Int J Data Sci Anal
DOI 10.1007/s41060-017-0079-5

REGULAR PAPER

Discovering co-location patterns with aggregated spatial transactions and dependency rules

Mohomed Shazan Mohomed Jabbar[1] · Colin Bellinger[1] · Osmar R. Zaïane[1] · Alvaro Osornio-Vargas[2]

Received: 1 February 2017 / Accepted: 27 October 2017

Abstract Co-location pattern mining focuses on finding associations among spatial features. Existing co-location pattern mining techniques mainly rely on frequency based thresholds which discard the rare patterns and find the noisy patterns. This could be avoided by evaluating co-location patterns based on their statistical significance. Recent studies focused on association rule mining have successfully adopted statistical tests to find significant rules. By transforming spatial data to transaction data, the co-location pattern mining problem can be reduced to an association rule mining problem and such methods can be used to find co-location patterns robustly. A transactionization mechanism has been recently proposed to achieve this. However, this method ignores the effect of general instances, with non-overlapping buffer regions, on the reference instances in their proximity. Addressing this, we propose a novel approach, AGT-Fisher, to robustly transform spatial data to transaction data and use statistically significant dependency rule searching methods to find co-location rules from them. Our work is motivated by an application in environmental health to investigate potential associations between air pollution and adverse birth outcomes in Canada. We used AGT-Fisher to find such associations from real datasets. The discovered co-location patterns were evaluated based on their statistical dependency and the empirical evidence, and results showed that our approach is more robust. Furthermore, we evaluated the resulting patterns to find spatial common and contrast sets, which are two special types of co-location patterns, to compare spatial regions and gain more insights.

Keywords Co-location patterns · Spatial data · Spatial contrast sets · Spatial common sets · Environmental health

✉ Mohomed Shazan Mohomed Jabbar
mohomedj@ualberta.ca; shazanj@gmail.com

Colin Bellinger
cbellinger@ualberta.ca

Osmar R. Zaïane
zaiane@ualberta.ca

Alvaro Osornio-Vargas
osornio@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[2] Department of Pediatrics, University of Alberta, Edmonton, AB, Canada

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 whose instances are often located together in spatial proximity.

Co-location pattern mining has been applied in many diverse disciplines. In particular, this work is a collaboration with researchers at the Canadian Neonatal Network[1] in which we perform data mining in an environmental health problem involving data from 21 cities. Our aim is to help advance the research question: “Do industrial air pollutants have any impact or associations with adverse birth outcomes?” In particular, we propose that our data mining approach can serve to generate hypotheses regarding the relationships between the presence of industrial air pollutants and adverse birth cases. We are particularly interested in the impact of chemical combinations, as these are difficult to study with traditional methods [11]. Environmental health researchers

[1] www.canadianneonatalnetwork.org.

Published online: 09 November 2017

can use these hypotheses in their motivation for their directed research.

There are many studies suggesting that associations between air pollutants and Adverse Birth Outcomes (ABOs) exist [13]. Discovering such associations can be directly converted to a co-location pattern mining problem where the objective is to find co-location rules of the form $X \rightarrow A$, where X is a set of air pollutants (i.e., chemical compounds) and A is an adverse birth outcome [e.g. low birth weight (LBW), small for gestational age (SGA) and preterm birth (PTB)].

Co-location pattern mining (CPM) and classical association rule mining (ARM) tasks have strong similarities. Most of the existing CPM and ARM techniques primarily rely on frequency based prevalence thresholds to find “interesting” patterns [17]. Given a strict prevalence threshold, such approaches could lose many rare but significant patterns [16]. Alternatively, with a lenient threshold, a large number of noisy patterns are likely to be detected. This leads to inefficient methods [12] that, in our domain, have a major impact on the practitioner’s ability to identify valuable hypotheses to explore.

Quantifying the strength of the statistical dependency between the antecedents and the consequent is a good alternative to measuring the significance or the “interestingness” of a rule. In other words, the antecedents and the consequent of a rule are associated with each other, or truly dependent, if the dependency in the observed data is not merely by chance. Statistical significance tests can be used to quantify this notion of true dependency. In the past, a few ARM techniques have been developed to use statistical significance tests to find relevant and rare patterns and also to reduce the number of noisy patterns [8,9].

If a spatial dataset can be effectively transformed into a non-spatial transaction dataset, the co-location pattern mining problem can be reduced to an association rule mining problem [14] and statistically significant association rule mining techniques can be used to find co-location patterns robustly. Once again, this is essential in our domain in order to increase the likelihood of sharing insightful hypotheses with our collaborators. This transformation process is called *Transactionization*. It is worth noting that transactionization enables the application of other ARM methods as well [3,12,20].

The Grid-based Transactionization method (GT) was previously proposed with the aim of exploiting statistically significant association rule mining methods. It addresses various limitations in the traditional transactionization, such as reference-centric and window-centric models [11]. Given a spatial dataset overlaid by a set of square grid cells, transactions are generated based on the features whose spatial instances’ buffer regions overlap a particular point of intersection of two grid cells. However, our application domain provides an excellent example of a limitation that we have identified with the GT method.

Our application involves two types of spatial instances: (a) patients with an adverse birth outcome and (b) air pollutants emitted from industrial facilities. The central feature of our study is the patient since we are interested in the effect of air pollution on the occurrence of adverse birth outcomes. In this domain, situations occur where the patient is exposed to multiple air pollution regions. When air pollution regions themselves are disjoint while still overlapping with the patient’s mobility region, we refer to them as *non-overlapping spatial regions*.

The GT transactionization method does not capture the effect of non-overlapping spatial regions because it generates a unique transaction each time a patient’s mobility region overlaps with the buffer region of an individual air pollutant instance. This ignores the fact that some of the patient-pollutant combinations involve the same patient. The transactions should reflect all of the chemical exposures an individual patient is subjected to.

An example of the above scenario is given in Fig. 1. Here, A is a patient with an adverse birth case, and B and C are air pollutant emitting facilities. The circles represent the spread of the air pollutants and the mobility region of the patient. The GT algorithm derives a transaction indicating that patient A is affected by both B and C in Fig. 1a, but it cannot derive such a transaction in Fig. 1b. However, the patient in Fig. 1b is clearly affected by both B and C. This inhibits the discovery of some associations involving combinations of air pollutants.

We propose Aggregated Grid Transactionization (AGT) to address the limitations of grid transactionization. We use this in conjunction with a Fisher’s test-based dependency rule search technique to find statistically significant co-location patterns. This combination forms a novel co-location pattern mining method, namely AGT-Fisher, which addresses many limitations of previous co-location pattern mining approaches.

Finally, with rich datasets that contain data from multiple spatial regions, it is useful to discover patterns that: (a) uniquely characterize a specific spatial region and contrast it with others, and (b) are common in many spatial regions. For instance, in our application, a valid research question is: “Are there any specific combinations of industrial air pollutants that are more associated to low birth weight in Greater Toronto Area than other Canadian cities?” In one of our previous studies, we proposed techniques to answer these questions. We refer to them as spatial contrast sets and spatial common sets [1]. This methodology can be applied within AGT-Fisher to discover spatial contrast sets and spatial common sets, which enables us to gain unique insight into the associations between chemicals, ABOs, and regions.

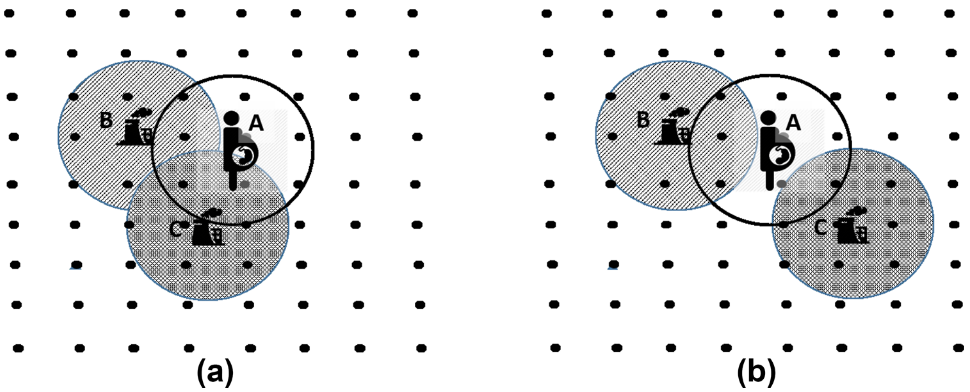

Fig. 1 The effect of overlapping versus non-overlapping spread regions when a central feature is given (A is a patient. B and C are air pollutant emitting facilities. The circles represents the spread of the air pollution and the mobility region of the patient): **a** air pollution spread regions overlap with one another and with the mobility region of the patient; **b** air pollution spread regions do not overlap with one another but overlap with the mobility region of the patient separately

We apply AGT-Fisher on adverse birth datasets collected by the Canadian Neonatal Network (CNN) from 21 major Canadian cities to effectively find co-location patterns of industrial air pollutants and adverse birth outcomes (ABOs) in Canada addressing our motivating challenge.

To summarize, our contributions in the current work are as follows:

- We identify a limitation in the existing transactionalization algorithm. Specifically, it does not capture the potential effect caused by instances of multiple general features (e.g. air pollutants) with non-overlapping buffer regions on instances of a reference feature (e.g. ABO cases).
- We propose Aggregated Grid Transactionization (AGT), to address this problem.
- Experimentally, we show that AGT with Fisher's exact test (AGT-Fisher) can find statistically significant co-location patterns more effectively with a higher lift.
- We apply AGT-Fisher to address a practical environmental health issue involving adverse birth outcomes and industrial air pollutants. Within this domain, our results constitute a resource for hypothesis generation and can lead to the discovery of new information. A particular advantage is our ability to generate hypotheses involving multiple chemicals.
- Finally, as an extension to our previous works [1], we demonstrate how to use AGT-Fisher to find two special cases of co-location patterns called spatial contrast and common sets.

The rest of the article is organized as follows. In Sect. 2, we discuss related work in the literature and in Sect. 3 we provide definitions and outline the preliminary concepts required to design our methods. We discuss our proposed AGT-Fisher method and its application in finding spatial contrast and common sets in Sect. 4. In Sect. 5, we present our experimental results and evaluation with CNN datasets. Finally, we conclude in Sect. 6.

2 Related work

Early spatial statistics approaches deploy techniques such as cross K-functions with Monte-Carlo simulations [26], mean nearest-neighbor distance, and spatial regression models [28] to evaluate and find co-location patterns between two features. Disadvantages of these approaches are their high computational requirements and the difficulty in applying them to patterns consisting of more than two spatial features.

On the other hand, traditional co-location rule mining techniques are based on the neighborhood relations and participation indices [18]. 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 neighborhood relation. To find rare patterns, previous studies have introduced a new measure called the max participation ratio $maxPR$. If the $maxPR$ instances of at least one of the features in the given pattern form a neighborhood relation with the instances of all the other features in the same pattern, then that co-location pattern is considered prevalent [15]. Most of these techniques depend on user-defined thresholds for the interestingness measure and detect a large number of noisy patterns when the threshold is too low. Moreover, they miss rare patterns if the threshold is high.

In contrast to the traditional co-location pattern mining approaches which rely on frequency based thresholds, more recently, transactionization-based techniques have been introduced to find co-location patterns more robustly by using statistically significant association rule analysis methods. Grid-based Transactionization is one such transactionization technique introduced to transform a spatial dataset into a transaction dataset, which addresses the limitations in early models such as reference-centric and windows-centric methods [3,11,12]. This approach is used with statistical significance tests to find more relevant co-location patterns in previous studies. One such approach exhaustively generates all possible patterns up to a certain length and computes the empirical p value based on the support-confidence value in the observed dataset against the simulated datasets satisfying the null hypothesis [3]. Methods for the discovery of co-location patterns based on computing empirical p values in simulated datasets have previously been used with non-transactionized datasets where the participation index (PI) is used as the preferred prevalence measure [19]. However, an Apriori-like search technique is not possible in any of the above cases due to the fact that the statistical significance is not a monotonic property. Once R number of simulated datasets are generated using a randomized test, the empirical p value is $\frac{R^{\geq \text{Observed\_confidence}}+1}{R+1}$. Due to the exponential growth of the computational complexity of this approach, the technique is not scalable beyond a certain pattern length.

Alternatively, in the past, statistically significant dependency searching techniques have been developed to find more relevant and statistically sound association rules. Due to various issues including redundancy and inefficiency in early methods, better approaches were suggested in recent years. StatApriori [8] and its successor Kingfisher [9] are two such searching methods proposed to find statistically significant dependency rules. StatApriori uses Z-score to approximate the upper bound for the p value and then uses Apriori-based strategies to search the solution space. However, the Z-score overestimates the p value, which leads to issues in redundancy, affecting the quality of the discovered rules [9]. In order to avoid these issues, the Kingfisher search method was proposed. I was proven to be more robust in finding non-redundant statistically significant dependency rules [9]. Given an association rule $X \rightarrow A$, Kingfisher estimates the statistical significance of the dependency between X and A using Fisher's exact test. If X and A are truly independent, the probability of an observed or stronger dependency occurring by chance, p_F (i.e., p value), can be computed using a cumulative hypergeometric distribution. Kingfisher uses enumeration trees, efficient search mechanisms, and pruning heuristics to efficiently search the solution space to find significant rules. Besides Fisher's exact test, the χ^2-test can also be used with Kingfisher to find the statistically significant dependencies. However, Fisher's exact test is empirically proven to be more effective, efficient and scalable compared to the χ^2-test [9].

CMCStatApriori was proposed to address scalability and efficiency (e.g. fixed size co-location patterns) in the transactionization-based co-location discovery methods that use the empirical p value to measure statistical significance. It uses a constrained version of the Z-score-based statistically significant association rule searching technique, StatApriori [8], on a transactionized spatial dataset [12]. However, as we previously discussed, the Kingfisher algorithm is a more efficient and effective approach to finding statistically significant dependency rules using Fisher's exact test [9]. Based on this, in our work, we introduce a novel grid-based transactionization technique to use with Fisher's exact test-based statistical dependency rules to more robustly find statistically significant co-location patterns.

To compare and contrast categories of data, a type of association patterns called contrast sets were first introduced through the STUCCO [5] algorithm. STUCCO can find association patterns which can contrast instances from a particular group to instances from other groups. Most of the existing contrast set mining techniques like STUCCO depend on two threshold values called support and confidence, and are prone to the limitations imposed by them. Hence, as an alternative, in a previous work, we proposed to use statistically significant association rules to mine contrast sets [2]. This can be applied in the context of spatial data to compare and contrast spatial groups as well. Recently, we extended our previous work to find such contrast sets in spatial data [1]. In the same work, we also proposed a new algorithm to find a novel type of pattern called spatial common sets to find patterns, which are equally (i.e., commonly) strong in multiple spatial groups. These common sets and contrast sets for spatial data help gain new insights and knowledge which was not accessible previously.

3 Preliminaries

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 (Table 1 introduces the basic notation used throughout this work). Let $A = \{A_1, A_2, \ldots, A_m\}$ 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*. A regular dataset with features and corresponding values can be discretized and/or binarized depending on the application to help represent each data instance as a set of items. Discretization involves converting a continuous attribute to a discrete or binary attribute. Some applications first convert continuous attributes to discrete attributes by defining multiple bins or intervals within the continuous range, and assigning the corresponding attribute value to one of the bins. Such a discretized variable can be further processed to binarize it. In such a case, each {original

Table 1 Basic notations

Notation	Meaning
A, X or Y	An itemset or a set of discrete attribute-value pairs where A_{i} is an item or an attribute-value pair
F	Set of features where $f_{\mathrm{i}} \in F$ and $v_{f_{\mathrm{i}}}$ is the value of the feature f_{i}
C	Set of classes where c_i is a member class
E^s	A spatial transaction or an itemset
D^s	Transactionized spatial dataset where $E^s \in D^s$
support(X) or m(X)	Frequency of transactions in D^s in which X itemset occur
conf$(X \rightarrow Y)$	Confidence of $X \rightarrow Y$ (i.e., support(Y) / support $(X \rightarrow Y)$)
N_{XA}	A random variable of the absolute frequency of X and A co-occurrence
p_{F}	Fisher's p value
M$(X \rightarrow A)$	Goodness measures for the rule (e.g. Fisher's p value)
$G_{x,y}$	Is a group where $G_{x,y} = \{C_x, L_y\}$; C_x is the class membership and L_y is the location of the group

attribute, discretized value} pair acts as a new binary/boolean feature for the dataset. In the context of association rule mining, such binary features can be considered as items where, if the binary feature is "True", the item exists and if it is "False" the item does not exist. Hence, in such an itemized dataset, a transaction E can be defined as a vector consisting of feature-value pairs or items $\{A_i, A_j, \ldots, A_m\} \subset A$. Given these, an association rule can be defined as in Definition 1.

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

Confidence conf in $X \rightarrow Y$ is the percentage of data instances in D containing X that also contain Y (i.e., $P(Y|X)$). Support *sup* for $X \rightarrow Y$ is the percentage of data instances in D containing $X \cup Y$. Traditional algorithms discover strong association rules by verifying that their *sup* and *conf* exceed some user-defined thresholds. Classification association rules (CAR) are a special case of general association rules [6]. Given a set of class labels $C = \{c_1, c_2, \ldots, c_q\}$ 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 dataset S, each instance $s \in S$ can be defined as a vector, $s =$ [long., lat., feature$_i$, other contextual data, …], consisting of longitude, latitude, spatial feature ID (e.g. $Pollutant_1$, $ABO_1, \ldots$) and other contextual data such as climate information (e.g. average wind speed and direction). This dataset can be transformed into a transaction dataset D^s using a suitable transactionization algorithm. For instance, a reference-centric model can be used as such a transactionization method [25] (other methods include window-centric models, spatial join-based approaches to co-location patterns, etc. [11]). This reference-centric model creates a transaction around a reference feature specified by the user. Each set of spatial features which forms a neighborhood relationship with an instance of the reference feature is considered as a transaction. The neighborhood relationship or spatial proximity between the features can be defined with a user-defined distance threshold. In such a scenario, $E^s \in D^s$ is a vector representing a single transaction, where $E^s = [ID, feature_1 \in \{0, 1\}, feature_2 \in \{0, 1\}, \ldots]$. E^s defines a neighborhood relationship based on the s spatial instances. Furthermore, A^s represents a set of all the possible spatial items [e.g. $(features_1, 0)$] in D^s as well. A transaction represents a set of spatial features whose instances from S are in the close spatial proximity. Under these conditions, an association rule mining technique can be applied to D^s as any other transaction database and find association rules which would be interpreted as co-location rules, defined in Definition 2.

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

Given a co-location rule, $X \rightarrow Y$, the dependency between X and Y is traditionally measured by using an empirical p value. When the prevalence measure or the confidence value of a pattern in a certain number of simulated datasets is larger than the observed confidence, then empirical p value suggests that pattern is statistically not significant. This cut-off number is determined by the level of significance. The simulated datasets are generated to comply with the null hypothesis which states that there is no dependency between the instances containing antecedent features with the instances containing the consequent features. However, other statistical significance tests such as Fisher's exact test and the χ^2 test are more flexible and extensively used in recent literature to measure the statistical significance of a rule.

Contrast sets are another class of associative patterns which are used to characterize a particular class and contrast it with the others. Contrast sets can be defined as shown 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 a set X in a class association rule $X \implies c_i$ meets STUCCO deviation conditions [5] as in Eqs. 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 Eq. 1 imposes that the support of a contrast set is significantly different across various groups. The second condition in Eq. 2 imposes that the difference in support of a contrast set across different groups is sufficiently large.

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

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

4 AGT-Fisher to find co-location patterns

We propose an improved co-location pattern mining approach, *AGT-Fisher*, based on a new grid transactionization method and Fisher's exact test. AGT-Fisher transforms a spatial dataset into a transaction dataset more effectively than prior approaches and uses statistically significant dependency rule searching techniques to find "interesting" and statistically sound co-location patterns. We further elaborate how this method can be used to find spatial contrast sets and spatial common sets to compare various spatial groups.

4.1 AGT-Fisher method

The algorithmic process of AGT-Fisher consists of two major steps: (1) Transactionizing the spatial dataset with AGT (Aggregate Grid Transactionization); (2) mining for statistically significant association rules with Fisher's exact test for statistical significance. Next we explain how this process can be accomplished.

4.1.1 AGT: aggregated grid transactionization

As we previously discussed, although GT solves the limitations posed by earlier co-location pattern mining methods, it has certain limitations when a reference or central feature is given. Specifically, it does not take into account the "combined effect" caused by instances of *non-overlapping* general features within the proximity of the spatial instance of the given reference feature. In the case of our application, this could avoid identifying how the "combination" of chemicals/air pollutants can affect the maternal health, leading to adverse birth outcomes.

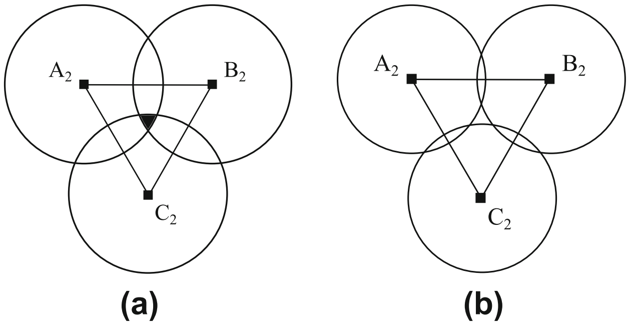

Fig. 2 Intersection of neighboring extended spatial objects: **a** an intersection of buffer regions of feature A, B, and C exists; **b** an intersection of buffer regions of feature A, B, and C does not exist [11]

To elaborate this further, consider the example scenario given in Fig. 2. In this example, there are three spatial features A, B and C. A_2, B_2 and C_2 are spatial instances of those features. Static circular buffer regions surrounding the spatial instances represent the area affected by them. The scenario given in Fig. 2a represents an occasion where the buffer regions of all three instances intersect. However, in the scenario presented in Fig. 2b, there is no intersection among the buffer regions of all three instances. Assume that feature C represents a patient (i.e., reference feature) and both features A and B represent some adverse environmental conditions (i.e., general features). Although the instances of C are exposed to adverse conditions B and A in both the scenarios, the original grid transactionization GT [11] is capable of capturing this relationship only when there are points in the overlaid grid which are overlapped by all three buffer regions, such as in Fig. 2a. Since there are no common overlaps in the scenario presented in Fig. 2b, GT is unable to find transactions which have all A, B, and C features. This leads to discovering less co-location patterns with multiple antecedents (i.e., combinations of air pollutants). In order to address this issue, we propose that when a reference feature such as C is given, the scenario given in Fig. 2b should produce valid transactions consisting of all A, B and C features. To achieve this, we propose an Aggregated Grid Transactionization (AGT) method.

Our proposed aggregated grid transactionization procedure is outlined in Algorithm 1. Given a spatial dataset *S*, Algorithm 1 initially generates a set of 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*, and taking the points at the intersection of the grid lines (line 2). Each such 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, then Algorithm 1 defines

Algorithm 1 ***GetAGTransactions(S)***

```
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: if Reference Feature Exists then
10:    for all set $T_g \in \{T$ Grouped By Reference Feature ID$\}$ do
11:        CombEffS = $min|$ set $T_{gf} \in \{T_g$ Grouped By General Feature
           ID(s)$\}|$
12:        CombEffT = CombEffS $\times$ Aggregate $T_g$
13:        for all set $T_{gf} \in \{T_g$ Grouped By General Feature ID(s)$\}$ do
14:            RemSet = CombEffS $\times$ TOP $T_{gf}$
15:            $T = T \setminus$ RemSet
16:        end for
17:        $T = T \cup$ CombEffT
18:    end for
19: end if
20: return $T$
```

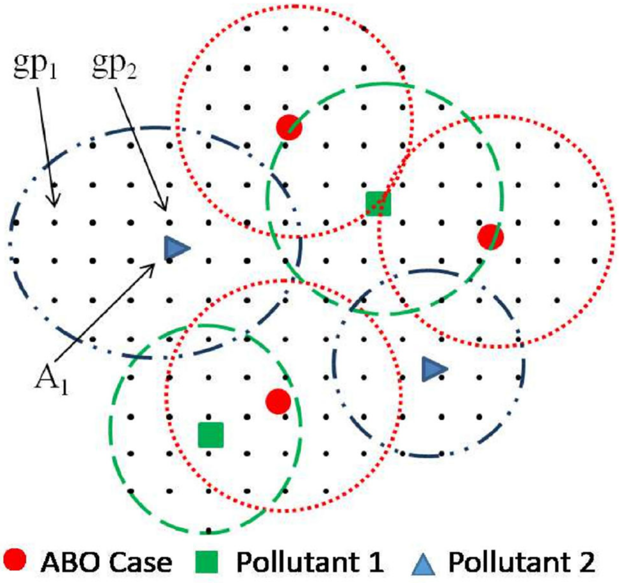

Fig. 3 Grid Transactionization: A grid with equal size cells is overlaid on a sample spatial dataset with point feature instances and their buffers. Three types of spatial instances are shown i.e., ABO cases, Pollutant 1 and Pollutant 2. Buffer regions representing the area of impact of spatial instances are drawn around them. Grid points, resulting from the intersection of grid lines, which intersect with those buffers are used to derive the associated spatial features of the corresponding instances and create transactions [11]

buffer zones around spatial objects in S. Defining such buffer zones is specific from problem to problem. We show how to define such buffers in the case of our motivating application in Sect. 5. In the dataset of our motivating application, we have two types of spatial objects: (1) ABO cases, and (2) Chemical emission points. The buffers are defined accordingly. In the next step of the algorithm, the constructed grid is imposed over the dataset S. Figure 3 illustrates an example dataset with buffers around spatial point instances where a grid is laid over it. Similarly, buffers can also be created around linear and polygonal spatial objects. In a two-dimensional space, points in the grid can represent a square geographical area with a length of a regular grid cell. Due to the spheroid shape of the Earth, a grid used for real-world applications becomes irregular. However, with a careful choice of a grid granularity, this fact should not considerably affect the accuracy of the results.

A point of two crossing grid lines 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 as shown in the Algorithm 1 (see lines 5–8).

In our experimental setup, we do not face the more challenging task of mapping continuous variables to multiple bins, instead we deal with discretizing variables into binary values. In particular, we consider each spatial object and their buffers as binary features associated to the grid point with which they overlap. To elaborate further, each grid point can only "sense" whether a particular spatial feature exists or not (i.e., True or False) in its current location. It does not quantify the size of the impact of that feature. This formalization of the binarization is based on the target domain. Moreover, it is required in the next step of mining statistically significant co-location patterns with Fisher's exact test.

Although, in the past, clustering-based co-location pattern mining approaches have been proposed which can handle continuous data [21], to our knowledge, no transactionization-based rules mining approaches have been proposed to find statistically significant co-location patterns in continuous data. However, recently, research efforts have been directed toward finding statistically significant itemsets with G-Test [22]. If proven effective, the modularity of our AGT-Fisher method can incorporate such novel methods in future versions of our approach to directly accommodate continuous data. Furthermore, the current version of the grid transactionization could also inherently help toward quantifying the impact of a spatial feature via capturing the spread of it. In other words, the larger the buffer region or spread of a particular spatial feature, the more grid points that would turn out as evidence.

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. In our previous work [11], we have conducted extensive experiments on evaluating two aspects of choosing a grid granularity level: computational complexity (i.e., the number of transactions) and effectiveness (i.e., the number of discovered quality rules). In our experiments, we used different grid

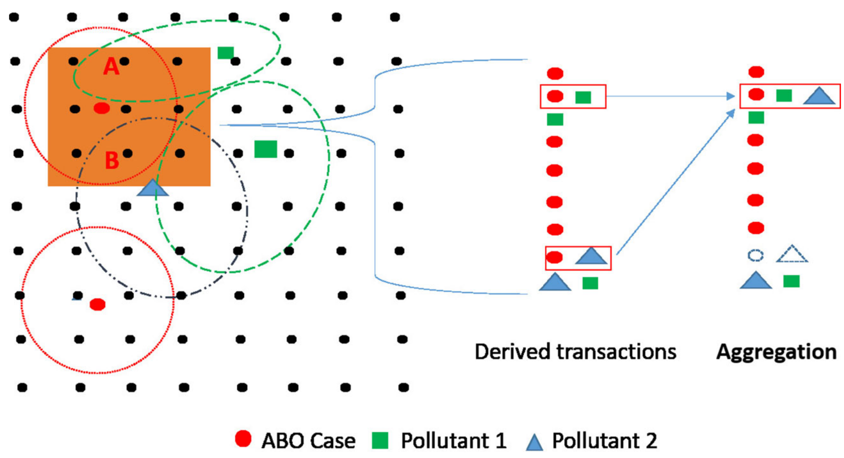

Fig. 4 The transaction aggregation process within the AGT method: Once the buffers are extended and the transactions are derived as previously explained, transactions with common spatial instances of the reference feature (i.e., ABO cases) are identified. Those identified transactions are aggregated by taking the union to derive the aggregated transactions which represent the combined effect

granularity levels such as 0.5, 1 and 2 km. As expected, lower granularity levels resulted in a higher number of transactions. To find more significant rules, having more transactions could be helpful. However, this can affect the computational complexity of the algorithm. Hence, selecting a moderate granularity level was important, especially when dealing with large spatial regions. Toward this end, we discovered that rules discovered with both granularity levels 0.5 and 1 km have the highest average p value, indicating that rules detected by them are the most statistically significant ones. Further analysis [11] revealed that rules detected under 1 km grid granularity were able to capture most of the rules with lower p value from both the other rule sets detected under 2 and 0.5 km grid granularity. This emphasizes that 1 km could be a better grid granularity choice because it captures most of the rules with lower p values while maintaining an efficient program execution. Our previous work has used the same dataset for air pollution data. Although the interested health variables are different in both previous and current studies, we used the same radius to define the patient mobility region as well. Given that the majority of the parameter configurations are similar in both of the applications, we also define that the grid granularity of our current application as 1 km.

Previous GT approach only consists of steps 1–8 in Algorithm 1. If a reference feature is given (e.g. adverse birth outcome), in the next part of the algorithm, our AGT method aggregates the set of obtained transactions to derive transactions representing the combined effect we previously explained. To perform that, initially, all the transactions in T are grouped by the distinct instance IDs of the reference feature, and the algorithm iterates over the resulting set of transaction groups (i.e., T_g) to aggregate them (see lines 10–18). In each iteration, T_g is again grouped by the general feature IDs other than the reference feature (e.g. chemical$_1$, chemical$_2$,…) and the minimum size of such a group T_{gf} is obtained. This is the maximum number of transactions (i.e., *CombEffS*; see line 11 in Algorithm 1) which can be aggregated to represent the combined effect of all the general non-overlapping features which only overlap with the same reference feature instance. All the features in T_g can be combined to obtain a single transaction representing the combined effect. This transaction is added *CombEffS* times to the final transaction set. Finally, *CombEffS* number of transactions from each of the group T_{gf} are removed from the final transaction set to be returned (see lines 13–16 of Algorithm 1).

A visual example of the aggregated grid transactionization process is depicted in Fig. 4. Three spatial object types are shown in Fig. 4: (1) ABO cases, (2) Pollutant 1, and (3) Pollutant 2. First, Algorithm 1 defines buffer regions around the spatial objects and overlays a grid covering them. Second, from each grid point, features of spatial instances whose buffer regions overlap with that particular point are derived as transactions. For instance, in this example, transactions which only contain the spatial feature ABO, transactions with the feature ABO and the Pollutant 1 or 2, and transactions with Pollutant 1 and 2 are derived. Third, in the aggregation step of the algorithm, it is revealed that two transactions have a common ABO case (i.e., point A and B share the same ABO source). Hence, we aggregate those two transactions into one by taking the union to show the combined effect of both the pollutants to the ABO case.

4.1.2 Fisher's test to find dependency rules

The usage of traditional association rule mining techniques, which are primarily based on the support and the confidence framework, to identify co-location rules, imposes some major limitations as we previously discussed. On the other hand, 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 rule mining frameworks, it has been proposed to adopt an association rule mining approach based on statistical significance tests. Given a rule, $X \rightarrow A$,

Table 2 2 × 2 contingency table for the X and A variables in rule $X \rightarrow A$

	A	¬A	Total
X	m (XA)	m (X¬A)	m (X)
¬X	m (¬XA)	m (¬X¬A)	m (¬X)
Total	m (A)	m (¬A)	$\mathrm{m}(X) + \mathrm{m}(\neg X) = n$

such tests are designed to test the dependency between X and A. The null hypothesis in those tests 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 of an observed or stronger dependency occurring 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.

Fisher's exact test is a statistical significance test which can assess whether two categorical variables are non-randomly dependent on each other or not. For instance, consider the two categorical variables X and A in the rule $X \rightarrow A$. X determines whether all the items in the antecedent of the given rule are present in a given transaction, whereas A determines whether or not the consequent of the rule is present in a given transaction. This can be represented in a 2 × 2 contingency table as in Table 2.

Given this table, the hypergeometric probability of obtaining this particular arrangement of values in the observed data when the null hypothesis is that A and $\neg A$ are equally likely to be co-occur with X is given in Eq. 3.

$$p = \frac{\binom{\mathrm{m}(X)}{\mathrm{m}(XA)}\binom{\mathrm{m}(\neg X)}{\mathrm{m}(\neg X \neg A)}}{\binom{n}{\mathrm{m}(A)}} \qquad (3)$$

Let N_{XA} be a random variable representing the absolute frequency of X and A occurring together. The dependency between X and A is stronger than observed in a given dataset if $N_{XA} > \mathrm{m}(XA)$, where $\mathrm{m}(XA)$ is the frequency of event XA in the observed data. Fisher's p value can be computed by accumulating all the probabilities of possible datasets containing at least m(XA) data instances confirming the co-occurrence of XA event. Hence, the Fisher's p value can be computed using the cumulative hypergeometric distribution shown in Eq. 4.

$$p_F(X \rightarrow A) = \sum_{i=0}^{J} \frac{\binom{\mathrm{m}(X)}{\mathrm{m}(XA)+i}\binom{\mathrm{m}(\neg X)}{\mathrm{m}(\neg X \neg A)+i}}{\binom{n}{\mathrm{m}(A)+i}} \qquad (4)$$

where $J = \min\{\mathrm{m}(X\neg A), \mathrm{m}(\neg XA)\}$, n is the number of total transactions, and $m(.)$ computes the frequency of transactions containing the given items. Given a level of significance (e.g. 0.05) this p value p_F could be used to determine whether a given rule is statistically significant or not. If the computed p value is lower than the level of significance the null hypothesis can be rejected and we can conclude that the dependency in the rule $X \rightarrow A$ is statistically significant. 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 $\mathrm{M}(Y \rightarrow A)$ is equally good or better than $\mathrm{M}(X \rightarrow A)$. Here, M is a goodness measure and in our specific case, can be considered as the Fisher's p value.

While Fisher's exact test can be applied directly to discover statistically significant dependency rules in transactionized spatial datasets, its application, like other statistical methods, is severely limited by the complexity of brute-force computations on large datasets.

To address this, the Kingfisher algorithm [9] implements efficient branch and bound search mechanisms on an enumeration tree to detect non-redundant and statistically significant association rules. In order to render it more efficient, the branch-and-bound search is supplemented with several pruning properties that significantly expedite the rule discovery process.

The Kingfisher algorithm is independent of the used goodness measure. However, in the original paper, the authors consider the application of either Fisher's exact test or χ^2. The study finds that using Fisher's exact test rather than χ^2 produces more reliable rules and leads to a faster search.

Following is a high-level description of how Kingfisher operates. More details on algorithms, pruning conditions and lower bounds can be found in [9].

- Prune out all the insignificant individual items / attribute-value pairs
- Order the rest in ascending order by frequency and add them to an enumeration tree. Using lower bounds for Fisher's p value (as proposed [9]) determine possible consequents of the rules where antecedents consist of the significant attributes added earlier to the enumeration tree.
- Expand attribute sets in the enumeration tree as long as new non-redundant and significant rules, as defined in [9], are found.
    - Create l-item-sets from (l-1)item-sets.
    - For each l-item-set, X, initialize possible consequences in node X in the enumeration tree, given possible consequences of it's parent nodes Y's, where $X = YA_m$; $A_m \in A$. Consequent A_j is possible in node X only if it's possible in all parent nodes'. Initialize the optimal value for the goodness measure M (e.g. Fisher's p value), for each $A_j \in X$, using the values from the parent node.

- Use lower bounds provided in [9] to determine any rule $XQ \rightarrow A_j$ can be non-redundant and significant.
- If any of the above tested rules are possible, calculate $M(X \rightarrow A_j)$ (i.e., Fisher's p value). If it is sufficiently good (among the best K rules and better than more general rules with consequent A_j), add it to the result rule set.
- If minimal rules were found mark all the redundant consequences as impossible.
- Use the Lapis Philosophorum principle to propagate consequence information to parent nodes.

Finally, it is worth noting the recent update to the Kingfisher method [27]. The authors propose a means of approximating Fishers exact test that can be more efficient. This new work deals with the cost of calculating Fishers exact p value by offering a bounded trade-off between speed and accuracy. This approximated test is also applicable in our framework.

We use a constrained version of this implementation in the second phase of our AGT-Fisher to successfully detect non-redundant and statistically significant co-location rules. We constrained the Kingfisher algorithm to only produce co-location rules of the form $X \rightarrow A$, where A is one of the desired adverse birth outcomes or groups according to our motivating application problem. Further information regarding the implementation of the search strategies, proofs, and mechanisms of the Kingfisher algorithm can be found in [9].

4.2 Discovering spatial contrast and common sets

Some of the statistically significant co-location rules we detected using the AGT-Fisher 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. Preterm Birth (PTB) cases in Vancouver, Low Birth Weight (LBW) cases in Edmonton, and Small for Gestational Age (SGA) cases in Hamilton can be considered as some of the spatial groups from our motivating application. The first type of rules are 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 are 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. To this end, we analyze the co-location rules detected previously, and 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.

We originally outlined the aforementioned problem and proposed algorithms to mine spatial contrast and common sets in [1] using a GT-based co-location pattern mining method. In this work, we apply those proposed algorithms with AGT-Fisher to discover statistically significant spatial contrast and common sets.

4.2.1 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., $A_i = V_{i_j}, \ldots, A_k = V_{k_l}$ 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 $G_{11}, \ldots, G_{1,p}, \ldots, G_{q,1}, \ldots, G_{q,p}$, 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 that 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 Eqs. 5 and 6 hold $\forall G_{p,q} \in G^s \setminus \{G_{x,y}\}$.

$$p_F(X \rightarrow G_{x,y}) \leq p_F(X \rightarrow G_{p,q}) \tag{5}$$

$$\max_{p,q} |\text{support}(X, G_{x,y}) - \text{support}(X, G_{p,q})| \geq \text{min\_dev} \tag{6}$$

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 $\forall y = q$, we can contrast data which belongs to the same spatial region but has different classes.
2. If we fix $\forall x = p$, we can contrast data which has the same class but belongs to different spatial regions.
3. $\forall x$ and $\forall y$, we can contrast data which belongs 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 lines 2–9). Then, for each group, it searches for contrast sets by imposing the conditions presented in the Eqs. 5 and 6 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 $\alpha$, Spatial-Groups $G^s$
1:  CANDS=2DHashTable()
2:  for all Location $l$ in $L$ do
3:    $t_l$ = GetTransactions($S_l$, $A$)
4:    $SCAR_l$ = AGT-Fisher($t_l$, C, $\alpha$)
5:    for all rule $X \rightarrow G_{c_i,l}$ in $SCAR_l$ do
6:      if $CANDS[l][c_i] == \emptyset$
7:         $CANDS[l][c_i]$ = HashTable()
8:      CANDS[l][$c_i$][X] = $M(X \rightarrow G_{c_i,l})$
9:    end for
10: end for
11: CSET=2DHashTable()
12: for all $G_{x,y}$ in $G^s$ do
13:   $CSET[L_y][C_x] = [\emptyset]$
14:   for all X in $CANDS[L_y][C_X]$.keys() do
15:     if $\forall\, G_{p,q} \in G^s \setminus \{G_{x,y}\}$ Equation 5 and 6 is TRUE
16:        $CSET[L_y][C_x].append(X)$
17:   end for
18: end for
    RETURN $CSET$
```

4.2.2 Spatial common sets

Other than spatial contrast sets, which are helpful in contrasting a particular spatial group from the others, sets that can characterize or represent a set of similar spatial groups are of equal interest. 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, $max - dev$, 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 Eqs. 7 and 8 can be satisfied and the $|G^{s'}| > MinFrac$ threshold.

$$p_F(X \rightarrow G_{x,y}) - p_F(X \rightarrow G_{p,q}) \leq max - pF - diff \tag{7}$$

$$|\,\text{support}(X, G_{x,y}) - \text{support}(X, G_{p,q})| \leq max - dev \tag{8}$$

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

Our proposed algorithm DiSComS to mine such spatial common 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 $\text{G}^{s'}$ where $|G^{s'}| > MinFrac$ and each pair of spatial groups in $\text{G}^{s'}$ satisfies Eqs. 7 and 8.

5 Experiments

We conducted experiments in order to find spatial association patterns between air pollutants and adverse birth cases with our proposed AGT-Fisher method on 19 real spatial datasets

Algorithm 3 DiSComS

```
INPUT: Database D, Attributes A, Classes C, Locations L, Level-of-Significance $\alpha$, Spatial-Groups $G^s$, MinFrac
CANDS=2DHashTable()
CANDP= $\emptyset$
for all Location $l$ in $L$ do
    $t_l$ = GetTransactions($S_l$, $A$)
    $SCAR_l$ = AGT-Fisher($t_l$, C, $\alpha$)
    for all rule $X \rightarrow G_{c_i,l}$ in $SCAR_l$ do
        CANDP = CANDP $\cup$ $X$
        if $CANDS[l][c_i] == \emptyset$
            $CANDS[l][c_i]$ = HashTable()
        CANDS[l][$c_i$][X] = $M(X \rightarrow G_{c_i,l})$
    end for
end for
CSET=$\emptyset$
for all Candidate Set $X$ in $CANDP$ do
    GrCnt = $|G^{s'}; G^{s'} \subset G^s, \forall(G_{p,q} \in Gs', G_{x,y} \in Gs')$ Equation 7 and 8 is TRUE}|
    if $\frac{GrCnt}{|G^s|} \geq MinFrac$
        CSET = CSET $\cup$ X
end for
RETURN $CSET$
```

from Canada. We used publicly available air pollution and climate datasets from Canadian agencies with CNN datasets to derive these 19 datasets. These association patterns could help environmental health scientists and pediatricians to find answers for our motivating application problem. We also used the patterns found by AGT-Fisher in the DiSConS and DiSComS algorithms to find spatial contrast and common sets. This extends the insights gained from the classical co-location patterns in our application problem, thus providing knowledge to users and practitioners.

To evaluate the effectiveness of our proposed aggregated grid transactionization method, we compared the results obtained with AGT-Fisher to the results obtained when the non-aggregated grid transactionization mechanism (i.e., GT) [11] is used with Fisher's test-based dependency rule search (i.e., GT-Fisher). We evaluated the effectiveness of using Fisher's test-based dependency rules by comparing the results obtained with AGT-Fisher to the results obtained when AGT is used with χ^2-test (AGT-Chi2) instead of Fisher's test to find dependency rules. This comparison allows evaluating the robustness of using Fisher's test-based dependency rules. Although in the literature Z-score-based techniques have been used to find dependency rules to discover co-location patterns [12], latest empirical and theoretical studies have proved that Z-score can overestimate the p value, and the rules obtained by it could be redundant [8,9]. These studies suggest that Fisher's and χ^2 tests could be more robust than other tests to find statistically significant rules. Hence, we focus on these two tests in our experiments. In our evaluation, we mainly focused on evaluating the proposed AGT-Fisher method. Further experiments, evaluations, and discussions on spatial contrast and common sets can be found in our previous works [1,2].

5.1 Data and preprocessing

We carried out experiments on 19 real spatial datasets coming from 19 census metropolitan areas (CMAs) in Canada, to evaluate our approach while answering our motivating research question: "what are the relationships between air pollutants released by industries and adverse birth outcome in Canada?" These datasets were collected by the Canadian Neonatal Network and are about babies admitted to Neonatal Intensive Care Units (NICUs) across 21 major 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. 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) [7] of Canada. More specifically, we chose industrial facilities within a 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 81 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 point 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, weather conditions (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 previous work [11], we defined the radius of these buffers as the natural logarithm

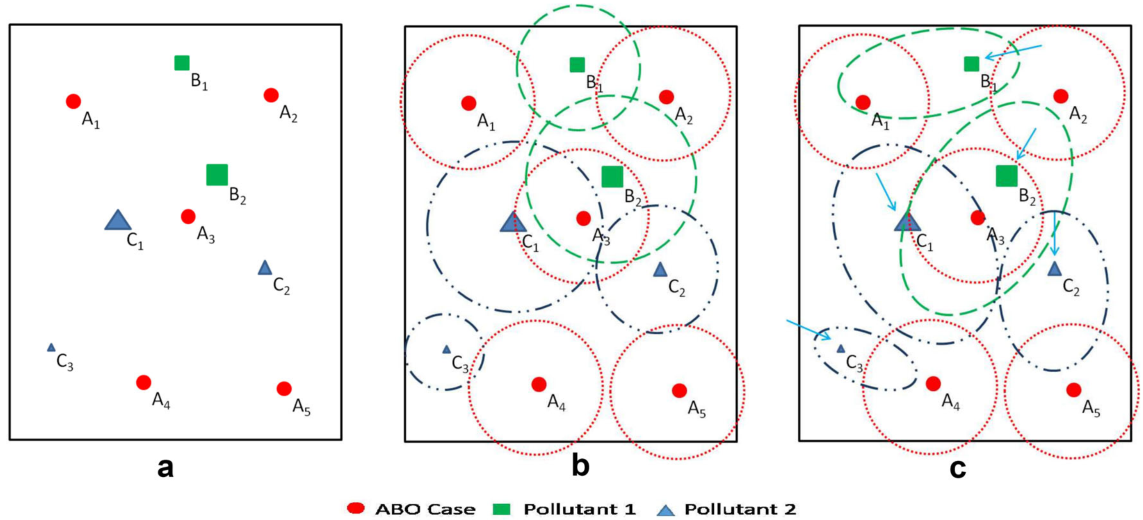

Fig. 5 Extending spatial objects: **a** an example spatial dataset (A—ABO, B and C—Pollutants); **b** buffer sizes of pollutants vary depending on the amount of release; **c** buffer shapes of pollutant emission points change with the wind direction and speed (as indicated by arrows) [11]

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 the model we used, it is assumed that, although the affected region can be different, the area affected by the pollutant is the same [11]. 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 Eq. [11].

$$a = r + \gamma|\mathbf{v}|, \tag{9}$$

$$b = \frac{r^2}{a}, \tag{10}$$

where r is the radius of the original circle, $\mathbf{v}$ 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 Fig. 5.

Table 3 Co-location rules found with AGT-Fisher

CMA	# of Rules
Calgary	259
Edmonton	376
Fredericton	26
Halifax	345
Hamilton	6511
Kingston	59
London	1108
Moncton	28
Montreal	771
OttawaGatineau	258
Quebec	209
Regina	250
Saint John	433
Saskatoon	156
St. John's	13
Toronto	2310
Vancouver	283
Victoria	4
Winnipeg	481

5.2 Results

When AGT-Fisher is applied on the datasets prepared for each CMA, the identified co-location patterns are of the form $X \rightarrow ABO_i$ where $\text{ABO}_i \in \{\text{SGA, PTB, LBW}\}$ and X is a combination of industrial air pollutants. The level-of-significance we used is 0.05. The summary of our obtained results is provided in Table 3.

On average, we discovered 730 co-location rules per census metropolitan area. The maximum number of co-location rules obtained for a single CMA was 6511 for Hamilton. For the given level of significance defined by the experts (i.e., 0.05), a minimum number of rules, 4, were obtained for the CMA of Victoria. It should be noted that with the previous GT method we were not able to find any strong and statistically significant co-location patterns for Victoria irrespective of the fact that it had sufficient cases of adverse birth outcomes and air pollutant emitting facilities. Interestingly, our AGT method increased the evidence and support for four previously insignificant rules, thus increasing their statistical significance, leading to their discovery with AGT-Fisher. We also observed that the average number of rules found by AGT (i.e., 730) is larger than the average number of rules obtained with GT method (i.e., 495). This is due to the fact that, through aggregation, AGT increases the support and significance of previously insignificant patterns, while the support or evidence for current rules and patterns remains the same.

Interestingly, Total Particulate Matter (i.e., TPM - Airborne Particulate Matter with an upper size limit of approximately 100 microns) is present in 1797 co-location rules from all the rules discovered with AGT-Fisher for different CMAs associating with one of the three adverse birth outcomes. Some of the other most common antecedents in the rules were NO_2, CO, Lead, Methanol, Toluene, Xylene, $PM_{2.5}$ (Particulate Matter $\leq$ 2.5 microns) and PM_{10} (Particulate Matter$\leq$10 microns), Arsenic, 2-Butoxyethanol, and Isopropanol. All these are well-known air pollutants causing many adverse health effects including adverse birth outcomes [4,10].

5.2.1 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 Sect. 4.2. 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 the Vancouver data, two contrast sets are found from the 81 unique antecedents (2.4%) included in statistically significant rules involving PTB as the consequence (i.e., $X \rightarrow PTB$ in Vancouver). Those two contrast sets are {Methanol, Toluene, Isopropanol, COl} and {Toluene, Isopropanol, CO}. In other words, when X is one of these two pollutant sets, X → PTB is more significant than X → SGA or X → LBW rules. 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 two contrast sets out of 92 (2.0%) air pollutant itemsets for LBW in Vancouver. Those two are as follows: {Methanol, Toluene NO_2} and {PM, Cadmium}. Similarly, these contrast sets can be reported for the CMAs with all three ABOs as well.

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 81 candidates (6.1%). Some of them are as follows:{Methanol, NO_2, Benzene}, {Benzene, CO, PM_{10}}, and {Benzene, $PM_{2.5}$, Methanol}. 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.

5.2.2 Spatial common sets

Based on the location set L (i.e., 19 CMAs in Canada) and the class set C, we focus on a single type of interesting spatial common sets out of the three described in Sect. 4.2. That is to find common sets for a specific ABO in different CMAs. To find such common sets, in addition to the MaxSig threshold, we use a MinFrac threshold of 0.3 (30%) to specify the minimum number of spatial groups a particular common set should exist in. For instance, let us consider the task of discovering common sets for PTB cases in different CMAs. We fond 42 spatial common sets which are associated with PTB cases in at least 30% of the CMAs. One such significant spatial common set we discovered is that {Lead (and its compounds)} is associated with PTB in 12 of 19 CMAs (63%) such as Toronto, Vancouver, Ottawa, Quebec, Montreal, Edmonton, etc. Other than that, in these 42 sets, interesting spatial common sets such as {PM_{10}, CO}, {TPM, CO}, $PM_{2.5}$, Toluene, Xylene and {Arsenic} exist. We observed that these common sets are also commonly associated with other ABO types as well.

5.3 Robustness evaluation of AGT-Fisher

There are many measures introduced in the association rule mining literature to quantify the "interestingness" or significance of a pattern and filter useful rules based on that [24]. These measures can be used to filter out interesting patterns and measure the quality of the filtered out patterns, thus effectively evaluating the performance of the pattern discovery method. The majority of these measures are based on the frequency of the items in the database (e.g. support). In our work, we use Fisher's p value to filter out the interesting patterns and evaluate the robustness of AGT-Fisher using a measure known as *lift*, as shown in Eq. 11. The lift is an interestingness measure used in the association rule mining community which can measure the dependency between the antecedent and the consequent of a pattern. If the lift is 1, it means that the antecedent and the consequent are independent of each other, whereas if it is larger than 1, they are dependent on each other. In other words, a higher dependency means better positive statistical dependency between the antecedent and the consequent.

The lift measure aligns well with our goal to find co-location rules in which the consequents and the antecedents are strongly dependent on each other. While the p value can measure whether a rule is statistically significant or not based

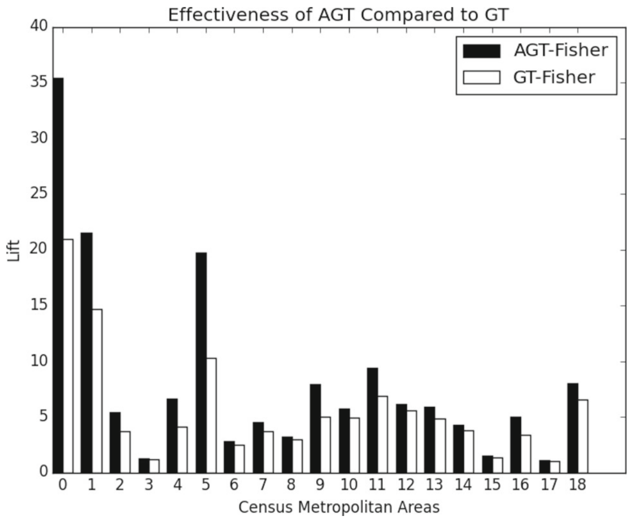

Fig. 6 Average lift of the co-location patterns obtained by AGT-Fisher and GT-Fisher in CNN datasets from 19 CMAs in Canada

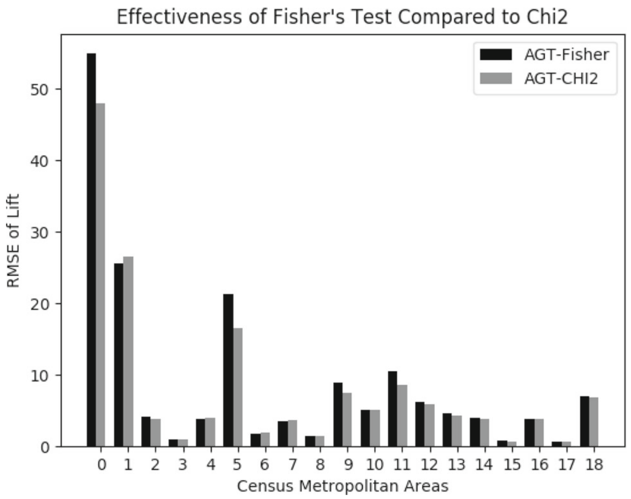

Fig. 7 Average $\text{RMSE}_{\text{lift}}$ of the fivefold cross-validation comparison of AGT-Fisher and AGT-Chi2 in CNN datasets from 19 CMAs in Canada

on a predefined level of significance, from a different perspective, the lift can help to evaluate the statistical significance of the filtered rules by quantifying the statistical dependency between their antecedents and the consequents. Moreover, previous statistically significant association rule search methods have also used p values to filter the rules, and lift to evaluate the effectiveness of the proposed approach [9,23]. Selecting an evaluation measure and a filtering criterion is primarily a domain specific problem. In our application, using p value-based rule filtering and lift-based evaluation was also well supported in discussions with our interdisciplinary collaborators.

$$\text{lift}(X \rightarrow A) = \frac{\text{support}(X \cup A)}{\text{support}(X) \times \text{support}(A)} \tag{11}$$

When we applied AGT-Fisher and GT-Fisher on the spatial datasets from 19 CMAs, the average lift of the rules found from each method is given in Fig. 6. It is clear that in all 19 datasets, AGT-Fisher achieves a higher lift than GT-Fisher, indicating the aggregated grid transactionization can find co-location patterns which are more statistically dependent. In all the datasets, both the approaches achieved an average lift greater than 1.

Another major aspect in the robustness evaluation is that the quality of the rules found should be held in unseen datasets as well. This quality may depend on the type of significance test being used. Hence, we performed experiments using Fisher's test and χ^2 test to find dependency rules with the AGT technique to evaluate the robustness of using Fisher's test. In particular, we performed a fivefold cross-validation test, as proposed in previous studies [8,9]. Specifically, each of the 19 spatial datasets is randomized and divided into five partitions. Fivefold cross-validation is an iterative process in which each partition is treated as the test set in exactly one iteration, and all other partitions are merged into a single training set for that iteration. Subsequently, AGT-Fisher and AGT-Chi2 are applied on this training data to obtain the best set of co-location rules. Finally, the lift of these rules is calculated in the testing data and the root mean squared error of the lift was obtained ($\text{RMSE}_{\text{lift}}$). We calculated this RMSE_{lift} as follows. First, the rules with the highest statistical significance are obtained from the training data. Then, their lift values are obtained for the training data and the test data. The RMSE is calculated for the difference of these train and test lift values for all of the statistically significant rules obtained from the training data. The goal behind this approach is to evaluate whether the statistical dependency of the significant rules obtained from the training data would hold over future / unseen data (i.e., test data). We averaged the obtained $\text{RMSE}_{\text{lift}}$ to obtain the error in the cross-validation experiment. This average $\text{RMSE}_{\text{lift}}$ is reported for all 19 spatial datasets as shown in Fig. 7. These results indicate that AGT-Fisher maintains a low or similar average $\text{RMSE}_{\text{lift}}$ in all 19 datasets compared to the AGT-Chi2 approach. This indicates that AGT-Fisher is more stable and is capable of discovering rules which can also be strong over the unseen data.

Efficiency is another important aspect to consider. All the algorithms we experimented had execution times below five seconds with the CNN datasets. However, previous studies have shown that when more demanding datasets are used, Fisher's test is much more efficient than the χ^2 test in finding dependency rules. Our results and these results from previous studies can prove that AGT-Fisher can be a more robust algorithm than other approaches by efficiently obtaining a set of co-location rules with better statistical dependency which will also hold in unseen data.

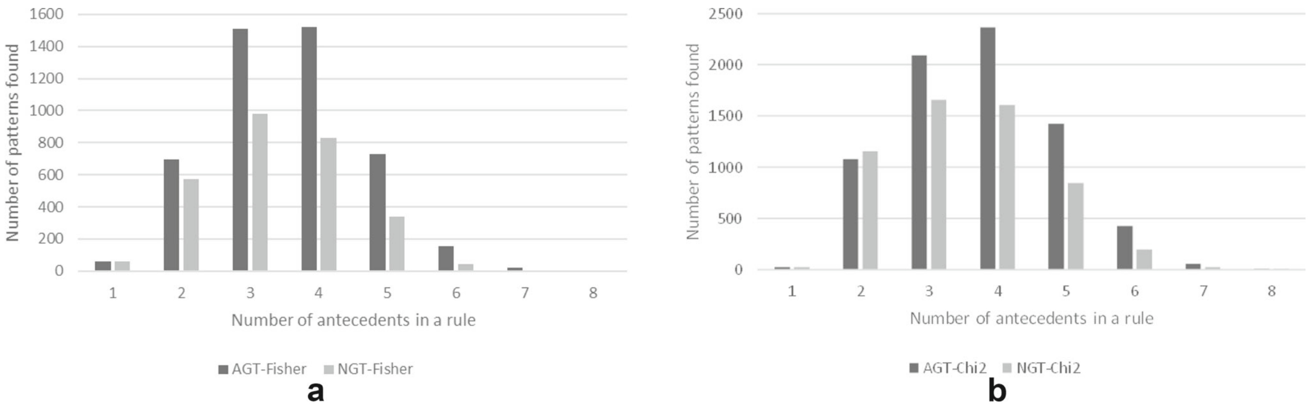

Fig. 8 Relationship of number of rules with higher number of antecedents: **a** AGT-Fisher versus NGT-Fisher; **b** AGT-Chi2 versus NGT-Chi2

One of the main objective of the proposed AGT transactionization method is to find associations between combinations of chemicals and the adverse birth outcomes. Hence, we evaluated the capability of AGT to find co-location patterns with more than one chemicals. As depicted in Fig. 8, the AGT method clearly outperforms NGT by finding more rules with a larger combination of chemicals. This is an indication that AGT is successful in detecting the "combined effect" of non-overlapping air pollution buffer regions ignored by the NGT.

5.4 Empirical evidence based evaluation

Associations between chemical exposure and adverse health effects have been a well-studied area in the environmental health and toxicology literature. To validate the quality and the impact of the associations we find between air pollutants and ABOs, we utilize these known associations from the literature. Comparative Toxicogenomics Database (CTD)[2] provides a comprehensive collection of manually curated known associations between chemicals and disease or adverse health conditions. We use these CTD associations to validate our methodology.

We considered 81 chemicals (i.e., air pollutants) from the NPRI database in our study. We queried these chemicals for known associations with adverse birth conditions in CTD. Out of the 81 chemicals, CTD has recorded 43 chemicals as known chemicals which form associations with adverse birth outcomes such as premature birth and low birth weight. The co-location rules we found, which are of the form $X \rightarrow A$ (i.e., X is a set of chemicals and A is an ABO), have 62 unique chemicals out of 81 we considered from NPRI in their antecedents associating with adverse birth outcomes such as low birth weight and premature birth. Our rules only missed 9 chemicals (i.e., we missed 21% of the 43 chemicals CTD recorded to have known chemical-abo associations) from the known associations from CTD. This could be due to a limitation in the dataset or other geographical factors. On the other hand, in the rules we found, there are 28 new chemicals indicating a potential association with ABOs which CTD does not indicate any known associations. Hence, from the co-location rules, we found that 54% of the chemicals conform to the known knowledge, whereas the remaining 45% can be considered as potential new knowledge which could be used to build research hypotheses to perform further research. We emphasize that our contribution is not only that we find these new potential associations of 28 chemicals with adverse birth cases, but we also find many mixtures of chemicals which could be associated with ABOs from our co-location rules with antecedents consisting of more than one item. Information on such chemical mixtures are not common in public databases such as CTD. Hence, the 45% of the chemicals missing from the CTD database could also be chemicals, of which the toxicity is activated when it is mixed or co-emitted with other known harmful chemicals. Identifying the toxicological impact of chemical mixtures on health is a major challenge that is receiving considerable attention in recent times. Manual curation of such combinations could be an arduous, and a sometimes impossible task. However, with the help of our co-location pattern mining algorithm, this task could be accomplished and the results might be used to assist researchers in the environmental health/toxicogenomics domains to narrow down the hypotheses space.

We also queried the chemicals from the rules discovered using the alternate methods (i.e., NGT-Fisher, NGT-Chi2, and AGT-Chi2). The results indicate that the difference between the outcome is insignificant when compared with the outcome of AGT-Fisher. This is expected as CTD primarily focuses on the hand curated relationships between individual chemicals and ABOs. Hence, we aggregated all the rules we found (indicating any type of ABO) and obtained the set of chemicals from each method which indicates any rela-

[2] www.ctdbase.org.

tionship to ABOs. This set of chemicals is then used against the CTD database for comparison. This, however, effectively discards all the information on the combination of chemicals. Hence, AGT-Fisher shows a similar behavior to the other methods. However, evaluation with the ground truth has provided us enough confidence on our approach in finding relationships between individual chemicals and ABOs so that we can use the multiple combinations of chemicals—which we discover more than others (as shown in Fig. 8)—to build new hypotheses.

6 Conclusion

In this work, we proposed a new co-location pattern mining technique, AGT-Fisher, to robustly find statistically significant co-location patterns. In AGT-Fisher, we addressed two major challenges in finding co-location patterns. First, we addressed the challenge of effectively transforming a spatial dataset into a transaction dataset by proposing an improved transactionization method called AGT. Second, we addressed the challenge of robustly finding statistically significant co-location patterns by using constrained versions of existing statistical dependency rule searching techniques based on Fisher's test to analyze the transactionized dataset. We extended the usage of AGT-Fisher by using it to find spatial contrast sets, a special type of co-location patterns which can discriminate one spatial group from another, and spatial common sets, another special type of co-location patterns which are commonly significant in many spatial regions. Our work is motivated by an important research problem in environmental health to find spatial associations between industrial air pollutants and adverse birth outcomes. However, the applicability of our methods extends to solve problems in many other domains including business, ecology, and transportation. To address our current application problem and to evaluate our approach, we performed experiments by applying the proposed AGT-Fisher on 19 real spatial datasets about adverse birth cases and air pollution in various spatial regions of Canada. In those experiments, we discovered a number of potential and interesting air pollutant(s) associations with adverse birth outcomes. The results we obtained with spatial contrast sets and common sets were able to provide insights beyond the traditional co-location patterns to help practitioners and knowledge users to further understand the application problem. Especially, we found that air pollutants such as NO_2, PM, CO and heavy metals such as Lead, Cadmium, and Arsenic are commonly associated with adverse birth outcomes in many spatial regions. We evaluated these findings with the known associations found in Comparative Toxicogenomics Database and the evidence from the literature. These evaluations suggest that majority of our findings conform to the existing knowledge while the others, especially the combinations of chemicals associated with adverse birth outcomes, could be helpful in forming new hypotheses leading to new knowledge. Our experiments further indicate that AGT can achieve better statistically dependent co-location patterns by having a higher lift in all the datasets we tested. The five-fold cross-validation experiment confirms that by having a better $\text{RMSE}_{\text{lift}}$, AGT-Fisher can discover rules which are similarly strong in unseen datasets as well. This proves that our approach is more robust than the other approaches.

For future research, we are currently working on extending our technique to work with uncertain or probabilistic datasets. In addition to that, if the necessary data are available, using the temporal factor could reveal interesting association patterns based on the seasonality. We are currently working on extending our co-location pattern mining methods to address such complex scenarios as well.

Acknowledgements The authors would like to acknowledge the funding of CIHR/NSERC Collaborative Health Research Program and the Canadian Neonatal Network for providing adverse birth outcome data.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Mohomed Jabbar, M.S., Zaïane, O.R., Osornio-Vargas, A.: Discovering spatial contrast and common sets with statistically significant co-location patterns. In: ACM Symposium on Applied Computing, ACM (2017)
2. Mohomed Jabbar, M.S., Zaïane, O.R.: Learning statistically significant contrast sets. In: 29th Canadian Conference on Artificial Intelligence, pp. 237–242. Springer (2016)
3. Adilmagambetov, A., Zaiane, O.R., Osornio-Vargas, A.: Discovering co-location patterns in datasets with extended spatial objects. In: International Conference on Data Warehousing and Knowledge Discovery, pp. 84–96. Springer (2013)
4. Brauer, M., Lencar, C., Tamburic, L., Koehoorn, M., Demers, P., Karr, C.: A cohort study of traffic-related air pollution impacts on birth outcomes. Environ. Health Perspect. **116**(5), 680 (2008)
5. Bay, S.D., Pazzani, M.J.: Detecting group differences: mining contrast sets. Data Min. Knowl. Discov. **5**(3), 213–246 (2001)
6. Antonie, L., Zaïane, O.R., Holte, R.C.: Redundancy reduction: does it help associative classifiers? In: ACM Symposium on Applied Computing, pp. 867–874. ACM (2016)
7. Canada, E.: National Pollutant Release Inventory. Tracking Pollution in Canada. http://www.ec.gc.ca/inrp-npri/
8. Hämäläinen, W.: Statapriori: an efficient algorithm for searching statistically significant association rules. Knowl. Inf. Syst. **23**(3), 373–399 (2010)
9. Hämäläinen, W.: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowl. Inf. Syst. **32**(2), 383–414 (2012)
10. Lavigne, E., Yasseen, A.S., Stieb, D.M., Hystad, P., van Donkelaar, A., Martin, R.V., Brook, J.R., Crouse, D.L., Burnett, R.T., Chen, H., et al.: Ambient air pollution and adverse birth outcomes: dif-

ferences by maternal comorbidities. Environ. Res. **148**, 457–466 (2016)

11. Li, J., Adilmagambetov, A., Mohomed Jabbar, M.S., Zaïane, O.R., Osornio-Vargas, A., Wine, O.: On discovering co-location patterns in datasets: a case study of pollutants and child cancers. GeoInformatica **20**, 1–42 (2016)
12. Li, J., Zaïane, O. R., Osornio-Vargas, A.: Discovering statistically significant co-location rules in datasets with extended spatial objects. In: Data Warehousing and Knowledge Discovery, pp. 124–135. Springer (2014)
13. Ha, S., Hu, H., Roussos-Ross, D., Haidong, K., Roth, J., Xu, X.: The effects of air pollution on adverse birth outcomes. Environ. Res. **134**, 198–204 (2014)
14. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Berlin (2015)
15. Huang, Y., Pei, J., Xiong, H.: Mining co-location patterns with rare events from spatial data sets. Geoinformatica **10**(3), 239–260 (2006)
16. Webb, G.I.: Discovering significant rules. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 434–443 (2006)
17. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoo, J.: A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proceedings of the 2004 SIAM International Conference on Data Mining (SDM), (2004)
18. Shekhar, S., Huang, Y.: Discovering spatial co-location patterns: A summary of results. In :Proceedings of the 7th International Symposium on Spatial and Temporal Databases (SSTD), pp. 236–256 (2001)
19. Barua, S., Sander, J.: Sscp: mining statistically significant co-location patterns. In: Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD), pp. 2–20 (2011)
20. Zaki, M.J.: Scalable algorithms for association mining. IEEE Trans. Knowl. Data Eng. **12**(3), 372–390 (2000)
21. Eick, C.F., Ding, R. Parmar W., Stepinski, T.F., Nicot, J.: Finding regional co-location patterns for sets of continuous variables in spatial datasets. In: Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems (2008)
22. Sugiyama, M., Borgwardt, K. M.: Finding Significant Combinations of Continuous Features. arXiv preprint arXiv:1702.08694, (2017)
23. Webb, G.I., Zhang, S.: K-optimal rule discovery. Data Min. Knowl. Discov. **10**(1), 39–79 (2005)
24. Jalali-Heravi, M., Zaïane, O.R.: A study on interestingness measures for associative classifiers In :Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1039–1046 (2010)
25. Koperski, K., Han, J.: Discovery of spatial association rules in geographic information databases In :Proceedings of International Symposium on Advances in Spatial Databases, pp. 47–66 (1995)
26. Cressie, N.: Statistics for Spatial Data (2015)
27. Hämäläinen, W.: New upper bounds for tight and fast approximation of fishers exact test in dependency rule mining. Comput. Stat. Data Anal. **93**, 469–482 (2012)
28. Chou, Y. H.: Exploring Spatial Analysis in Geographic Information Systems (1997)