $V_{k_l} \in \{0, 1\}$) defined on mutually exclusive groups $G_{11}, ..., G_{1,p}, ..., G_{q,1}, ..., G_{q,p}$, where $G_{x,y} = \{C_x, L_y\}$; C_x is the class membership and L_y is the location of the group. Furthermore, q is the number of mutually exclusive classes and p is the number of mutually exclusive spatial regions exist in the given dataset.

Given a statistically significant co-location rule of the form $X \rightarrow G_{x,y}$, X is a spatial contrast set for the group $G_{x,y}$ over any other groups of interest $G_{p,q} \in G^s \setminus \{G_{x,y}\}$ if Equation 6 and 7 hold $\forall G_{p,q} \in G^s \setminus \{G_{x,y}\}$.

$$p_F(X \rightarrow G_{x,y}) \leq p_F(X \rightarrow G_{p,q}) \quad (6)$$

$$\max_{p,q} |support(X, G_{x,y}) - support(X, G_{p,q})| \geq min\_dev \quad (7)$$

where the $p_F(X \rightarrow G_{x,y})$ is the Fisher's p-value for the co-location pattern and $support(X, G_{x,y})$ is the support of X in the subset of data that belongs to $G_{x,y}$. The first constraint tests whether a candidate contrast set is more *statistically significant* in the associated spatial group than in the other groups. The second constraint tests whether the support of a candidate contrast set is *sufficiently large* in the associated spatial group than in the other groups. These constraints can be used to find contrast sets among three different types of spatial groups as follows:

1. If we fix that $\forall y = q$ we can contrast data which belong to the same spatial region but in different classes.
2. If we fix that $\forall x = p$ we can contrast data which are in the same class but belong to different spatial regions.
3. $\forall x$ and $\forall y$ we can contrast data which belong to different classes in different spatial regions.

Based on the type of application these conditions can be used interchangeably to find interesting spatial contrast-sets. Our proposed algorithm DiSConS to mine such spatial contrast sets is shown in Algorithm 2. DiSConS first discovers statistically significant classification co-location rules of the form $X \rightarrow G_{x,y}$ using the approach we proposed previously for each spatial region $l \in L$ in the dataset (see line 2-9). Then for each group it searches for contrast sets by imposing the conditions presented in the Equation 6, and 7 on the candidate co-location patterns found in the previous step.

Algorithm 2 DiSConS

INPUT: Database S, Attributes A, Classes C, Locations L, Level-of-Significance α, Spatial-Groups G^s

```
1: CANDS=2DHashTable()
2: for all Location l in L do
3:    t_l = GetTransactions(S_l, A)
4:    SCAR_l = ContrainedKingfisher(t_l, C, α)
5:    for all rule X → G_{c_i,l} in SCAR_l do
6:       if CANDS[l][c_i] == ∅
7:          CANDS[l][c_i] = HashTable()
8:       CANDS[l][c_i][X] = M(X → G_{c_i,l})
9:    end for
10: end for
11: CSET=2DHashTable()
12: for all G_{x,y} in G^s do
13:    CSET[L_y][C_x] = [∅]
14:    for all X in CANDS[L_y][C_X].keys() do
15:       if ∀ G_{p,q} ∈ G^s \ {G_{x,y}} Equation 6 and 7 is TRUE
16:          CSET[L_y][C_x].append(X)
17:    end for
18: end for
RETURN CSET
```

4.3.2 Spatial Common Sets

As opposed to spatial contrast sets which are helpful in contrasting a particular spatial group from the others, another type of patterns of interest would be the ones which can characterize or represent a set of similar spatial groups. For example, a particular feature value combination set X can be consistently significant in all or a majority of the spatial groups, (PTB, Toronto), (LBW, Edmonton), (SGA, Calgary), etc. Such patterns could be useful to identify important feature sets which are associated with many adverse birth outcomes in various spatial regions. We define such sets as *Spatial Common Sets* and the same formal definition for spatial contrast sets (i.e. Definition 4) can be used to define spatial common sets as well. Given a co-location pattern $X \rightarrow G_{x,y}$, a set of spatial groups, G^s, a MinFrac threshold and a maximum deviation threshold, maxdev, X is a spatial common set if $\exists G^{s'} \subset G^s$ where for all $G_{x,y} \in G^{s'}, G_{p,q} \in G^{s'}$ the constraints given in Equation 8 and 9 can be satisfied and the $|G^{s'}| > MinFrac$ threshold.

$$p_F(X \rightarrow G_{x,y}) - p_F(X \rightarrow G_{p,q}) \leq max - pF - diff \quad (8)$$

$$|support(X, G_{x,y}) - support(X, G_{p,q})| \leq max - dev \quad (9)$$

max-pF-diff is a user defined threshold to control the variation of the significance of a common set among the given set of spatial groups. *max-dev* is the maximum support difference, allowed to be between any two different groups in the given set of groups. These two constraints make sure that the *statistical significance* and the support of the common set does not vary significantly across spatial groups. Similar to spatial contrast sets, we can find common sets for three different types of spatial groups:

1. If we fix that $\forall y = q$ we can find patterns common in data which belong to the same spatial regions but different classes.
2. If we fix that $\forall x = p$ we can find patterns common in data which belong to different spatial regions but in the same class.
3. If $\forall x$ and $\forall y$ we can find patterns common in data which belong to different classes in different spatial regions.

Our proposed algorithm DiSComS to mine such spatial contrast sets is shown in Algorithm 3. DiSComS first generates all the classification co-location rules of the form $X \rightarrow G_{c_i,l}$ for each location $l \in L$ using the co-location pattern mining approach we previously discussed. Antecedents of each of the retrieved rules are added to the candidate spatial common set pool. In the next step, the algorithm performs spatial common set mining by searching for patterns that have at least one subset of spatial groups $G^{s'}$ where $|G^{s'}| > MinFrac$ and each pair of spatial groups in $G^{s'}$ satisfies Equation 8 and 9.

Algorithm 3 DiSComS

INPUT: Database D, Attributes A, Classes C, Locations L, Level-of-Significance α, Spatial-Groups G^s, MinFrac

```
CANDS=2DHashTable()
CANDP= ∅
for all Location l in L do
    t_l = GetTransactions(S_l, A)
    SCAR_l = ContrainedKingfisher(t_l, C, α)
    for all rule X → G_{c_i,l} in SCAR_l do
        CANDP = CANDP ∪ X
        if CANDS[l][c_i] == ∅
            CANDS[l][c_i] = HashTable()
        CANDS[l][c_i][X] = M(X → G_{c_i,l})
    end for
end for
CSET=∅
for all Candidate Set X in CANDP do
    GrCnt = |G^{s'}; G^{s'} ⊂ G^s, ∀(G_{p,q} ∈ Gs', G_{x,y} ∈ Gs')
    Equation 8 and 9 is TRUE}|
    if GrCnt / |G^s| ≥ MinFrac
        CSET = CSET ∪ X
end for
RETURN CSET
```

5. EXPERIMENTS AND RESULTS

We used our proposed framework on the CNN dataset to find answers to our motivating question: "are there any relationships between industrial air pollutants and ABO cases?" and on a synthetic dataset to assess the accuracy of our techniques. All the experiments were performed in an off the shelf modern personal computer which is equipped with a multi-core CPU and a solid state drive to ensure efficient processing of in-memory database operations.

5.1 CNN Dataset

CNN dataset contains ABO data in 19 local CMAs in Canada. Hence, the set of locations, L is consisted of these 19 CMAs. In this dataset, there are three main ABOs of interest:1) PTB; 2) LBW; and 3) SGA, which together form the set of classes, C, used to label the data. In our experiments, we first extend the point spatial objects as we explained in Section 4. Then we divide that processed dataset based on the CMAs and perform grid transactionization on each of the sub datasets. Following that, we mine co-location patterns in each of those datasets as previously described. This resulted in a set of co-location rules of the form $X \rightarrow ABO$ for each of the CMAs where $ABO \in \{SGA, PTB, LBW\}$ and Xs are some combination of air pollutants. In Fisher's exact test we used a level of significance of 5% when mining for rules.

5.1.1 Co-location Rules

On average, we discovered 495 co-location rules per CMA. The maximum number of co-location rules found for a single CMA was 3371 for Hamilton. No rules were found for the CMA of Victoria. Interestingly, Total Particulate Matter (i.e. TPM - airborne Particulate Matter with an upper size limit of approximately 100 microns) is present in 1849 co-location rules from all the rules from different CMAs associating with one of the three ABOs. Some of the other most common antecedents in the rules were Methanol, Toluene, NO_2, CO, Xylenes, $PM_{2.5}$ (Particulate Matter$\leq$2.5 microns) and PM_{10} (Particulate Matter$\leq$10 microns), respectively. We devised the visualization scheme presented in Figure 2 to navigate through these rules we discovered.

Figure 2: Visualizing Co-location patterns in CMAs. In bubble chart, Y-axis=CMAs and X-axis=rules. Each bubble is a rule and size of a bubble represents the support whereas the color represents the statistical significance (red=high and yellow=low)

5.1.2 Spatial Contrast Sets

Based on the location set L and the class set C we focus on two out of three variations of interesting spatial contrast sets described in Section 4. Those are as follows.

1. Patterns contrasting ABO groups in the same location
2. Patterns contrasting same ABO in different locations

Let us consider the CMA of Vancouver as an example of the first type. In Vancouver, PTB has only two contrast sets out of 66 candidate rules (3%) (i.e. $X \rightarrow PTB$ in Vancouver), which contrast PTB cases from LBW and SGA cases in Vancouver. Those two contrast sets are: {Methanol, CO, 2-Butoxyethanol} and {Methanol, 2-Butoxyethanol}. The significant reduction in patterns using this method can be helpful in efficiently locating specific associations for a particular adverse outcome in a given location. For the LBW cases, we found four contrast sets out of 80 (5%) air pollutant itemsets for LBW in Vancouver. Two of those are: {$PM_{2.5}$, NO_2, CO} and {$PM_{2.5}$, TPM, CO}. On average type 1 spatial contrast mining can result in 18% of the candidates for the spatial group of interest as contrasting.

On the other hand, as an example of the second type, let us consider the CMA of Vancouver and the class PTB again. When contrasted with PTB cases in other 18 CMAs in Canada, we discovered five contrast sets for PTB cases in Vancouver out of 66 candidates (7.5%). Some of them are as follows:{Hexachlorobenzene, HCL}, {Benzene, PM_{10}, CO}, and {Methanol, CO, 2-Butoxyethanol}. These five sets can contrast PTB cases in Vancouver from PTB cases in other CMAs. Similarly, we can detect spatial contrast sets of type 1 and type 2 for any set of spatial groups of interest to locate more specific patterns, effectively narrowing down the hypothesis space.

5.1.3 Spatial Common Sets

Based on the location set L and the class set C we focus on a single type of interesting spatial common sets out of the three described in Section 4. That is to find common sets for a specific ABO in different CMAs. To find such common sets we used a MinFrac threshold of 0.4 (40%) to specify at the minimum in how many spatial groups we would like to see a particular common set exist. For instance, let us consider the task of discovering common sets for PTB cases in different CMAs. We find 24 spatial common sets which are associated with PTB cases in at least 40% of the CMAs. One such spatial common set we discovered is that {Lead (and its compounds)} is associated with PTB in 13 of 19 CMAs (68%) such as Toronto, Vancouver, Ottawa, etc. Other than that, in this 24 sets, interesting spatial common sets such as {PM_{10}, CO}, {Total Particulate Matter, CO}, and {Arsenic} exist. When we compared the spatial common sets for PTB with the sets discovered for LBW and SGA in CMAs in Canada, it is revealed that more than 54% of them are shared with each other.

5.1.4 Empirical and Expert Evaluation

We evaluated our findings by comparing them against the results from the environmental health and pediatrics literature, and with the help of experts in the domain. Most of the studies in the literature discovered monitored urban criteria pollutants like CO (Carbon Monoxide), NO_2 (Nitrogen Dioxide) and Particulate Matter (i.e. $PM_{2.5}$, PM_{10} and Total Particulate Matter) [12, 5] are associated with ABOs such as SGA, PTB and LBW. The majority of the rules we found include these urban pollutants conforming to the existing knowledge. This provides a good indication to the quality of the patterns we discovered. Furthermore, experts and knowledge users collaborating with us from various fields such as environmental health, epidemiology, public health and pediatrics have validated that most of the rules we discovered are interesting and worth further investigation.

5.2 Synthetic Dataset

To assess correctness, we evaluate our methods on synthetic datasets we generated based on previous works in the literature [13]. Similar to the real dataset, this synthetic dataset contains point features that appear in the antecedent part of the co-location rules ("pollutant" features $P_i s$), and ABO features $ABO_i s$ which appear in the consequent part. The study spatial region is a 100x100 unit square. We define buffer radius of each of the point objects as 1 unit. We create three variations of this dataset to evaluate our framework on the three aspects: 1) Detecting statistically significant co-location patterns, 2) Detecting spatial contrast sets, and 3) Detecting spatial common sets.

5.2.1 Co-location Rules

In the first variation of the synthetic dataset, the features P_1 and P_2 have 20 instances each and are co-located with each other. The features P_3 and P_4 have 30 instances each. 20 of P_3 and P_4 are co-located with each other, while remaining 10 instances are uniform randomly placed in the area under study. These associations represent co-located air pollutant items. The ABO feature ABO_1 is co-located with subsets of $\{P_1, P_2\}$ and $\{P_3, P_4\}$, and with 30 out of 40 instances of feature P_5. It does not co-locate with the feature P_6 (30 instances), and negatively associate with the feature P_7 (30 instances), so that no pair of instances ABO_1 and P_7 are co-located with each other. In addition there are 30 ABO_1 cases uniform randomly placed in the area under study. We look for co-location rules of the form $X \rightarrow ABO_1$ where X is a some combination of P_i features. When we mine for co-location rules in the above dataset we obtained 7 co-location rules. All of these rules comply with the positive associative constraints and negative associative constraints imposed and do not include any insignificant rule. Rules such as $(P_1, P_2) \rightarrow ABO_1$, $(P_3, P_4) \rightarrow ABO_1$, $(P_1) \rightarrow ABO_1$, $(P_2) \rightarrow ABO_1$, $(P_3) \rightarrow ABO_1$, $(P_4) \rightarrow ABO_1$, and $(P_5) \rightarrow ABO_1$ are detected.

5.2.2 Spatial Contrast Sets

We slightly modify the above dataset to inject associations for a second ABO variable, ABO_2, as follows. Instead of ABO_1 we co-located ABO_2 with $\{P_3, P_4\}$. We introduced a new feature $P_{5.1}$ and uniform randomly placed 40 of its instances in the area under study. We co-located ABO_2 with 30 instances of $P_{5.1}$. Similar to ABO_1, ABO_2 also does not co-locate with C_6. We introduce another feature $P_{7.1}$ and negatively associate 30 instances of it with ABO_2. With this dataset we contrast two groups ABO_1 and ABO_2 in the same spatial region using our approach. We only detected spatial contrast sets such as (P_1), (P_2) and (P_1, P_2) to contrast ABO_1 from ABO_2. On the other hand when contrasted ABO_2 from ABO_1 we obtained sets, (P_3), (P_4),

and (P_3, P_4) as expected.

5.2.3 Spatial Common Sets

As in spatial contrast sets we used the same dataset in the spatial common sets experiments. The only change we did was modifying all the co-located pairs consisting of P_3 and P_4 instances to pairs consisting of P_1 and P_2 instances. This would result in a co-location relationship among P_1,P_2 and ABO_2. Since ABO_1 already forms a co-location relationship with P_1 and P_2 this should result in several spatial common sets. As expected, using our approach, we find only (P_1), (P_2), and (P_1, P_2) as spatial common sets for groups ABO_1 and ABO_2.

Rediscovering the patterns artificially injected irrespective of their prevalence in the synthetic data shows that the approach we proposed is indeed able to extract the patterns we seek to find.

6. CONCLUSIONS AND FUTURE WORK

We proposed a novel set of co-location pattern mining methods leading to discover two novel types of co-location patterns, spatial contrast and common sets, which allowed to compare and contrast various spatial groups. Our work is motivated by an application problem in the environmental health domain, where the goal is to find interesting co-location patterns between industrial air pollutants and adverse birth outcomes in Canada. We primarily focused on addressing three major challenges: 1) Finding rare but statistically significant co-location rules, 2) Finding discriminant co-location rules, and 3) Finding common co-location rules. Although there are few existing works addressing the first challenge, no significant work has been proposed to address the latter two challenges. Hence, in our work we introduced two new kinds of patterns called *spatial contrast sets* and *spatial common sets*, to address the challenges of finding discriminant spatial patterns and common spatial patterns. We proposed two new algorithms DiSConS and DiSComS to discover those patterns as well. We applied our methods to the ABO dataset from CNN and discovered a number of potential and interesting air pollutant patterns which are either associated with specific spatial groups or significant in many spatial groups. Empirical evidence and expert opinions suggested that the majority of the patterns we discovered are conforming to existing knowledge and the rest provides interesting hypotheses which should be explored further to confirm as new knowledge. Similarly, our framework can be easily applied to find such interesting co-location patterns in many general application problems. We also used a synthetic dataset to further validate our methods. The results confirm that we successfully discovered intended co-location patterns. We are currently working with domain experts in environmental health to further validate the new knowledge we discovered. Our framework also can detect significant negative co-location patterns. This could open up a whole new avenue of research to find interesting spatial contrast and common sets. The proposed framework can also easily be extended in the future to deal with uncertain and temporal datasets.

7. REFERENCES

[1] A. Adilmagambetov, O. R. Zaiane, and A. Osornio-Vargas. Discovering co-location patterns in datasets with extended spatial objects. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 84–96. Springer, 2013.

[2] L. Antonie, O. R. Zaïane, and R. C. Holte. Redundancy reduction: does it help associative classifiers? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 867–874. ACM, 2016.

[3] S. Barua and J. Sander. Sscp: mining statistically significant co-location patterns. In *SSTD*, pages 2–20, 2011.

[4] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[5] M. Brauer, C. Lencar, L. Tamburic, M. Koehoorn, P. Demers, and C. Karr. A cohort study of traffic-related air pollution impacts on birth outcomes. *Environmental Health Perspectives*, 116(5):680, 2008.

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

[7] S. Ha, H. Hu, D. Roussos-Ross, K. Haidong, J. Roth, and X. Xu. The effects of air pollution on adverse birth outcomes. *Environmental research*, 134:198–204, 2014.

[8] W. Hämäläinen. Statapriori: an efficient algorithm for searching statistically significant association rules. *Knowledge and information systems*, 23(3):373–399, 2010.

[9] W. Hämäläinen. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. *Knowledge and information systems*, 32(2):383–414, 2012.

[10] Y. Huang, J. Pei, and H. Xiong. Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, 10(3):239–260, 2006.

[11] M. S. M. Jabbar and O. R. Zaïane. Learning statistically significant contrast sets. In *Canadian Conference on Artificial Intelligence*, pages 237–242. Springer, 2016.

[12] E. Lavigne, A. S. Yasseen, D. M. Stieb, P. Hystad, A. van Donkelaar, R. V. Martin, J. R. Brook, D. L. Crouse, R. T. Burnett, H. Chen, et al. Ambient air pollution and adverse birth outcomes: Differences by maternal comorbidities. *Environmental research*, 148:457–466, 2016.

[13] J. Li, A. Adilmagambetov, M. S. Mohomed Jabbar, O. R. Zaïane, A. Osornio-Vargas, and O. Wine. On discovering co-location patterns in datasets: a case study of pollutants and child cancers. *GeoInformatica*, pages 1–42, 2016.

[14] J. Li, O. R. Zaïane, and A. Osornio-Vargas. Discovering statistically significant co-location rules in datasets with extended spatial objects. In *Data Warehousing and Knowledge Discovery*, pages 124–135. Springer, 2014.

[15] S. Shekhar and Y. Huang. Discovering spatial co-location patterns: A summary of results. In *SSTD*, pages 236–256, 2001.

[16] H. Xiong, S. Shekhar, Y. Huang, V. Kumar, X. Ma, and J. S. Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *SDM*, 2004.