Discovering Spatial Patterns using Statistically Significant Dependencies

by

Mohomed Shazan Mohomed Jabbar

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Co-location pattern mining is a class of techniques to find associations among spatial features. It has a wide range of applications varying from business to science. 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 a potential cause to 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. However, the increasing complexity of the application problems poses new challenges that traditional approaches are unable to address well. For instance, comparing and contrasting spatial groups is one such complex task posed as a research question in our application problem. Furthermore, traditional co-location pattern mining techniques heavily rely on frequency based thresholds which discard underrepresented rare patterns and find exaggerated noisy patterns which may not to be equally prevalent in unseen data. To address limitations in frequency based methods, some association studies propose to use statistical significance tests. The use of a spatial data transactionization mechanism helps exploiting such statistically significant association mining methods to find strong co-location patterns more efficiently. Towards this end we propose a novel approach, AGT-Fisher, to achieve the task of transactionization and using statistically significant dependency rules to find strong co-location patterns more efficiently. Our experiments reveal that the proposed AGT-Fisher could indeed help in finding co-location patterns with a better statistical significance. Furthermore to compare spatial groups we introduce two new spatial patterns: *spatial contrast sets* and *spatial common sets*, and techniques based on AGT-Fisher to mine them efficiently. Our evaluation reveals that the contrast sets we found can successfully

distinguish one group from the others. We also propose a new visualization framework, VizAR, to interactively visualize complex spatial patterns such as the ones we intend to discover. With the proposed methods and the VizAR tool, we discovered that air pollutants such as heavy metals, NO_2, $PM_{2.5}$, PM_{10} and TPM are frequently associated with adverse birth outcomes.

Preface

Some part of the preliminary methods used in Chapter 3 of this thesis has been published as Jundong Li, Aibek Adilmagambetov, Mohomed Shazan Mohomed Jabbar, Osmar R. Zaïane, Alvaro Osornio-Vargas and Osnat Wine, "On Discovering Co-Location Patterns in Datasets: A Case Study of Pollutants and Child Cancers," Geoinformatica, vol. 20, issue. 4, 651-692. I contributed by performing experiments as well as by writing and editing parts of the manuscript. Osmar R. Zaïane was the supervisory author and was involved with concept formation and manuscript composition. A. Adilmagambetov and J. Li contributed on composing a preliminary version of the manuscript via their MSc research work that precede this paper. A. O. Vargas and O. Wine contributed their insights from the application domain (i.e. Pediatrics) perspective in collecting data and designing experiments.

Part of Chapter 4 is published as Mohomed Shazan Mohomed Jabbar and Osmar R. Zaïane, "Learning Statistically Significant Contrast Sets," In Proceedings of the 29th Canadian Conference on Artificial Intelligence, 237-242. I was responsible for the data collection, analysis as well as the manuscript composition. Osmar R. Zaïane was the supervisory author and was involved with concept formation and manuscript composition.

Acknowledgements

I would like to express my deepest gratitude to my supervisor and mentor, Prof. Osmar R. Zaïane for his advice, support, encouragement and guidance throughout my journey. I am extremely fortunate to have such a supervisor who genuinely cared about my well being and success in my work, and who always kept his doors open whenever I ran into any trouble. Thank you Osmar for all your invaluable insights and knowledge, constructive feedback, motivation, support and encouragement on this thesis.

I am also grateful to Prof. Alvaro O. Vargas for his invaluable feedback, continuous support and encouragement throughout this research. My special thanks also goes to Jesus and Charlene for helping me in preparing the datasets, Saeed for his programming contributions in VizAR program and Jundong for his support in the work done related to Geoinformatica research paper. I also acknowledge the support of Osnat and Leslie in many occasions. I also would like to thank the whole DoMiNO team for their feedback and the support.

I also would like to thank Dr. Prakeshkumar Shah and the Maternal infant Care (MiCare) research team at the University of Toronto for their support given on analyzing adverse birth data from Canadian Neonatal Network and for hosting me at their Lab in Toronto during my short stay there. I am also especially indebted to my committee members, Prof. Mario Nascimento and Prof. Yutaka Yasui, for taking time from their busy schedules to read my thesis, and for providing me with great comments and insightful advices.

Last but not least, I must express my heart-felt gratitude to my dear friends and family for their continuous support and unconditional love without whom I would not be where I am today.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ABO Adverse Birth Outcome

SGA Small for Gestational Age

LBW Law Birth Weight

PTB Preterm Birth

CNN Canadian Neonatal Network

APHP Alberta Perinatal Health Program

CMA Census Metropolitan Area

DoMiNO Data Mining and Neonatal Outcomes

AGT Aggregated Grid Transactionization

CAR Classification Association Rule

NGT Non-aggregated Grid Transactionization

List of Algorithms

List of Publications

Refereed Conference Papers:

1. Mohomed Shazan Mohomed Jabbar and Osmar R. Zaïane. Learning Statistically Significant Contrast Sets. In *Proceedings of the Canadian Conference on Artificial Intelligence*, 2016.

Refereed Journal Papers:

1. Jundong Li, Aibek Adilmagambetov, Mohomed Shazan Mohomed Jabbar, Osmar R. Zaïane, Alvaro Osornio-Vargas and Osnat Wine. On Discovering Co-Location Patterns in Datasets: A Case Study of Pollutants and Child Cancers. *International Journal of Geoinformatica*, 2016.

Chapter 1

Introduction

Recent advancements in science and technology have led to a massive collection of unconventional and rich datasets in various fields varying from telecommunication to environmental health. Spatial data sets are one such highly important, rich dataset type that have recently started to gain attention. Although traditional spatial statistics and GIS analysis techniques have been around for some time they are unable to cope with the new challenges imposed by increasingly complex spatial data mining tasks. This necessitates the need to implement new data mining methods which are not only capable of handling the massive size of the big spatial data but also is capable of addressing non-conventional knowledge discovery tasks in spatial datasets.

Spatial data mining can be defined as a branch of data mining which intends to discover previously unknown interesting patterns from spatial datasets. Spatial outliers (e.g. detection of bad traffic sensors), co-location patterns (e.g. symbiotic relationships between species based on location), spatial classifiers (e.g. prediction of habitats of endangered species), and spatial clustering (e.g. crime hotspots) are four of the most important types of tasks of interest in spatial data mining [38]. There are a wide range of important applications to discover these patterns in many domains, such as earth and atmospheric sciences, environmental health, and telecommunications. The significance of the impact of some of these applications and the complexity of the research questions posed by them require to think beyond the traditional methods and to develop novel techniques to solve them.

1.1 Motivation

Environmental health—a branch of public health—is a very important research field which concerns about all the aspects of natural and man made environments that can affect the health of humans. This is closely related to the environmental protection which in general is concerned with protecting the natural environment to safeguard the whole ecosystem. Addressing the challenges posed in environmental health can make a huge impact on improving the lives of humans while benefiting the whole ecosystem. Challenges in environmental health heavily involves analysing datasets with natural and artificial geographic features such as communities where people live, facilities which emit industrial air pollutants as well as studying the impact of climate in the geographic regions which are affected by those features. This brings us to spatial data mining which can accomplish such complex geographic analysis tasks.

Particularly in our current work, we are motivated by a challenging research question in environmental health: "Do air pollutant emissions play any role in adverse birth outcomes?" We collaborate with the researchers at the Canadian Neonatal Network (CNN)[1] and the Department of Pediatrics at the University of Alberta to discover such potential relationships between industrial air pollutant emissions and adverse birth cases in 21 Canadian cities. When forming hypotheses to answer this question, maternal exposure to air pollutants during the pregnancy has to be well understood. In fact there are many studies [17] suggesting that associations between air pollutants and adverse birth outcomes exist. Most of such studies follow traditional statistical models [17], epidemiological methods, or cohort studies. However, discovering such associations turns out to be a spatial data 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. Such co-location patterns can explain which combination of industrial air pollutants are co-located or in near proximity with adverse birth outcomes hinting possible associations. A sample dataset with co-location patterns discovered is shown in Fig-

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

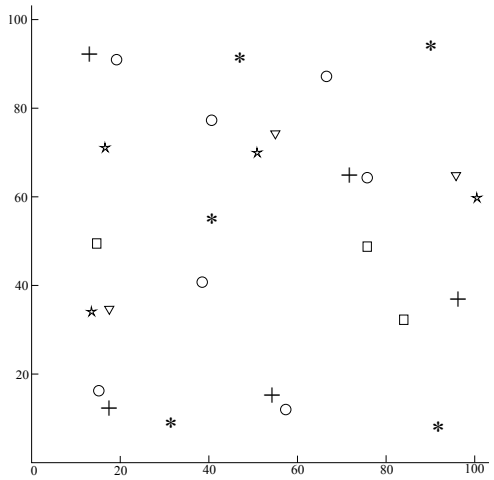


Figure 1.1: Sample dataset with point spatial features. Instances of feature sets $\{+, \circ\}$ and $\{\star, \triangledown\}$ are often located close to each other [29].

ure 1.1. When dealing with rich datasets which contain data from multiple spatial regions, which have to be treated separately to identify locally and globally significant co-location patterns, one has to look beyond the traditional co-location patterns such as the ones shown in Figure 1.1. For instance, given adverse birth cases 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 that are more significantly associated to low birth weight in Toronto area than any other city in Canada?" To answer such questions not only the classical co-location patterns, but also discriminative co-location patterns which can contrast a particular spatial group from the others or as we define it, spatial contrast sets could be of great interest. On the other hand some other researcher might be interested to know about which co-location patterns that are co-located with instances of a particular adverse birth outcome such as LBW is significant or prevalent in at least half of the geographic areas under study (i.e. spatial common sets). There are three major challenges when dealing with interdisciplinary problems as complex as above: 1) Finding rare but statistically sound co-location patterns; 2) Comparing various spatial groups to discover locally unique or globally consistent significant co-location patterns; and 3) Visualizing discovered co-location patterns. Traditional co-location pattern mining techniques, which are primarily based on neighbourhood relationships and join based approaches [37], are not capable of effectively and ef-

ficiently finding rare but significant patterns owing to the fact that they heavily rely on defining a global prevalence or neighbourhood distance threshold [44]. Given a strict prevalence threshold value this could lose many rare but statistically significant patterns while a low threshold is given a large number of patterns detected could become noisy and not useful. Addressing this limitation, a transaction based approach relying on association rule mining techniques has been proposed to find statistically significant co-location patterns [29]. Although this approach is capable of dealing with extended spatial objects and was able to find rare but statistically significant patterns, it was unable to find patterns beyond traditional co-location patterns which were restricted to treat all the given data instances as belonging to a single spatial region. In such methods, patterns which are only significant in a sub-region of the larger spatial region can be falsely ignored as insignificant in the whole spatial region. Moreover such global pattern mining approaches cannot be extended to compare spatial groups and regions to find discriminative or common co-location patterns which would be valuable to researchers who are interested only in a particular spatial subgroup or only in a particular group in multiple locations. Although there exists a class of data mining techniques called contrast set mining to discover similar discriminative associative patterns to characterize a particular class from other classes in non-spatial datasets [33], a variant which can discover significant contrast sets to contrast spatial groups (i.e. spatial contrast sets) does not exist. Moreover, in the literature, no significant approach has been proposed to discover association patterns which are consistently significant in many spatial regions (i.e. spatial common sets). In any knowledge discovery task, visualization plays a major role to convey the discovered patterns. Hence, several approaches have been suggested in the past to visualize co-location patterns as well. Most of such approaches are restricted to a single level of abstraction such as visualizing patterns in the geographic space (i.e. pattern level abstraction) [14]. When visualizing a complex co-location pattern results set, a single abstraction visualization scheme might not be sufficient. Developing a visualizing scheme which provides multi levels of abstractions to better understand, compare and contrast co-location patterns in multiple regions is an open challenge in the literature.

1.2 Problem Definition

Co-location pattern and contrast set mining have strong foundations in the association rule mining problem domain. Hence, we formulate our core framework to discover spatial patterns 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 $\{A_i, A_j, ..., A_k\} \subset A$. 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 [6]. Classification Association Rule (CAR) are a special case of general association rules [8, 30]. 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 suitable transactionization technique if it can be transformed into a transaction database D^S where multiple spatial instances are aggregated into one transaction based on proximity, 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 [31].

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 important associative patterns which are used

to characterize a particular class and contrast it from the others. It can be defined as in Definition 3 [36].

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 CARs. STUCCO algorithm [10] is one of the foremost technique to mine contrast sets. Originally, if set X in class association rule $X \implies c_i$ suffices STUCCO deviation conditions as defined in Equation 1.1 and 1.2, then X is considered as a contrast set for class c_i which can distinguish c_i from the other classes. Condition in Equation 1.1 imposes that the support of a contrast set is significantly different across various groups. The second condition in Equation 1.2 imposes that the difference of support of a contrast set across different groups is sufficiently large.

$$\exists_{i,j} P(X|c_i) \neq P(X|c_j) \tag{1.1}$$

$$\max_{i,j} |support(X, c_i) - support(X, c_j)| \geq min\_dev \tag{1.2}$$

Spatial contrast sets can be recognized as a specific case of general contrast sets where the classes which are used to contrast are in fact groups in geographic space. This problem is further explored in Chapter 4. **Spatial common sets** are another type of association patterns which also can be defined as conjunctions of attribute value pairs on mutually exclusive classes similar to contrast sets as in Definition 3. However, the significance conditions of common sets are slightly different to that of contrast sets. Conditions to find common-sets as defined in Chapter 4 verify that the difference of statistical significance of a particular pattern among various groups are below a certain threshold and the pattern exists in a user defined fraction of the groups under study. No significant work exists to define or find common sets in spatial or non spatial datasets. Lessons learned from existing works in contrast set mining such as STUCCO are useful in devising methods to find spatial contrast sets and spatial common sets.

1.3 Thesis Statements

To address the drawbacks and limitations posed by previous works, in this thesis we investigate the application of statistically significant dependency rules to discover advanced spatial patterns. In doing so, we pose the following theses:

Thesis 1 Statistically significant dependency rules can be used to efficiently and effectively detect statistically sound co-location patterns irrespective of their prevalence.

Thesis 2 Statistically significant dependency rules can be used to efficiently and effectively discover statistically sound spatial contrast and common sets.

Thesis 3 Visualization tools can be devised to effectively explore a large number of co-location patterns in various spatial regions and can be helpful in discovering as well as interpreting spatial contrast and common sets.

1.4 Thesis Contributions

Summaries of our major contributions while investigating the theses we posed are as follows:

1. A proposal of a novel grid based transactionization algorithm, **AGT**, for spatial datasets.

2. We introduce two novel types of spatial patterns: 1) Spatial contrast sets; and 2) Spatial common sets, and propose two new algorithms, **DiSConS** and **DiSComS**, to efficiently and effectively mine those patterns that are statistically significant, in a spatial dataset.

3. A proposal and implementation of a visualization system, **VizAR** (Visualizing Spatial Association Rules), to visualize spatial patterns such as co-location patterns, spatial contrast and common sets.

4. A proposal of a design for a novel spatial pattern discovery framework: **Di3SP** (Discovering Statistically Significant Spatial Patterns), to mine and visualize statistically significant co-location patterns, spatial contrast and common sets.

5. We present the spatial patterns discovered by using the above spatial pattern discovery tools and techniques, indicating potential association between sets of various industrial air pollutants and adverse birth outcomes in 21 cities in Canada. This is the first time that co-location pattern mining techniques have been applied to solve this particular problem in environmental health.

1.5 Research Methodology

The work in this dissertation is motivated by an interdisciplinary research problem in Environmental Health—finding relationships between industrial air pollutants and adverse birth outcomes. This is one of the primary goal of the Data Mining and Neonatal Outcomes (DoMiNO) project team which we are part of. DoMiNO project involves an interdisciplinary team of researchers including investigators (from computing sciences, neonatology, pediatrics, epidemiology, bio-statistics and knowledge translation areas) from five universities across Canada / US including University of Alberta and knowledge users from government (Health Canada), Canadian Perinatal Programs Coalition and social organizations (the Canadian Partnership for Children's Health & Environment). Due to the interdisciplinary nature of the project and the motivating problem we base our research methodology on the CRISP-DM (CRoss-Industry Standard Process for Data Mining) process. In the following, we discuss the various phases of our methodology.

1.5.1 Problem Understanding

Discovering potential associations between industrial air pollutants and adverse birth outcomes is realized as a spatial pattern mining task. Most closest pattern mining approach which could be of interest in the context of this problem would be co-location pattern mining methods. However, the stakeholders of the DoMiNO project are interested in not only finding rare but statistically sound relationships

between air pollutants and Adverse Birth Outcome (ABO)s, but also on comparing various spatial groups to discover discriminant or common association patterns. This requirement has been emphasized mainly due to the fact that the DoMiNO works with various levels of data from provincial (e.g. the province of Alberta) to National (e.g. Canada). To understand unexplained phenomenon in such levels it is necessary to look beyond traditional co-location pattern mining techniques. With this initial understanding of the problem we surveyed the existing literature on spatial co-location patterns mining techniques to understand the drawbacks and limitations in existing works. Furthermore we also surveyed another set of important techniques called contrast set mining to understand how to compare two groups. The lessons learned in this phase were useful in coming up with new pattern mining methods in the later stages of the research.

1.5.2 Data Collection and Preprocessing

We are primarily interested in two levels of data: Provincial level and City level, at this stage of our work. ABO data was collected by practitioners from the participating health organizations of the DoMiNO project. Provincial ABO data for Alberta was obtained from Alberta Perinatal Health Program (APHP) and ABO data for major cities in Canada was obtained from CNN. An expert guided data collection was carried out to obtain industrial air pollutants from NPRI (National Pollutant Release Inventory) and climate data from Environment Canada to carry out the research.

Once we obtained the raw datasets it has been transformed into a relational schema to ease the storage and preprocessing. Due to the privacy concerns of the patient data, we undertook necessary anonymization steps before further analysis. Following that we applied tokenization, aggregation, table lookups and joins, redundancy operations, spatial indexing, and etc. to clean and construct a focused rich spatial dataset.

1.5.3 Designing Analytical Methods

Based on the lessons learned during our literature survey, we designed a new grid transactionization technique—to transform spatial datasets into an easy-to-mine transaction database—and a set of novel co-location pattern mining techniques which are capable of building hypotheses to the questions asked by the stakeholders of the DoMiNO project. Those pattern mining techniques are designed to discover rare but statistically significant co-location patterns and discriminating (i.e patterns which can contrast a particular group from the others) or common co-location patterns for various spatial groups. Most of these novel techniques utilize the insights from the existing techniques and are inspired by the questions posed by the DoMiNO stakeholders. Especially all the methods are designed to utilize statistically significant dependency rules to test their efficacy in finding relevant associations. Furthermore, our analytical methods can easily incorporate temporal information if available. However, the chemical emission data we use do not contain such information yet. Hence, we do not explore the temporal aspects of our methods in current study. Due to the abstract nature of the original problem during the research cycle, we had to revisit the problem understanding phase more than once during designing of the analytical methods.

1.5.4 Evaluation

Our evaluation of the developed methods to find significant rules is carried out in two phases:1) We validate the efficiency and effectiveness of the usage of statistically significant dependency rules in the pattern mining methods we propose; and 2) With the help of external experts and knowledge users in DoMiNO project we evaluate the quality of the associations we discovered to verify the applicability of our techniques in solving the intended real world problems.

Internal Evaluation

Internal evaluation is primarily carried out on well known transaction or frequent itemset mining datasets from public domain to validate the efficacy of our proposed methods in discovering intended patterns successfully.

External (Expert) Evaluation

External evaluation primarily focuses on the success of the proposed algorithms in discovering associations which can form hypotheses that are actually relevant to the stakeholders of the DoMiNO project. This evaluation is performed with the help of the investigators and experts in pediatrics, neonatalogy, medicines and environmental health from DoMiNO team. Their feedback is helpful in tweaking the parameters and configurations of the proposed models.

1.5.5 Dissemination

Some of the methods and results proposed in this thesis are published in peer-reviewed journals and conferences [33, 29]. Some other manuscripts are still under review. We also presented the associations of ABOs and air pollutants we discovered at the two-day DoMiNO full team workshop held at University of Alberta in August, 2016.

1.6 Outline of the Thesis

We present our literature survey on related association pattern mining techniques in general in Chapter 3. Here, we provide an overview of the work on spatial association patterns such as co-location patterns and its counterpart association rules. In addition a review on contrast set mining to find discriminating patterns between groups is also presented.

In Chapter 3 we present our surveys, methods and experiments on devising a statistically significant co-location pattern mining approach using statistically significant dependency rules. We first present the survey we carried out to review existing work on co-location pattern mining to learn the drawbacks and the ways to overcome the limitations in them. There we introduce our novel grid-transactionization technique and how can statistically significant dependencies can be effectively used in combination with grid transactionization to find statistically significant co-location patterns. The foundation of our Di3SP framework is based on these techniques. We apply our proposed co-location pattern mining approach

to mine association patterns of industrial air pollutants and ABOs on APHP provincial ABO dataset from Alberta. We present our evaluation of the discovered patterns with the help of domain experts. In this chapter we also validate our approach with the experiments in public transaction datasets. This is the first component in our spatial pattern discovery framework.

In Chapter 4, we extend our co-location pattern mining framework proposed in Chapter 3 by introducing two new algorithms, DiSConS and DiSComS, to mine two novel types of patterns called spatial contrast sets and spatial common sets. We initially present our survey of the literature to understand existing techniques to compare and contrast two groups. Then we proceed to introduce what it means to compare spatial groups and define spatial contrast and common sets. We propose two novel algorithms to mine statistically significant spatial contrast and common sets based on the approach suggested in Chapter 3. We also present the results produced by applying the implemented proposed techniques on CNN ABO dataset from 21 Canadian cities to find discriminative or common spatial association patterns between industrial air pollutants and adverse birth outcomes among CNN cities. We also present proof to the validity of our approach using a synthetic dataset. This is the second component in our spatial pattern discovery framework Di3SP.

We outline our proposed spatial pattern visualization scheme **VizAR** in Chapter 5. How to overcome challenges when visualizing co-location patterns in multiple spatial regions, spatial contrast sets and spatial common sets are explored in that Chapter. This is the third and the final stage of our spatial pattern discovery framework.

In Chapter 6 we conclude our work. There we summarise our results and discuss the impact of the discoveries we made addressing one of the major research challenge faced by the health informatics community. We also discuss the impact and the applicability of our proposed frameworks, methods, and tools in various applications, contexts and in datasets. We also discuss what can be done to extend the current work and any limitations in our proposed work.

Chapter 2

Related Work

This thesis is mainly focused on discovering three types of spatial association patterns: 1) Co-location patterns; 2) Spatial contrast sets; and 3) Spatial common sets. These three types of patterns have strong ties with patterns of interest in the areas of association rule mining, co-location pattern mining and contrast set mining. In this chapter we review some of the major methods in those areas to understand the above connection and to learn limitations in them.

2.1 Association Rule Mining

Spatial association patterns, or in particular co-location patterns, have strong roots in traditional association rule mining techniques. Hence, understanding the strengths and weaknesses in existing association rule mining techniques is essential to design new spatial association pattern mining techniques. An association rule is an implication of the form $X \rightarrow A$ where X and A are subsets of items from $I = i_1, i_2, ...i_m$. I is the set of unique items in a given transaction database D. These rules intend to discover relationships between variables or items in datasets. For instance the $X \rightarrow A$ rule explains that if the items or variables in the set X occur or exist together, then the items in set A will also co-occur or coexist along with X. Various measures can be used to determine the strength of such an implication rule. Two commonly used measures are called support and confidence. Given a transaction database D where each transaction or data instance T (i.e. $T \in D$) is a subset of items I existing in that transaction, support and confidence can be defined as in

Equation 2.1 and 2.2 respectively.

$$support(X \rightarrow A) = \frac{|T; X \cup A \subset T|}{|D|} \tag{2.1}$$

$$confidence(X \rightarrow A) = \frac{|T; X \cup A \subset T|}{|T; X \subset T|} \tag{2.2}$$

The first algorithm proposed to mine association rules based on support and confidence is called Apriori [7]. Apriori exploits the monotonicity property of the support measure and mine for frequent itemsets. Monotonicity property states that if an itemset is frequent in a dataset all the subsets of that itemset are also frequent. Apriori starts by discovering all frequent single items and expand them to larger itemsets as long as their support is above the predetermined threshold. Once all such frequent itemsets are found, discovering association rules is straightforward given the confidence threshold. However due to computational inefficiency of this approach in high dimensional large datasets, more robust approaches such as FP-Growth [21] and ECLAT [47] have been proposed. These approaches either use efficient tree data structures or efficient methods to compute the support without generating candidate subsets to achieve a much higher performance improvement over the traditional Apriori algorithm. Although efficient, the underlying measures (i.e. support and confidence) used in FP-Growth and ECLAT are as same as in Apriori. Hence the discovered patterns are similar in any association rule mining algorithm based on support and confidence measures given that same threshold values are applied.

All of the above support-confidence based rule mining techniques face several downsides. For instance, determining the support and confidence threshold is harder. If a low threshold for support or confidence is used, the resulting set might be consisted of a very large number of noisy and useless rules. On the other hand if a stringent threshold is used, the model may face the risk of loosing rare but significant patterns [40]. To mitigate some of these drawbacks several of previous works introduced additional rule quality measures such as the lift to measure the dependency between antecedent and the consequent. Despite having these additional measures most of the traditional association rule mining techniques are strongly dependent on support and confidence thresholds.

2.2 Co-location Pattern Mining

Many techniques have been proposed in the past to discover co-location patterns. Understanding the drawbacks and limitations of those techniques leads to design better co-location mining methods. Co-location patterns are a specific type of associative patterns which represent sets of features whose data instances are co-located in the geographic space. In other words co-location patterns can be defined as association patterns or rules in spatial datasets. Traditionally frequency or prevalence based techniques have been used to detect co-location patterns. However, in recent years statistical tests based co-location pattern mining techniques have gained much more attention. We discuss about some of the statistical tests based approaches in Chapter 3.

Traditional co-location rule mining techniques are mainly based on the neighborhood relations and participation indices [37]. In such methods, co-location patterns were of 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. Similar to association rule mining, these techniques use (k-1) candidate sets to generate (k) candidate sets. 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 [23]. Some methods have also considered extended spatial objects to find co-location patterns. One such method prunes the candidate patterns if the region covered by the features in the given pattern is below a certain coverage ratio threshold [44]. 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 but interesting patterns if the threshold is high. In one of our works [29] we

have provided a comprehensive review and drawbacks in the traditional methods. Furthermore we continue our discussion on statistical test based co-location pattern mining in Chapter 3.

2.3 Contrast Set Mining

Previous works show that characterizing a group or a class to contrast it from others is a challenge as important as clustering or classification tasks. Such a characterization can be meaningfully performed with the help of association patterns. For instance, when given a clinical dataset with two clusters, healthy and non-healthy, an intriguing question researchers would like to ask is: "what are the key variables that can differentiate between healthy and non-healthy people?" The difference between contrasting groups in such situations can be described using conditional probabilities [36]. As an example consider, $P(Healthy|Smoking \wedge Exposure - to - carcinogens)$ and $P(Non - healthy|Smoking \wedge Exposure - to - carcinogens)$. These conditional probabilities can be interpreted as association rules as follows: $(Smoking \wedge Exposure - to - carcinogens) \implies Healthy$ and $(Smoking \wedge Exposure - to - carcinogens) \implies Non - healthy$. The antecedents of these rules could be representing a contrast set [36]. Although contrast set mining is the primary technique used in the literature to find such patterns, there are several other related techniques such as emerging pattern discovery, subgroup discovery and treatment learning which has attempted to pursue goals close to that of contrast set mining but with different specific objective / optimization functions.

In contrast set mining literature there are two main pattern discovery approaches followed. The first approach—the traditional approach—is focused on implementing dedicated techniques to find contrasting patterns whereas the second approach is relied upon association rule mining methods to find contrasting patterns. Some techniques in both of the above approaches adapt statistical significance tests in different levels and forms to confirm the significance of the patterns found, while the others primarily rely on frequency based measures.

2.3.1 Traditional Approaches

Contrast sets were first introduced through the STUCCO [10] algorithm as a way to contrast a specific group from the others. As explained in Section 1, STUCCO uses two deviation conditions to find such strong contrasting patterns. The first condition enforces that the support of a contrast set is significantly different across various groups while the second condition makes sure that difference is large enough. The CIGAR method [22] follows up the work of STUCCO by adding more pruning conditions based on correlation and support. Although some form of statistical tests are used to quantify the significance of support difference, most of these traditional approaches like STUCCO and CIGAR depend on frequency based thresholds such as group support, and prone to the limitations imposed by them.

2.3.2 Association Rule based Methods

The contrast set learning problem is intrinsically connected to the association rule mining problem where several of the previous methods directly exploit this connection. One advantage of this approach is that this allows us to use the existing vast array of well developed association rule mining techniques to find more effective contrast sets efficiently. Traditional approaches such as STUCCO and CIGAR emphasize on representing the contrast sets as *first-kind* of association rules which take the form: $Group \implies Contrast - set$ [10, 22]. However, techniques such as MagnumOpus that rely on association rule mining techniques are primarily based on the *second-kind* of rules which take the form: $Contrast - set \implies Group$ [42, 36]. Recent works in the literature have empirically proven [36] that only the *second-kind* of contrast sets are possible.

Previously it is argued that contrast set mining is simply a specific case of association rule mining task [42]. To prove that, MagnumOpus software has been used to mine *second-kind* of association rules and the antecedents of the mined rules were compared against the contrast sets mined by the STUCCO. The results proved that MagnumOpus was able to extract all the contrast sets mined by STUCCO plus some additional rules. However, the results were unable to clearly conclude that

having more rules is better or not. MagnumOpus software tests the improvement of the confidence of a rule over its immideate generalizations using binomial sign tests or Fisher's exact test among other measures to prune rules. It is claimed that to find statistically significant rules the criterion used by MagnumOpus is insufficient [18]. Especially it is being proved that the rules found in MagnumOpus are redundant. These findings further challenge the usage of rules found by MagnumOpus as contrast sets. On the contrary, in a different data set it is found that MagnumOpus only detects a subset of the contrast sets found by STUCCO. On the otherhand, Terry Peckam [2] disagrees and argues that MagnumOpus only detects a subset of the contrast sets detected in STUCCO [22]. Other approaches have been proposed recently to use traditional association rule mining techniques such as Apriori in combination with the STUCCO deviation conditions to discover contrast sets. The approach was simply to mine classification association rules of the form $X \rightarrow Group$ (i.e. *second-kind* association rules) with user defined support-confidence thresholds and use STUCCO constraints to prune out the irrelvant rules. In another work Apriori based association rules have been used to find contrast sets in brain stroke datasets [26]. All such techniques which are based on Apriori like algorithms [6] inherit limitations imposed by their support and confidence thresholds.

2.3.3 Other Related Methods

There are two other related classes of techniques called, emerging pattern discovery and subgroup discovery, which shares some of the objectives of contrast set mining task [34]. Similar to contrast sets, emerging patterns are also association patterns which aim to "capture emerging trends in time stamped data" [16]. Typically, an emerging pattern in class C_1 over class C_2 is recognized using *growth rate* measure as defined in Equation 2.3.

$$growth\ \ rate(X) = \frac{support(X \cup C_1)}{Support(X \cup C_2)} \tag{2.3}$$

where $support(X \cup C_1)$ is the relative frequency of data instances having X itemset and belonging to C_1 over number of data instances belonging to class C_1. If this

growth rate is larger than a given threshold it is considered as an emerging pattern in class C_1. On the other hand sub group discovery methods also use Apriori like algorithms to mine for association rules but they use different objective functions to optimize for the classification accuracy [34]. Both emerging pattern mining and subgroup discovery tasks are aimed at building a classification model using the patterns discovered. Hence, to quantify the quality of the patterns, often classification accuracy related measures have been used such as the one shown in Equation 2.4 .

$$WRAcc(X, C) = \frac{p+n}{P+N} \times \left(\frac{p}{p+n} - \frac{P}{P+N} \right) \tag{2.4}$$

where *p* is the number of true positive, *n* is the number of false positive, *P* is the number of actual positives and *N* is the number of actual negatives.

2.4 Discussion

Existing association rule mining frameworks have been primarily focused on frequency based measures such as support and confidence to measure the quality of the rule. This causes certain drawbacks such as high noise or loss of rare patterns. To mitigate this, alternative rule mining approaches have been proposed recently which uses statistical significance tests to measure the rule quality. This is further discussed in Chapter 3. Usage of such such statistical tests could eliminate the limitations posed by frequency based methods and able to find rare but statistically significant patterns.

Traditional co-location pattern mining frameworks also pose several drawbacks mostly due to the dependency with frequency or prevalence based measures to quantify the quality of the patterns [29]. Hence, such traditional approaches could either miss rare patterns or add lot of noise to the resulting pattern set depending on the level of threshold selected. This motivated the usage of statistical significant tests, similar to the association rule mining literature, to find co-location patterns. Moreover, instead of developing specific co-location mining techniques it is encouraged to exploit existing association rule mining frameworks with a suitable transactionization approach due to various advantages it brings [5, 29]. For instance, given the vast array of existing AR mining techniques catering various concerns such

as efficiency and effectiveness, designing a co-location pattern mining algorithm which can address similar concerns becomes relatively less complex if built on AR mining techniques. Similarly, exploiting association rule mining techniques to design better contrast set mining algorithms would allow to focus on better pruning mechanisms by being able to rely on the efficiency and the quality of the candidate sets produced by using a suitable association rule mining algorithm. We emphasize the usage of statistical significance tests in such association rule mining algorithms when discovering spatial patterns. Mainly because it can efficiently eliminate patterns which do not show true statistical significance and it can find patterns which are rare but statistically significant.

On the contrary, techniques in emerging pattern mining and subgroup discovery show much less concerns about statistical significance of the patterns. They primarily focus on building better classification systems [34]. Although the pattern quality measures in these different pattern mining techniques are indirectly compatible to some level, contrast set mining techniques primarily vary because of the statistical significance tests used while catering for different trade-offs with other additional frequency based measures.

In this thesis our primary objective is efficient and effective discovery of three types of patterns as we explained above. Although a number of studies have been performed on one of those pattern types—co-location patterns—no significant prior works have been carried out to find other two types of patterns, contrast sets and common sets for spatial data. Although some work exist in emerging pattern mining literature to discover spatial patterns due to the differences in objective functions and lack of statistical significance in the patterns found requires further work to be carried out in the problem domain [13, 15]. On the other hand in the literature not many works have been carried out to find common association patterns among different groups. A proper framework and evaluation criterion become essential to tackle the common set discovery problem when it comes to complex datasets such as the spatiotemporal data.

Chapter 3

Statistically Significant Spatial Co-location Patterns

One of the primary objectives of our work is to find significant co-location patterns. In this chapter we outline the problem and current state-of-the-art in finding "significant" co-location patterns, discuss our proposed approach and present our experimental results.

3.1 Background

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". As an example consider Oxpecker, a bird species which forms a symbiotic relationship with large mammals such as Impalas and Zebras. Because of this relationship, Oxpeckers are restricted to the neighbourhood of such large mammals forming a co-location pattern. Discovering such patterns among other similar species is helpful in learning unknown symbiotic relationships. Similarly, there is a wide range of applications of co-location pattern mining in earth and atmospheric sciences, environmental and health sciences, telecommunications and many more. However, in this thesis we are specifically interested in answering a particular research question in environmental health: "Do industrial air pollutants have any impact or associations with adverse birth outcomes?" This problem can be directly converted to a co-location

pattern mining problem where the objective is to find co-location rules of the form: $industrial\_air\_pollutant\_set \rightarrow ABO$ where the instances of participating antecedents and consequent features can be seen to co-occur in the near geographic proximity.

When searching for co-location patterns, frequency based methods fail due to the usage of ambiguous prevalence thresholds and inability to capture rare patterns. On the other hand, in co-location patterns, quantifying the dependency between the antecedents and the consequent could be a great way to measure the significance of the rule. Although the confidence measure attempts to capture this dependency in terms of conditional probability, there is no guarantee that the measured confidence level of a rule would hold in unseen data. In other words the dependency of a rule in observed data should not be merely by chance but it should have a true dependency in future data or in other sampled datasets as well. Statistical significance tests can be used to quantify this notion of true dependency. In Chapter 2 we also concluded that association rule based techniques could be of great use when designing new co-location pattern mining methods. Hence in this Chapter we discuss how statistically significant dependencies (rules) could be used to accomplish this task.

3.1.1 Problem Definition

Given a spatial dataset S where $s \in S$ (i.e. instances in spatial dataset) can be defined as a vector, $s = [long., lat., feature_i, othercontextualdata, ...]$, consisting of longitude, latitude, spatial feature ID (e.g. $Pollutant_1, Pollutant_2, ABO_1, ...$) and other contextual data such as climate information (e.g. average wind speed and direction). This S dataset can be transformed into a transaction dataset D^S using a suitable transactionization algorithm [29]. In such an event, $E^S \in D^S$ is a vector representing a single transaction, where $E^S = [ID, feature_1 \in \{0, 1\}, feature_2 \in \{0, 1\}, ...]$. E^S defines a neighborhood relationship based on the s spatial instances. A transaction represent a set of spatial features whose instances from S are in the close spatial proximity (e.g. a spatial clique based on a given distance threshold). 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 inter-

preted as co-location rules as defined in Definition 2.

Given a co-location rule, $X \rightarrow A$, the dependency between X and A is traditionally measured by using an empirical p-value. Empirical p-value of a co-location rule is computed by taking the fraction of simulated datasets which yield a higher prevalence measure or the confidence value of that rule than in the observed data. If this fraction (i.e. p-value) is lower than a given level of significance then the rule can be identified as statistically significant. 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 χ^2 test are more flexible and extensively used in the recent literature.

3.1.2 Related Work

In recent years, some methods were designed to use statistical significance tests to find dependency rules (i.e statistically significant association rules) in association rule analysis as well as co-location pattern mining. In the following we discuss some of the important work in this area.

Statistically Significant Association Rules

MagnumOpus is one of the early notable algorithms to consider statistical significant tests to find association rules [43]. However, due to various issues such as redundancy and inefficiency in early methods, better approaches were suggested in recent years. StatApriori [19] and its successor Kingfisher [18] are two such algorithms proposed recently to mine statistically significant dependency rules. Out of these the Kingfisher algorithm is proven to be far more efficient and effective in finding non-redundant statistically significant dependency / association rules [18]. 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 p_F (i.e. p-value) of occurring observed or stronger dependency by chance can be computed using a cumulative hypergeometric distri-

bution in Fisher's test as depicted in the Equation 3.1.

$$p_F(X \rightarrow A) = \sum_{i=0}^{J} \frac{\binom{m(X)}{m(XA)+i}\binom{m(\neg X)}{m(\neg X \neg A)+i}}{\binom{n}{m(A)+i}} \tag{3.1}$$

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 [20]. ***Kingfisher*** also identifies $X \rightarrow A$ as a redundant rule 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 such as Fisher's p-value. Kingfisher uses enumeration trees, efficient search mechanisms and pruning heuristics to efficiently search the solution space to find such significant rules. Other than Fisher's exact test, χ^2-test is also used to find the statistical significance of association rules in the literature. However, Fisher's exact test is empirically proven to be effective, efficient and much scalable compared to the χ^2-test [18].

Statistically Significant Co-location Patterns

SCCP algorithm was originally introduced to use empirical p-values to find statistically significant co-location patterns [9]. SCCP defines "a co-location patterns is statistically significant at level α if the probability (p-value) of seeing, in a dataset conforming to the null hypothesis, a participation index value of C larger than or equal to the observed PI-value is not greater than α". Null hypothesis in this case would be that instances belonging to different features are distributed across the geographic space independent of each other. PI-value can be defined for a given pattern, $C = P, Q, R$, where P, Q and R have n_P, n_Q and n_R instances respectively, as follows. If n_P^C, n_Q^C and n_R^C are number of distinct instances of P, Q and R which participate in pattern C then the participation ratio can be defined for each feature as $\frac{n_P^C}{n_P}$, $\frac{n_Q^C}{n_Q}$ and $\frac{n_R^C}{n_R}$. Given these PI-value of C can be defined as the minimum participation ratio out of all three participation ratio calculated for the P, Q and R features. Once an R number of simulated datasets are generated using randomized test, the empirical $p - value = \frac{R^{\geq PI_{observed}}+1}{R+1}$. However, the effectiveness and efficiency of this approach can be dependent on the way the randomized datasets are generated when given the null hypothesis. To avoid such issues, standardized

significance tests such as χ^2 test and Fisher's exact test could be good alternatives.

More recently a novel grid transactionization mechanism was introduced to transform the spatial dataset into a transaction dataset and to use traditional support-confidence threshold values, instead of PI-value, to compute the empirical p-value [4]. This approach took extended spatial objects as well into account. However, it was less scalable, due to the fact that it performed a brute-force search on all the possible patterns to find the statistically significant ones. The main reason of this brute-force search strategy was that the statistical significance is not a monotonic property. Hence, usage of Apriori like search algorithms is not possible to cut off the search space. This has constrained the algorithm to define a fixed length of three for rules in order to deal with the high computational complexity.

To eliminate the limitations (e.g. fixed size co-location patterns) in the above transactionization-based method another algorithm, CMCStatApriori, has used a constrained version of a statistically significant association rule mining technique called StatApriori [19] on a transactionized spatial dataset [31]. This approach uses Z-scores to apprxoimate an upperbound for the p-value and employs efficient search strategies to prune the large search space. However, a more efficient and effective algorithm to StatApriori, named Kingfisher has been suggested in the recent literature to find statistically significant dependency rules using Fisher's exact test [20]. Based on this, in our work we primarily employ Fisher's exact test based statistical dependency rules in combination with a novel transactionization approach to find co-location patterns.

3.2 AGT-Fisher to Mine Co-location Patterns

We propose an improved co-location pattern mining approach, *AGT-Fisher*, based on a new grid based transactionization method and Fisher's exact test. AGT-Fisher transforms a spatial dataset into a transaction dataset and uses statistically significant dependency rule analysis techniques to find co-location patterns. Algorithmic process of AGT-Fisher is consisted of two major steps: 1) Transactionizing the spatial dataset; 2) Mining for statistically significant association rules.

3.2.1 Aggregated Grid Transactionization

Transactionization helps to transform a spatial data set consisted of extended spatial objects 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, a grid based transactionization, Non-aggregated Grid Transactionization (NGT), method was proposed [29]. This method also has certain limitations such as when there is a reference feature, the "combined effect" from multiple non-referential general features which do not overlap with each other was ignored. More specifically, the NGT method derives transactions based on the features whose buffer regions overlap a particular grid point. To elaborate this consider an example scenario given in Figure 3.1. In this example let us consider three spatial features A, B and C. A_2, B_2 and C_2 are spatial instances of those features. A static circular buffer region surrounding the spatial instances represent the area affected by them. The scenario given in Figure 3.1(a) represents an occasion where the buffer regions of all three instances intersect. However in the scenario presented in Figure 3.1(b) there is no intersection among the buffer regions of all three instances. Assume that C feature represent a patient (i.e. reference feature) and both A and B features represent some adverse environmental conditions (i.e. general features). Although instances of C is exposed to adverse conditions B and A in both scenarios, the original grid transactionzation NGT [29] is capable of capturing this relationship only when there are grid points which are overlapped by all three buffer regions, such as in Figure 3.1(a). Since there are no common overlaps, in the scenario presented in Figure 3.1(b) NGT is unable to find transactions which has all A,B and C features. Addressing this issue we propose that when given a reference feature such as C, the scenario given in Figure 3.1(b) should produce valid transactions consisting of all A,B and C features. To achieve this, we propose Aggregated Grid Transactionization (AGT) method avoiding the limitation in the previous transactionization method.

Our proposed aggregated grid transactionization procedure is outlined in Algorithm 1: *GetAGTransactions(S)*. Given a spatial dataset *S*, Algorithm 1 initially

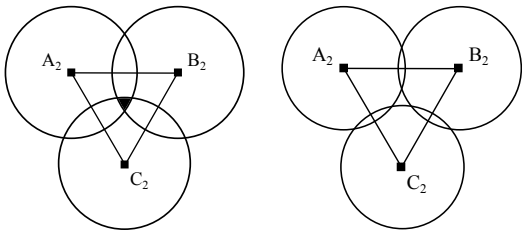


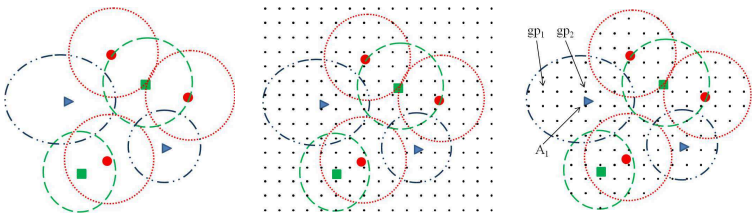
Figure 3.1: Intersection of neighboring extended spatial objects: (a) An intersection of buffer regions of feature A,B, and C exist; (b) An intersection of buffer regions of feature A,B, and C does not exist

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, then Algorithm 1 defines 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 Section 3.3 of this Chapter. 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 3.2(a) illustrates an example dataset with buffers around spatial point instances, and a grid is laid over it in Figure 3.2(b). Similarly, buffers can also be created around linear and polygonal spatial objects. In a two-dimensional space, grid points represent a square regular grid. Due to the spheroid shape of the Earth, a grid used for real-world applications becomes irregular. However, with a careful choice of a grid granularity this fact should not considerably affect the accuracy of the results as we explained in our previous works [29].

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 as shown in the Algorithm 1 (see line 5-8). 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. According to our previous work [29], with

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



(a) (b) (c)

Figure 3.2: Grid Transactionization: (a) A sample spatial dataset with point feature instances and their buffers; (b) A grid imposed over the space; (c) Grid points which intersect with buffers are used to create transactions [29]

careful consideration, we have chosen 1 km as the grid granularity level. Previous NGT approach is only consisted of the steps from 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 line 10-18). In each iteration, T_g is again grouped by the gen-

eral 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 that only overlap with the same reference feature instance. All the features in the 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. *CombEffS* number of transactions from each of the group T_{gf} is removed from the final transaction set (see line 13-16 of Algorithm 1) to balance out the support of newly added aggregated transactions.

3.2.2 Fisher's Test to Find Significant Rules

The usage of traditional association rule mining techniques, which are primarily based on the support-confidence framework, to identify co-location rules, imposes few 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 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. 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.

Fisher's exact test is a statistical significance test which can assess two categorical variables are non randomly dependent on each other or not. For instance consider the two categorical variables X and A in $X \rightarrow A$ rule. X determines either all the items in the antecedent of the given rule are present or not in a given transaction whereas the A determine either the consequent of the rule is present or

not in a given transaction. This can be represented in a 2×2 contingency table as in Table 3.1. Given this table, the hypergeometric probability of obtaining this partic-

Table 3.1: 2×2 contingency table for the X and A variables in rule $X \rightarrow A$

	A	**¬A**	**Row Total**
X	m(XA)	m(X¬A)	m(X)
¬X	m(¬XA)	m(¬X¬A)	m(¬X)
Column Total	m(A)	m(¬A)	m(X)+m(¬X)=n

ular 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 Equation 3.2.

$$p = \frac{\binom{m(X)}{m(XA)}\binom{m(\neg X)}{m(\neg X \neg A)}}{\binom{n}{m(A)}} \tag{3.2}$$

Let N_{XA} is a random variable representing the absolute frequency of XA occurring together. Dependency between X and A are stronger than observed in a given dataset if $N_{XA} > m(XA)$ where $m(XA)$ is the frequency of event XA in 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 following cumulative hyper-geometric distribution as given in Equation 3.3.

$$p_F(X \rightarrow A) = \sum_{i=0}^{J} \frac{\binom{m(X)}{m(XA)+i}\binom{m(\neg X)}{m(\neg X \neg A)+i}}{\binom{n}{m(A)+i}} \tag{3.3}$$

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. 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 can conclude 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 M($Y \rightarrow A$) is equally good or better than M($X \rightarrow A$). Here the M is a goodness measure and in our specific case it can be considered as the Fisher's p-value.

We use the above explained Fisher's exact test based framework to discover statistically significant dependency rules in a transactionized spatial dataset. The final result would yeild us a set of statistically significant co-location patterns. Kingfisher algorithm [20] 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 framework we described. Hence we use a constrained version of this implementation 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 outcome or groups such as Small for Gestational Age (SGA),LBW or PTB. Further information regarding the implementation of the search strategies, proofs and mechanisms of the Kingfisher algorithm can be found in [20]. As same in Kingfisher, in our work too we do not attempt to solve the multiple comparison issue and define a level which a rule could be called significant in a statistical sense [20]. Instead, we mainly focus on using Fisher's p-value as a goodness measure to rank and compare rules. Solutions to the multiple test could be found in [41].

3.3 Results and Evaluation

We conducted experiments to validate our proposed AGT-Fisher approach and to discover potential associations between industrial air pollutants and adverse birth outcomes to address our motivating application problem. In this section we discuss these experiments and results.

3.3.1 Datasets

We primarily worked with two sets of data: 1) Spatial data for adverse birth outcomes and industrial air pollution; and 2) Association pattern mining datasets. The first dataset was used to address our application problem and the second dataset was used to validate the statistical soundness of the rules we detected.

Adverse Birth Case Analysis Data

When addressing our motivating research question regarding the associations between air pollutants and adverse birth cases, our primary dataset is an adverse birth outcome dataset which is collected by Alberta Perinatal Health Program (APHP) during the time period of 2006-2012 from the province Alberta. We compiled this original row dataset to obtain 333,247 birth cases with their geolocations. In this dataset there are three main adverse birth outcomes of interest to researchers: 1) Preterm birth (PTB) - a birth that takes place more than three weeks before the baby is due (22,733 cases); 2) Low birth weight at term (LBW) - cases when the weight of the baby is less than 2500g and the gestational age is on or above 37 weeks (5,485); 3) Small for Gestational Age (SGA) - those whose weight is on or below 10^{th} percentile for the gestational age according to Kramer statistics [27] (29,679 cases); and 4) Pregnancy outcome - describing the birth is a still birth or a live birth.

To obtain the air pollutant information in Alberta, we used the datasets from the National Pollutant Release Inventory (NPRI) of Canada [12]. We only considered the air pollutant emissions from each of the industrial facilities within the time period of 2005-2012. This dataset contains data on estimated yearly releases of 60 chemicals/pollutants in approximately 1400 locations. The minimum and maximum average yearly release of any chemical in the dataset is 1 kg and 85,000 tonnes, respectively.

Finally, to model the air pollutant dispersion and to extend chemical release points to buffer regions we used weather data from Environment Canada. In particular we are interested in historical data for wind speed and direction in Canada. We obtained this data from 26 monitoring stations in Alberta from Environment Canada.

Association Pattern Analysis Data

In the statistically significant rule analysis literature, publicly available association pattern analysis datasets have been used to validate the quality of the rules found [20]. Similarly, we also use association rule mining datasets from FIMI Repository

[1] to evaluate the capability of our approach to find statistical sound patterns which would also hold significant in unseen data. Table 3.2 presents a summary of the FIMI datasets we used. These datasets vary in their characteristics and used by other researchers in the domain to test the robustness of frequent itemset mining methods. Hence, given a association rule mining method which performs well in these datasets, it could be expected to perform well in other transaction datasets as well.

Table 3.2: FIMI Datasets: n=no. of rows, k=no. of items, tlen=avg. transaction length

Data	**n**	**k**	**tlen**
Chess	3196	75	37.0
Mushroom	8124	119	23.0
T10I4D100K	100000	870	10.1
Accidents	340183	468	33.8
Pumsb	49046	2113	74.0
Retail	88162	16470	10.3

3.3.2 Preprocessing

We mainly preprocess the spatial datasets to discover associations between air pollutants and ABO cases. In this problem we deal with two types of spatial data: 1) Adverse birth outcome cases (ABOs), and 2) Chemical emission points. Originally, both of these are point spatial objects. We extend these two types of points objects to represent the maternal mobility range of adverse birth outcome cases and the dispersion of air pollutants more accurately. We define buffer regions around these spatial points as proposed in our previous work [29]. For ABO cases we define a circular buffer region with a fixed radius (e.g. 5 km) originating from the maternal geolocation. This buffer region represents the maternal mobility range during the pregnancy. The example dataset provided in Figure 3.3 is consisted of ABO cases and chemical emission points. Figure 3.3(b) visualizes the static circular buffer regions of ABO cases in contrast to the pollutant emission points.

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, 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 to consider 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. On Figure 3.3 (b) buffer zones around chemical points are based on the amount of their yearly release at that location. For example, the instance C_1 affects a larger zone than the instance C_3 which has a smaller amount of emission. As we previously presented [29] we defined the radius of these buffers as the natural logarithm function of the amount of chemicals released at the given location.

Secondly, we consider the wind speed and direction when modelling the dispersion of a chemical. In particular we assume that the original circle buffer will be morphed into an elliptical buffer region based on the average wind speed and direction in that location. Figure 3.3(c) depicts these elliptical buffer regions. This model is further explained in our previous work [29]. In this model we assume that the area affected by the pollutant is the same irrespective of the wind speed and direction. However the affected region can be different based on the wind speed and the direction. We interpolate the wind fields data from Environment Canada, as we suggested in our previous study [29], to obtain the average wind speed and direction in chemical emission points. Subsequently, 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.

$$a = r + \gamma|\vec{v}|, \tag{3.4}$$

$$b = \frac{r^2}{a}, \tag{3.5}$$

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. This

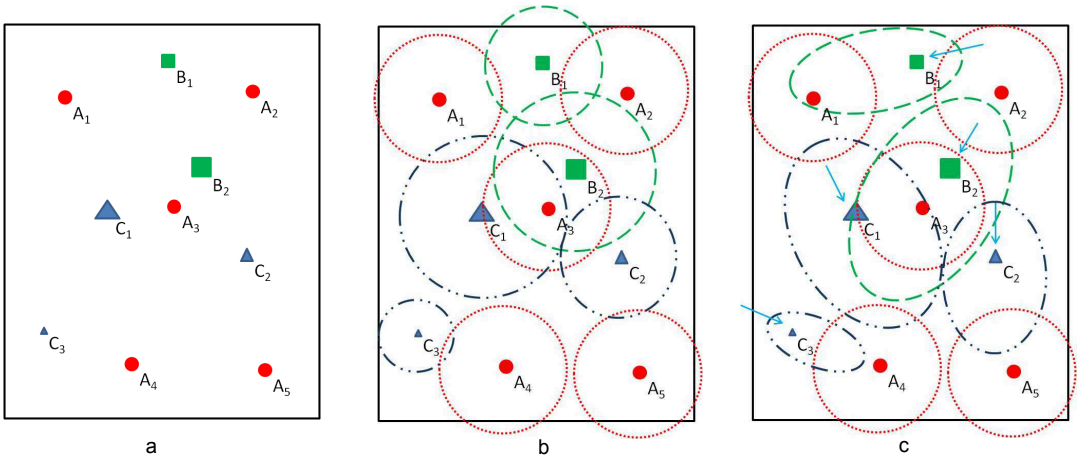


Figure 3.3: Extending spatial objects: (a) An example spatial dataset (A - Adverse Birth Outcomes, B and C - Pollutants); (b) Buffer sizes of pollutants vary depending on the amount of release; (c) Buffer shapes of pollutant emission points change with the wind direction and speed (as indicated by arrows) [29]

dataset can be directly used in the AGT method we described previously to obtain a transaction dataset.

3.3.3 Experimental Results

We performed the AGT method on the above preprocessed spatial dataset consisting of ABO cases and industrial air pollutant emission cases in Alberta and obtained a transaction dataset. The grid granularity measure we used was 1 km. Then we applied the constrained Kingfisher algorithm to obtain statistically significant dependency rules from the above transaction dataset. We used a level of significance of 3.7×10^{-44} to obtain a set of 237 very significant co-location patterns. We recommend to relax this level of significance up until 0.05. However the final decision should be based on the application problem and the expert opinions. A summary of findings is presented in the Table 3.3.

HCL, Xylene, toluene, Isopropanol and Chromimum or pollutants under the heavy metal category are frequently presented in the antecedents of patterns we found (at least in a 10% of the patterns). Some interesting co-location patterns we discovered for Alberta are $Chromimum$ & $Toluene \rightarrow PTB$, $SulfurDioxide$ & $Chromimum \rightarrow LBW$ and $Chromium \rightarrow SGA$. However, in APHP data, we have not considered widely known air pollutants such as NO_2 and Particulate

Table 3.3: Summary of the co-location patterns found in APHP data

ABO Variable	Results	Most Common Antecedents
Pregnancy Outcome	53 rules were detected indicating pregnancy outcome is a stillbirth	Hydrochloric Acid, Xylenes, Toluene, 2-Butoxyethanol, Chromium metallic, Isopropanol, Ethylene
LBW	22 rules were detected indicating a low birth weight	Hydrochloric Acid, Chromium metallic, Toluene, Isopropanol, Xylenes
SGA	65 rules were detected for small for gestational age variable	Hydrochloric Acid, Chromium metallic, Isoporpanol, 2-Butoxyethanol, Toluene, Xylenes, Ethylene
PTB	97 rules were detected for preterm birth variable	Hydrochloric Acid, Chromium metallic, Toluene, Xylenes, Isopropanol, Ethylene, 2-Butoxyethanol

Matter (PM) which causes adverse health effects. In our city wide analysis with CNN dataset (refer to Chapter 4) we expanded the industrial air pollutant set under study to include these important pollutants.

3.3.4 Evaluation

We evaluate our proposed approach to find statistically significant co-location patterns under two different aspects. Firstly, we evaluate the internal mechanisms of the approach to produce quality patterns and secondly we empirically investigate the quality of the industrial air pollutant and adverse birth case associations we found, with the help of DoMiNO experts from the application domain.

Effect of Aggregated Grid Transactionization

We compared our AGT transactionization method with a grid transactionization method without aggregation [29] to evaluate the effectiveness of our proposed approach. We used APHP ABO dataset with the level of significance 0.05 to discover co-location patterns with AGT-Fisher and NGT-Fisher (Non Aggregated Grid

Transactionization with Fisher's Test). A summary of our results is given in Table 3.4. We used *lift*, as shown in Equation 3.6, to measure the quality of the two

Table 3.4: Effect of Aggregated Grid Transactionization

Method	**Transactions**	**Rules**	**Avg. Lift**
AGT-Fisher	30200	3594	16.9
NGT-Fisher	31412	1270	8.9

different rules set. Lift can measure the dependency between the antecedent and the consequent of the rule. If the lift is 1 it means that the antecedent and the consequent are independent of each other whereas if it is larger than one they are dependent on each other. The average lift of the rules found when NGT was used is lower than the average lift of the rules when AGT was used. There are 1119 common patterns found by both methods. The average lift of those is 14.49. This indicates that the rules NGT shares with AGT method are statistically sound ones. Moreover, the average lift of the rules, discovered only when AGT method is used is 17.98. This clearly indicates that the AGT method can yield statistically sound rules.

$$lift(X \rightarrow A) = \frac{support(X \cup A)}{support(X) \times Support(A)} \tag{3.6}$$

$$leverage(X \rightarrow A) = support(X \cup A) - (support(X) \times Support(A) \tag{3.7}$$

Effect of Statistically Significant Dependencies

The objective of our approach is not only to discover patterns prevalent in observed dataset, but also to find patterns which are significant in unseen data. To test this we carried out a set of experiments using the association rule mining dataset from FIMI repository. We used the following cross-validation scheme: Each dataset was partitioned 5 times to a training set and a testing set. In each of these 5 times, we used 2/3 of the randomly picked data as the training dataset and 1/3 of the randomly picked data as the testing dataset. In each training set we used Fisher's test to find 100 best rules, with the constrained Kingfisher algorithm. The level of significance used is 0.05. This set of 100 rules were then evaluated in testing set for their quality and statistical soundness. We used lift (see Equation 3.6) and leverage (see Equation 3.7) to assess the statistical dependency of the rules in the testing

data. In this experiment, root mean squared error of lift ($RMSE_{lift}$) and leverage ($RMSE_{lift}$) can quantify the quality of our approach. If the discovered rules are equally strong in the future/test data, then the $RMSE_{lift}$ and $RMSE_{leverage}$ values should be close to zero. On the other hand larger values indicate that the dependency can either be significantly stronger or weaker. A summary of our results is shown in Table 3.5. Our results have shown that in four out of six datasets we obtained near optimal values (i.e. close to zero) for the $RMSE_{lift}$ and for $RMSE_{leverage}$ we achieved near optimal values in all the datasets indicating that the statistical dependencies of the rules found will hold in future data as well. In Table 3.5 we also report the average support, confidence, lift and leverage in training data as well. A similar set of experiments performed in one of the previous works [20] proved that Fisher's exact test achieves better RMSEs in lift and leverage than χ^2 test. Although χ^2 test helped to achieve better average lift values in training rules, the RMSE is very high in most of the datasets, indicating that exaggerated training lift values might not hold strong in unseen data. Above study concludes that Fisher's test is more robust and effective than the χ^2 test or z-score based tests (such as the ones used in StatApriori [19, 31]) in finding statistically significant dependency rules.

Table 3.5: Summary of the evaluation in FIMI data

Dataset	**Sup.**	**Conf.**	**Lift**	**Leverage**	$RMSE_{lift}$	$RMSE_{leverage}$
Mushroom	0.2213	0.9701	4.4349	0.1678	0.2017	0.0058
Chess	0.3558	0.6952	1.9933	0.1502	0.0842	0.0071
T10I4D100K	0.0073	0.8867	68.2132	0.0073	5.4311	0.0007
Accidents	0.2394	0.8695	3.8535	0.1112	0.3199	0.0007
Pumsb	0.4492	0.9889	2.2726	0.2369	0.0304	0.0011
Retail	0.0097	0.6142	137.4570	0.0043	39.1123	0.0006

Expert and Empirical Evaluation

We evaluate the relevance of the co-location patterns discovered, regarding the adverse birth outcomes and the industrial air pollutants, with the help of experts in environmental health and pediatrics. Most of the co-location patterns discovered with AGT-Fisher indicate the involvement of chemicals, such as Xylene and Toluene,

that are proven to cause various health hazards [25]. On the other hand heavy metals such as Chromium, Cadmium, Lead and Arsenic are well known carcinogens and toxic air pollutants to cause various health hazards including Cancer [39, 24]. In our results we also discovered that many patterns associate heavy metals with adverse birth outcomes. When presented our results to the experts in the domain and in the DoMiNO from environmental health and related disciplines, it is concluded that many patterns discovered are worth further investigating.

3.4 Discussion

In this chapter we presented our approach to use statistically significant dependencies to find significant co-location rules. In our approach we introduced a novel grid transactionization mechanism and discussed the applicability of Fisher's exact test to find statistically significant co-location patterns. In our experiments we showed that our proposed grid-transactionization can be helpful in finding statistically sound rules better and the rules found with Fisher's exact test can be statistically significant in not only in observed data, but also in future data as well. This proposed AGT-Fisher approach function as the foundation to the analytical methods we introduce in Chapter 4.

In grid transactionization, a particular grid point might capture more than one buffer region for a particular feature, implying more contribution from that feature than the other features in that particular location. On the other hand in a probabilistic dataset, an existential probablity might be assigned to a particular feature to indicate its impact on a specific grid point. In our proposed AGT approach, we do not assign weights to items in a transaction to represent such probabilities or recurrent items. We simply consider each item present in a transaction holds an equal weight in that transaction. However, the usage of the weighted itemsets or accounting for the recurrence of items could improve the accuracy of the model and can better represent the real world conditions [46]. The application of Fisher's exact test to handle such probabilistic or recurrent datasets should be further investigated.

The rules we discovered to address our motivating problem present some in-

teresting patterns to investigate further. However the results are subjected to the level of significance used in the experiments. Determining the level of significance should be done with the help of the domain experts. Based on this number of rules discovered can be varied. However this does not affect the efficiency or the scalability of the algorithm.

In next chapter we discuss how the patterns we discovered using AGT-Fisher can be used to discover more advanced types of spatial patterns such as spatial contrast sets and spatial common sets.

Chapter 4

Spatial Contrast and Common Sets

To address our motivating application problem of finding associations between air pollutants and adverse birth outcomes, we work with different levels of spatial data varying from provincial to national. Specifically in our collaboration with the CNN team, we work with 21 cities across all Canada. Although the co-location patterns we discovered in the previous chapter can explain which combination of industrial air pollutants are co-located with adverse birth outcomes hinting possible associations in a given spatial region, they are inadequate to understand the similarities and differences between multiple spatial regions such as the cities in the CNN dataset. To address this, we introduce two novel spatial patterns called spatial contrast sets and spatial common sets, and techniques to mine them. Furthermore we present our experimental results and evaluation of the techniques.

4.1 Background

Some of the statistically significant co-location rules we detected using the proposed AGT-Fisher approach for various spatial regions could be used to uniquely characterize and contrast a particular spatial group (e.g. preterm birth cases in Toronto or low birth weight cases in Vancouver) from the others. On the other hand some co-location rules can be useful to represent patterns which are consistently statistically significant across various spatial regions. The former type of rules are useful in identifying unique patterns specific to various spatial groups while the latter helps to recognize global co-location patterns which can be generally seen

in many spatial groups of interest. In this context, spatial groups can be defined as mutually exclusive groups represented by a particular class and associated with a particular spatial location. Preterm birth cases in Vancouver, Low birth weight cases in Edmonton, and Small for gestational age cases in Hamilton can be considered as some of the spatial groups existing in our motivating application. Especially, the first type of rules are useful to discover associations between air pollutants and some ABOs, which are specific to a particular spatial region leading to take necessary actions to handle the condition locally. On the other hand, the second type of co-location patterns are useful to recognize globally common co-location patterns between industrial air pollutants and ABOs leading to take necessary actions and creating policies to affect a larger spatial region or a country as a whole. Towards this goal we further analyze the co-location rules we detected with AGT-Fisher under the following two pattern classes: 1) Spatial contrast sets: to identify unique patterns which can characterize or contrast a particular group in a given spatial region; and 2) Spatial common sets: to identify patterns which can commonly be seen across various spatial regions/groups.

4.1.1 Problem Definition

As we explained in Chapter 2, 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 sets of 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 and non-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* $G_{x,y} \in G^s$ *and* $G_{x,y} = \{C_x, L_y\}$*;* $C_x \in C$ *is the class membership and* $L_y \in L$ *is the location of the group. Further-*

more, 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 4.1 and 4.2 holds $\forall G_{p,q} \in G^s \setminus \{G_{x,y}\}$.

$$p_F(X \rightarrow G_{x,y}) \leq p_F(X \rightarrow G_{p,q}) \tag{4.1}$$

$$\max_{p,q} |support(X, G_{x,y}) - support(X, G_{p,q})| \geq min\_dev \tag{4.2}$$

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 belongs to the same spatial region but in different classes.

2. If we fix that $\forall x = p$ we can contrast data which is in 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.

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 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 for all groups $G_{x,y} \in G^{s'}, G_{p,q} \in G^{s'}$ the constraints given in Equation 4.3 and 4.4 can be sufficed and the $|G^{s'}| > MinFrac$ threshold where the $G^{s'} \subset G^s$.

$$p_F(X \rightarrow G_{x,y}) - p_F(X \rightarrow G_{p,q}) \leq max\_pF\_diff \tag{4.3}$$

$$|support(X, G_{x,y}) - support(X, G_{p,q})| \leq max\_dev \tag{4.4}$$

max_pF_diff is a user defined threshold to control the variation of 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. The first constraint makes sure that the *statistical significance* of the common set does not vary significantly across spatial groups. The second condition makes sure that the *support* of the spatial 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 belongs to the same spatial regions but different classes.

2. If we fix that $\forall x = p$ we can find patterns common in data which belongs to different spatial regions but in 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.

4.1.2 Related Work

In the contrast set mining literature, spatial contrast sets mining is an under explored area and no significant prior work has been done. However, some work exist in the emerging pattern mining literature to find emerging patterns in spatial data.

One approach proposes a multi-relational approach to find emerging patterns in spatial datasets [13]. This approach assumes that spatial data are given in a relational database with a schema defining the spatial relationships between the tuples in the tables. Then they use a frequent pattern mining based emerging pattern discovery program to discover patterns which have a higher growth rate in one class than the other. However this approach does not attempt to distinguish between spatial groups in terms of the spatial region combined with the class. Instead it assumes that all the data belongs to the same spatial region and focuses on characterizing classes in that region. This contrasts with our definition of spatial groups as we explained previously. In addition, this approach, similar to emerging pattern discovery, completely is based on frequency based thresholds and any statistical dependencies between groups and patterns are ignored. Another approach was proposed to mine emerging patterns for binary classes in spatial data [15]. They use a reinforcement learning based method with growth rate to find optimal emerging patterns. However, as the other emerging pattern mining approaches, the above method also follows a frequency based approach and targets goals different from ours.

Common sets mining in general is an underexplored area in the literature. No significant prior work has been performed to define the problem, let alone devising techniques to find such patterns. However, lessons learned from contrast set mining literature can be largely exploited to develop common set mining techniques.

4.2 Spatial Contrast/Common Set Mining Algorithms

Addressing limitations in the previous approaches, we propose, two novel techniques to efficiently mine statistically significant spatial contrast and common sets. As similar to an association rule based contrast set mining approach, our proposed spatial contrast/common set mining techniques are based on the AGT-Fisher co-location pattern mining approach we proposed in the previous chapter.

4.2.1 DiSConS: Discovering Spatial Contrast Sets

To discover statistically significant spatial contrast sets we propose a novel algorithm, DiSConS (**Di**scovering **S**patial **Con**trast **S**ets). A pseudo code of the DiSConS algorithm is provided in Algorithm 2. DiSConS first generates all the classification co-location rules (i.e. counterpart of the classification association rules in the spatial context) of the form $X \rightarrow G_{c_i,l}$ for each location $l \in L$ using AGT-Fisher. Only the subset of the dataset belonging to the spatial region l is used when the AGT-Fisher procedure is invoked. Evaluation metrics of interest (e.g. Fisher's p-value p_F and $support(X, G_{c_i,l})$) is saved for each rule, $M(X \rightarrow G_{c_i,l})$. In the next step the algorithm performs spatial contrast set mining for each spatial group. In this step we check whether a candidate contrast set of a particular spatial group satisfies the constraints provided in Equation 4.1 and 4.2 against all the other spatial groups under analysis. If these constraints are satisfied then the candidate set is added to the results as a spatial contrast set.

Algorithm 2 DiSConS

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

CANDS=2DHashTable()
for all Location l in L **do**
  $SCAR_l$ = AGT-Fisher(S_l, A, C, α)
  for all rule $X \rightarrow G_{c_i,l}$ in $SCAR_l$ **do**
    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=2DHashTable()
for all $G_{x,y}$ in G^s **do**
  $CSET[L_y][C_x]$ = [$\emptyset$]
  for all X in $CANDS[L_y][C_X]$.keys() **do**
    if $\forall\ G_{p,q} \in G^s \setminus \{G_{x,y}\}$ Equation 4.1 and 4.2 are TRUE
      $CSET[L_y][C_x].append(X)$
  end for
end for
RETURN $CSET$

4.2.2 DiSComS: Discovering Spatial Common Sets

To discover statistically significant spatial common sets we propose a novel algorithm, DiSComS (**Di**scovering **S**patial **Com**mon **S**ets). A pseudo code of the DiSComS algorithm is provided in Algorithm 3. Similar to DiSConS, DiSComS also first generates all the classification co-location rules of the form $X \rightarrow G_{c_i,l}$ for each location $l \in L$ applying AGT-Fisher in the subset of the dataset belonging to that spatial region. Antecedents of each of the retrieved rule are added to the candidate spatial common set pool. Evaluation metrics of interest is also saved for each retrieved rule. In the next step, the algorithm performs spatial common set mining . For each candidate common set, the algorithm searches for subsets of the spatial group set of which the member spatial groups can suffice the Equation 4.3 and 4.4. If any such subset exists where the size of that subset is larger than or equal to a MinFrac fraction of all the spatial groups under analysis, then the candidate is added to the final results set.

Algorithm 3 DiSComS

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

CANDS=2DHashTable()
CANDP= $\emptyset$
for all Location *l* in *L* **do**
  $SCAR_l$ = AGT-Fisher(S_l, A, C, α)
  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 4.3 and 4.4 is TRUE$\}|$
  if $\frac{GrCnt}{|G^s|} \geq MinFrac$
    CSET = CSET $\cup$ X
end for
RETURN $CSET$

4.3 Results and Evaluation

We carried out a set of experiments to evaluate the DiSConS and DiSComS algorithms as well as to discover discriminating or common air-pollutant-ABO co-location patterns across 21 Canadian cities, using those algorithms. In this chapter we discuss these experiments and the results we obtained.

4.3.1 Datasets

We primarily worked with two sets of data: 1) Spatial data for adverse birth outcomes and industrial air pollution; and 2) Contrast set mining datasets. We used the first dataset to find interesting patterns while addressing our application problem and the second dataset to validate the analytical methods we use.

Adverse Birth Case Analysis Data

As part of the attempt to address our motivating application problem we investigated on discriminative/common air-pollutant-abo co-location patterns in 21 Canadian cities using our proposed spatial contrast and common set mining algorithms. In this investigation we primarily used adverse birth outcome datasets from CNN. CNN adverse birth outcome dataset is collected during the time period of 2006-2010 from NICUs (Neonatal Intensive Care Units) across 21 cities in Canada. We compiled this original row dataset to obtain 32,836 adverse birth cases with their geolocations. We grouped this dataset according to 19 Census Metropolitan Areas (CMAs) in Canada. In this dataset there are three main adverse birth outcomes of interest to researchers: 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 CMAs of interest in Canada, we used the datasets from the National Pollutant Release Inventory (NPRI) [12] of Canada. More specifically we chose industrial facilities within the 100 km radius of each of the Census Metropolitan Area (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. The minimum and maximum average yearly release of any chemical in the dataset is 1 kg and 85,000 tonnes, respectively.

Finally, to model the air pollutant dispersion and to extend chemical release points to regions we used weather data from Environment Canada. In particular we are interested in historical data for wind speed and direction in Canada. We obtained this data from 47 National Air Pollutant Surveillance stations in Canada.

Contrast Set Mining Data

To validate the capability of statistically significant dependencies to function as better candidate contrast sets, we have used 16 association pattern mining datasets from the UCI Machine Learning Repository [3]. All the instances in these datasets are labeled and belong to multiple classes. The minimum number of classes in any dataset is 2 and the maximum number of classes in any dataset is 10.

4.3.2 Preprocessing

As in the APHP dataset, used in the experiments in Chapter 3, in the CNN dataset we also deal with two types of point spatial data: 1) Adverse birth outcome cases (ABOs), and 2) Chemical emission points. Originally, both of these are point spatial objects. We followed the same procedure explained in Chapter 3 to extend these spatial objects to represent the maternal mobility regions and the distribution of the industrial air pollutants. However, although we proposed to interpolate the wind fields data from Environment Canada in Chapter 3 and in our previous study [29] to obtain the average wind speed and direction in chemical emission points; in CNN dataset, due to insufficient data, an interpolation is not possible. Hence we simply attribute the wind speed and direction (i.e. $\vec{v}$) of the nearest weather station to each of the chemical emitting facilities respectively.

We also have preprocessed all the contrast set mining datasets from UCI ML repository by discretizing their numerical attributes using the LUCS -KDD Software Library [2].

4.3.3 Experimental Results

According to DiSConS and DiSComS, we initially applied the AGT-Fisher method on each one of the subsets of the air-pollutant-abo spatial data belonging to CMAs. This meant that we perform the aggregated grid transactionization and co-location pattern discovery on each of the CMA mutually exclusively. The identified co-location patterns are of the form $X \rightarrow ABO_i$ where $ABO_i \in SGA, PTB, LBW$ 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 4.1.

Table 4.1: Summary of the rules found with AGT-Fisher in CMAs in Canada

CMANAME	# SGA Rules	# PTB Rules	# LBW Rules	# Total Rules
Calgary	84	77	98	259
Edmonton	109	128	139	376
Fredericton	9	1	16	26
Halifax	111	92	142	345
Hamilton	2076	2181	2254	6511
Kingston	20	17	22	59
London	311	315	482	1108
Moncton	12	12	4	28
Montreal	230	236	305	771
OttawaGatineau	57	71	130	258
Quebec	89	120	0	209
Regina	83	85	82	250
Saint John	137	157	139	433
Saskatoon	53	53	50	156
St. John's	5	5	3	13
Toronto	730	846	734	2310
Vancouver	104	81	98	283
Victoria	0	0	4	4
Winnipeg	180	110	191	481

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) minimum number of rules, 4, were obtained 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 1797 co-location rules from all the

rules from 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, Xylenes, $PM_{2.5}$ (Particulate Matter$\leq$2.5 microns) and PM_{10} (Particulate Matter$\leq$10 microns), Arsenic, 2-Butoxyethanol and Isopropanol.

Spatial Contrast Sets

Based on the location set L consisting of 19 CMAs and the class set C consisting of three ABOs, using DiSConS, we discovered two types of interesting spatial contrast sets out of the three described previously. Those two types are as follows.

1. Patterns contrasting ABO groups in the same location

2. Patterns contrasting same ABO in different locations

As an example for the first type of spatial contrast sets let us consider the CMA of Vancouver. In Vancouver, PTB has only two contrast sets out of all the 81 unique antecedents (2.4%) in the candidate rules found (i.e. X in the rule $X \rightarrow PTB$), which contrast PTB cases from LBW and SGA cases in Vancouver. Those two contrast sets are: {Methanol & Toluene & Isopropanol & CO} and {Toluene & Isopropanol & CO}. The significant reduction in patterns (i.e. from 81 to 2) 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 98 (2.0%) air pollutant itemsets in Vancouver. Those two are as follows: {Methanol & Toluene & NO_2} and {PM (Total Particulate Matter) & Cadmium}. Similarly, these contrast sets can be reported for other CMAs with all three ABOs as well.

On the other hand for the second type of spatial contrast sets let us consider the CMA of Vancouver and 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 significant patterns (6.1%). Some of them can be reported as follows:{Methanol & NO_2 & Benzene}, {Benzene & CO & PM_{10}}, and {Benzene & Methanol & $PM_{2.5}$}. These five sets can contrast the PTB cases in Vancouver from the PTB cases in other CMAs. On the other hand for PTB cases in Calgary, we

detect eight sets out of 77 significant patterns (10.3%) which can contrast the group from PTB cases in other CMAs such as Vancouver and Toronto. Some of those sets are as follows: {Asbestos}, {2-Butoxyethanol, Toluene,CO}, and {NO_2, Isopropanol, Xylenes}. 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 results set.

Spatial Common Sets

Similar to spatial contrast sets, based on the location set L and the class set C, we focus on discovering a single type of interesting spatial common sets out of the three types described previously, using DiSComS algorithm (i.e. to find common sets for a particular adverse birth outcome in different CMAs). To find such spatial sets, in addition to the MaxSig threshold we also use a MinFrac threshold of 0.3 (30%) to specify at 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 found 42 spatial common sets which are associated with PTB cases in at least 30% of the CMAs. One top such 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, Edmonton, Montreal, etc. Other than that, in this 42 sets, interesting spatial common sets such as {PM_{10}, CO}, {Total Particulate Matter, CO}, {$PM_{2.5}$}, {NO_2, Total Particulate Matter}, {Arsenic and arsenic compounds}, {Toluene}, and {Xylene} exist. We observed that these common sets are also commonly associated with other ABO types as well.

4.3.4 Evaluation

We evaluate our algorithms based on the analytical methods we used and the knowledge they discover. Two main analytical methods we used in our proposed spatial contrast and common set mining methods are, aggregated grid transactionization and the Fisher's exact test to find statistically significant dependencies. We evaluated the effectiveness of the above methods in the previous chapter. However,

in DiSConS and DiSComS algorithms we use the statistically significance dependencies found with AGT-Fisher to discover contrast set and common set in spatial data. Hence, in this chapter we evaluate the effectiveness of statistically significant dependencies to find contrast sets. Furthermore, we evaluate the quality of discovered spatial contrast and common sets in CNN dataset with empirical evidence and expert opinions.

Effect of Statistical Significant Dependencies to Find Contrast Sets

To investigate the applicability of the statistically significant dependencies to find contrast sets, we used the contrast sets obtained by applying DiSConS on 16 datasets fron the UCI-ML repository. Since, these datasets have no spatial attributes, we skipped the transactionization process in the DiSConS algorithm and found contrast sets for groups of the type 1 (i.e. contrasting data belong to the same spatial region but different classes) to build an associative classifier, CS^2 (**C**lassification based on **S**tatistically Significant **C**ontrast **S**ets), similar as CBA [32]. Bypassing transactionization procedure and searching for contrast sets of type 1 effectively allows to find contrast sets for non spatial datasets using DiSConS algorithm. Having a better classification accuracy with CS^2 than with CBA would mean that we have identified contrast sets which can meaningfully differentiate classes than the general association rules. As shown in Algorithm 4 (line 3-6), CS^2 first recognizes the subset of contrast sets which can contribute to classify a given data instance. Then, based on the class, it categorizes this subset of rules. Aggregate function *sum* can be applied to each of these categories to obtain a representative measure for each class. Subsequently a class can be assigned to the data instance by using another aggregate function *min* on the representative measures.

We have compared the classification accuracy of CS^2 with several other standard classifiers on UCI datasets to evaluate the quality of our resulting contrast sets. However, we emphasize that our target is not to build a classifier and outperform the accuracy of dedicated classifiers. Instead, our goal is to achieve a moderate or similar accuracy compared to the standard classifiers to indicate that we discover accurate contrast sets. In our experiments we used a level of significance of 0.05

Algorithm 4 CS^2

INPUT: Database D, Object O, Attributes A, Classes C, Level-of-Significance α

CSet = DiSConS(D, A, C, $\emptyset$, α, C)
CSet$_{new}$ = $\emptyset$
for all c-set *c* in *CSet* **do**
  if $c.antecedent \subseteq O.antecedent$
    CSet$_{new}$ = CSet$_{new}$ $\cup$ c
end for
Divide CSet$_{new}$ to subsets based on class labels: $S_1, S_2, ...S_n$
for all S_i in $S_1, S_2, ...S_n$ **do**
  sum all the *ln* p_F values in each subset
end for
Assign the class with lowest some of p_F to O

RETURN $O.label$

and a minimum groups support difference of 10%. The classification accuracies of the methods we used in our experiments are reported in Table 4.2. We chose rule based classifiers, C4.5 [35], CBA [32] and CPAR [45], to compare. Average accuracy in all 16 datasets indicates that CS^2 outperforms one out of three other standard rule based classifiers while having a very close accuracy to the other two. This is a strong indication that statistically significant dependencies can provide quality contrast sets.

Expert and Empirical 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 from the DoMiNO team at the University of Alberta. Most of the studies in the literature emphasises the involvement of 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) [28, 11] with adverse birth outcomes such as SGA, PTB and LBW. Most of the spatial common sets we discovered for CMAs in Canada include these air pollutants within them. This provides a good indication to the quality of the patterns we discovered. Furthermore we presented our discovered contrast patterns to the domain experts who validated that some of the rules we discovered are interesting and worth further investigation. In particular, we found that

Table 4.2: Comparison of classification results: C4.5, CBA, CPAR and CS^2.

Dataset	#**cls**	#**rec**	**C4.5**	**CBA**	CPAR	CS^2
adult	2	48842	78.8	**84.2**	77.3	80.88
anneal	6	898	76.7	94.5	**95.1**	83.4
breast	2	699	91.5	**94.1**	93.0	80.98
flare	9	1389	82.1	**84.2**	63.9	79.99
glass	7	214	65.9	68.4	64.9	**69.18**
heart	5	303	**61.5**	57.8	53.8	57.16
hepatitis	2	155	**84.1**	42.2	75.5	81.54
horseColic	2	368	70.9	78.8	**81.2**	70.33
ionosphere	2	351	84.6	32.5	**88.9**	74.92
iris	3	150	91.3	93.3	**94.7**	94.67
led7	10	3200	**73.8**	73.1	71.3	72.75
mushroom	2	8124	92.8	46.7	**98.5**	94.97
pageBlocks	5	5473	92.0	90.9	**92.5**	89.77
pima	2	768	71.7	**74.6**	74.0	65.11
wine	3	178	75.8	49.6	88.2	**90.95**
zoo	7	101	91.0	40.7	**94.1**	94
Average			80.2	69.1	**81.7**	80.04

heavy metals such as Arsenic, Lead and Cadmium can be attributed as industrial air pollutants and potential causes for adverse birth outcomes, through the discovered patterns. These should be further investigated and studied with the help of domain experts to find them as qualitatively sufficient to formulate research hypotheses.

4.4 Discussion

We introduced two novel type of spatial patterns called spatial contrast sets and spatial common sets in this chapter. DiSConS and DiSComS are two efficient novel techniques we propose to mine those patterns. These two algorithms are built on AGT-Fisher approach we outline in Chapter 3. We evaluate our analytical methods for their effectiveness in using statistically significant dependencies to find intended patterns. Our results show that our methods can produce contrast sets which can distinguish two or more classes better.

With the help of domain experts we collaborated from the DoMiNO team, we investigated the spatial contrast and common sets we discovered to find out whether

they can form plausible research hypotheses for further investigation by environmental scientists. From our patterns, we discovered that chemicals and substances such as NO_2, Particulate Matter, CO, and heavy metals such as Lead, Cadmium and Arsenic are commonly associated with many of the adverse birth outcomes across many Canadian cities. Most of this chemicals are empirically well known to cause adverse health effects and the environmental health community carry out many studies to assess their involvement in such cases. Hence, our results conform to this existing knowledge and hypotheses. In addition to these, we also produce many patterns which could explain the adverse health effect when combination of industrial air pollutants exist. This could be really helpful in future studies for the researchers in application domain because most of the traditional studies are designed to focus on the effect of one variable at a time.

Our proposed methods are easily extensible to consider temporal information and find contrasting patterns in spatio-temporal groups as well. On the other hand usage of other measures such as leverage in place of support or lift in place of Fisher's p-value can be suggested when using the conditions in the algorithms. Interpretation of the patterns we discover can be changed based on this. Hence, by discussing with the experts in the domain, more insights can be obtained and a better model can be designed to serve the goals of the application problem.

Chapter 5

VizAR: A Visualization Framework for Co-location Patterns

In any knowledge discovery task visualization plays a key role in transferring the discovered knowledge to the end users (in our case pediatricians, medical practitioners, environmental scientists, etc.). In this chapter we propose a novel framework, **VizAR**, to visualize various spatial patterns we have explained in the previous two chapters.

5.1 Background

Visualization is traditionally recognized as any technique to create images, diagrams, maps, videos or animations to convey a message. It is a very effective tool which can help to communicate complex ideas and experimental results across disciplines since humans are more visual learners. However, with recent advancement in data science, the traditional visualization paradigm is shifting towards more interactive visualization systems to allow the end knowledge users to more actively engage in understanding knowledge discovered through the data mining process. In this chapter we explore on devising such an interactive visualization framework for spatial patterns.

5.1.1 Problem Definition

In this thesis we primarily focus on three types of spatial patterns: co-location patterns, spatial contrast sets and spatial common sets. The latter two patterns are in

fact two special cases of the former pattern type. Our motivating problem demands to work with single spatial regions such as provinces as well as multiple spatial regions such as cities or census metropolitan areas in Canada. Hence, we outline the problem of visualizing spatial patterns, considering the different levels of complexities spatial datasets could have. We define that visualizing spatial patterns in multiple spatial regions can be done under the following three abstraction levels.

1. Lay of the Land: Visualize all the co-location patterns discovered for all the spatial regions under study.

2. Pattern/Regional Level: Investigate the characteristics of a specific co-location pattern in various spatial regions or investigate all the co-location patterns in a particular spatial region.

3. Instance Level: Investigate a specific co-location pattern in a specific spatial region to understand instance distribution and the nuances of the contributing geographic factors.

When devising a visualization framework to implement the above abstraction levels an interactive approach should be followed to allow the user to navigate through various levels of interest to gain knowledge.

5.1.2 Related Work

Traditionally to analyze and visualize geographic information or spatial data, proprietary software systems with a comprehensive set of tools have been built (e.g. ESRIs ArcGIS). However when the visualization is not coupled with the analytical methods provided by such systems it is not a trivial task to visualize patterns discovered by a third party algorithm or software.

Although the association rule mining community—a closely related field to co-location pattern mining—have extensively worked on various ways to visualize association patterns, only a limited number of approaches have been proposed to visualize co-location patterns [14]. One such approach [14] propose a clustering based approach to visualize co-location patterns in a map. Most of such proposed schemes

are focused on developing comprehensive visualization schemes for instance level visualization or clustering approaches to visualize patterns as clusters of instances on a given spatial region. Such approaches can be classified as pattern/regional level of abstraction under our problem definition. This limits the levels of abstractions provided by those scheme. Moreover, user interactivity in such systems are also minimal.

On the other hand, to visualize contrast sets, few approaches have been proposed based on horizontal bar charts. An example of a such a visualization is provided in Figure 5.1. However, these approaches have limitations when it comes to visualizing contrast sets in multiple classes [34]. On the other hand there are no known work in the literature focusing on visualizing spatial contrast or common sets.

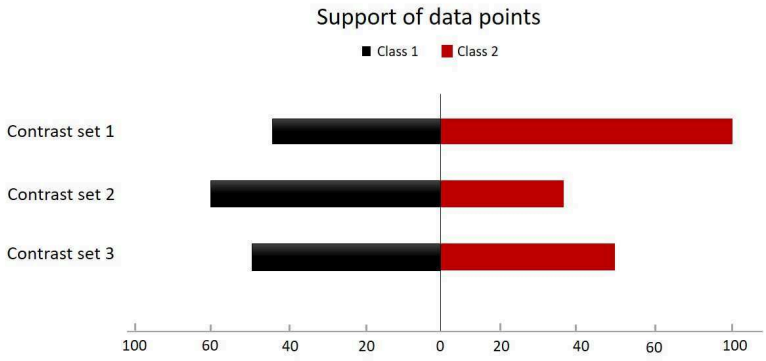


Figure 5.1: Visualizing Contrast sets with bar charts

5.2 VizAR Framework

To allow the knowledge users or other researchers to explore the patterns we discover, we propose a novel interactive visualization framework called VizAR to visualize spatial co-location, contrast and common sets. In contrast to the traditional approaches, that target pattern/regional level abstraction, VizAR provides a simplified, yet a complete visualization scheme targeting all three levels of abstractions provided in our problem definition.

5.2.1 System Design

The VizAR system consists of three main visualization modules. Each module represent one of the abstraction layers previously defined. We devise various visualization tools in each of these modules to visualize the patterns accordingly. As presented in the system design diagram in Figure 5.2, following are the three main modules of the VizAR system:

1. Overview Level
2. Pattern/Regional Level
   (a) Regional Level
   (b) Pattern Level
3. Instance Level

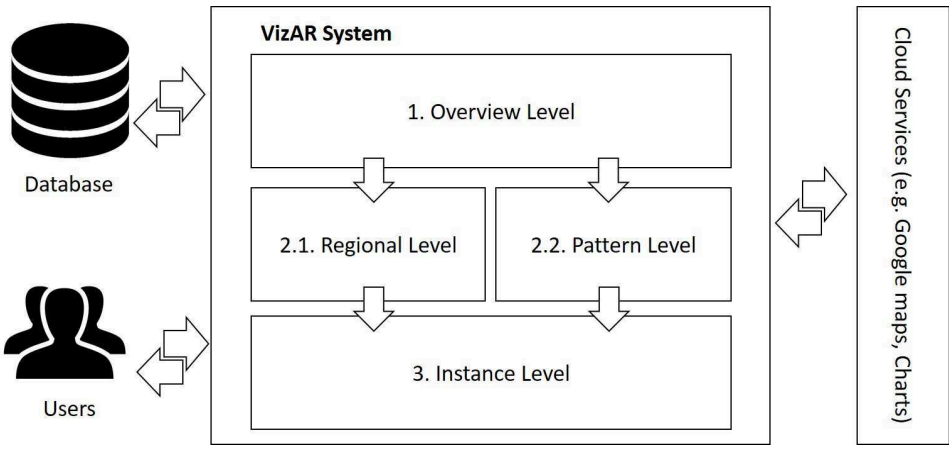


Figure 5.2: System Design of the VizAR Framework

VizAR communicates with a central database to access the patterns to visualize, the transactions which support a selected pattern, and various other meta data. It also interacts with cloud services to access various kinds of resources such as maps. User interaction with the VizAR system starts from the overview level. The overview level is designed to provide an interface to the user, presenting all the patterns for a selected spatial dataset. Various types of interfaces can be implemented to provide such overviews. It could either be a filter allowing a user to browse only a subset

of selected patterns leading to the pattern level module in the next layer or a bubble chart representing all the rules in all the spatial regions under study. Selecting a specific bubble from such a chart could lead to the regional level module in the next layer. Regional level module is designed to visualize the significance of the selected pattern in various spatial regions. From the regional or pattern level a particular pattern for a particular spatial region could be singled out to further investigate. This leads to the instance level, where the distribution of the transactions contributing to a particular patterns and the other geographic features could be closely investigated on an interactive map. In the next section we discuss few of the interfaces we implemented under these modules and how one can navigate through the patterns to discover .

5.2.2 Implementation

We implemented a prototype of the VizAR framework as a web application. The VizAR system is completely independent of the algorithms or programs used to discover the spatial patterns. The only input required to the system is the set of discovered rules, their quality measures and the set of transactions. All these can be stored and queried back from a database.

Interactive Visualization Tools for Overview Level

In our prototype we implemented two types of visualization tools for a user to start with in the overview level. The first prototype tool, as shown in Figure 5.3, provides an interactive filter to apply constraints for the antecedents set and the consequent to obtain a desired subset of the patterns. Based on the user requirements, further constraints can be implemented to target specific subsets of the patterns to investigate. This particular tool implements the idea of constraint based rule analysis. Instead of applying the constraints during the mining process we apply the constraints in the post-mining stage. Some examples of the constrains which can be implemented are, item exclusions, item inclusions, and thresholds on item recurrence or the quality measure. Once the constraints are selected the result will display a bar chart visualizing the distribution of the rules subset based on the size of the antecedent set. If

a particular bar is selected only the subset of rules with the selected antecedent set length would be displayed as the output.

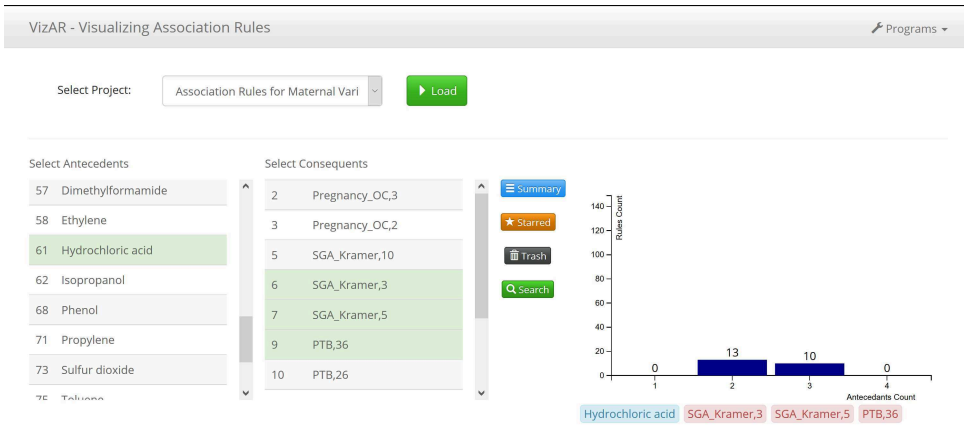


Figure 5.3: A prototype of an interactive filter to be used in the overview level

The second tool we implemented is more suitable to visualize all the patterns discovered in all the spatial regions under study (i.e. CMAs in Canada). As shown in Figure 5.4, we use bubble charts to implement a prototype for this tool. The y axis of the bubble chart represents the CMAs where the relative distances between x axes represent the approximate relative geographical distances between CMAs. More sophisticated charts, such as the ones that uses a primary vertical axis for the longitude and a secondary vertical axis for the latitude where a line drawn connecting two points in the vertical axes would represent a specific location, exist. However, since the CMAs we consider does not have drastic changes over the latitude lines, we simply considered a single vertical axis to represent the locations. The x axis in the given bubble chart represents the rules. A bubble in this overview visualization represents a co-location pattern. The size of the bubble vary according to the support of that rule in the corresponding geographic area or CMA. All the bubbles are color coded based on the statistical significance of the rule (i.e. log(p-value)). Yellow is the lowest statistical significance level where red is the highest statistical significance levels. If the chart is analyzed along the vertical lines, the support and the statistical significance variation of a specific rule in various spatial regions can be clearly observed.

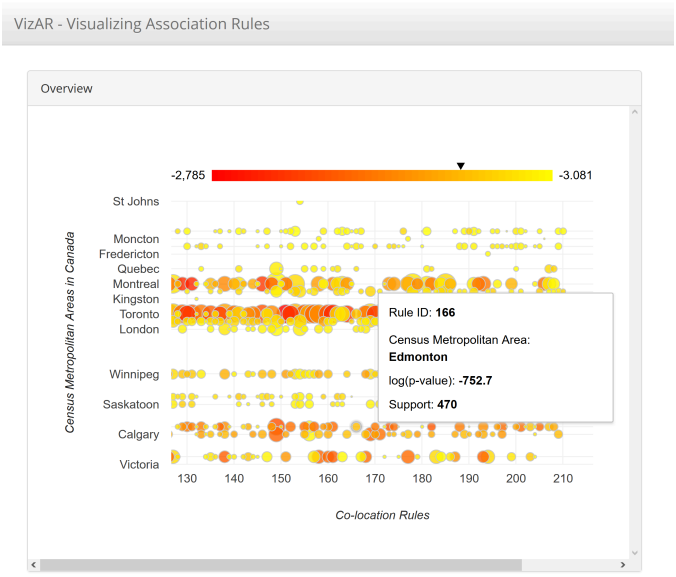


Figure 5.4: A prototype of an interactive bubble chart to be used in the overview level (each bubble represents a pattern where the size is corresponding to the support and the color is corresponding to the statistical significance)

Interactive Visualizations for Pattern/Regional Level

Once the filtering constraints are applied in the first prototype tool we described for the overview layer, it would lead to the next level where the selected subset of patterns are shown and allowed to be explored. This tool would simply provide a flat representation of the patterns in the subset. However when an individual pattern is selected it would lead to the next level where the instance distribution of that pattern is visualized. On the other hand if a bubble is clicked in the second visualization tool we explained for the overview module, it would lead to two different interactive visualization tools as shown in Figure 5.5. One visualization would represent the distribution of a selected pattern in various CMAs with the corresponding support and color coded statistical significance level. The other visualization uses radar charts to show how the support contrasts from one CMA to another.

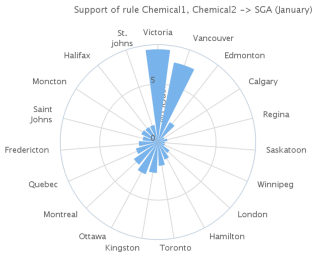
In particular radar charts are a good alternative visualization scheme to represent (spatial) contrast sets. Especially because it can distinguish the support of a particular pattern in one region/class than the others, in a more contrasting manner. Hence we further explore the usability of such chart types to discover contrast/com-

Figure 5.5: A prototype of a geochart and a radar chart to be used in the regional level

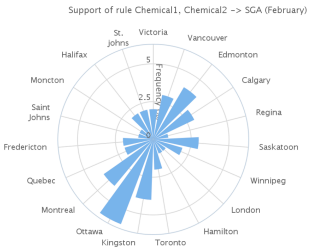
mon sets. Wind roses are an interesting type of charts, very similar to radar chart, which are used to visualize the relative frequency of wind speed at a place. These type of radar charts or wind rose graphs could be used to visualize spatial as well as temporal variation of support/strength of a particular pattern more meaningfully. For example consider a scenario where variation of the support of a particular co-location pattern of industrial air pollutants and an ABO has to be visualized for various spatial regions in different months. Based on the climate changes, impact caused by an air pollutant also can change. Hence, understanding how the strength of a spatial contrast set can change over the time is very important and can be helpful in changing environmental and industrial policies. An example for a such a scenario is given in Figure 5.6.

Interactive Visualizations for Instance Level

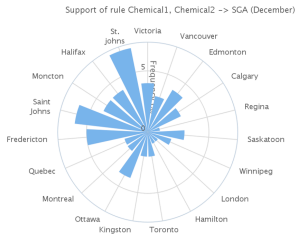
When a specific pattern for a specific spatial region is selected from any one of the visualization tools discussed in the previous level, it would lead to the final instance level visualizations. In this level, we mainly visualize the supporting transactions or data instances in a map so that their distribution can be understood well. An example is given in Figure 5.7. Furthermore, additional information such as chemical dispersion can be added to this visualization to better understand the impact of the



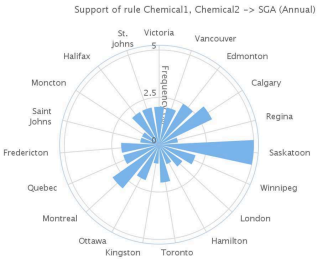
(a) Support distribution in January



(b) Support distribution in February



(c) Support distribution in December



(d) Annual Support distribution

Figure 5.6: How the support distribution of a contrast set vary across time

pattern. One such visualization is depicted in Figure 5.8. In both these visualizations green circles represent that the rule is valid in that instance, where as the red circles represent that the rule is invalid in those transactions.

5.3 Discussion

In this chapter we outlined the system design of a novel interactive visualization framework to visualize spatial patterns. We also presented our prototype visualization tools developed under this framework. These tools we devised can be effectively used to find spatial contrast sets, common sets and can be helpful in further investigating on the distribution of the instances of a particular pattern in the local region. VizAR does not depend on the rule mining technique. Any rule mining technique can be used to find co-location or association patterns for a given transaction dataset. This rules found can be used as the input to the VizAR system. Our proposed three layer level of abstraction can provide a solid outline to anyone who wishes to extend the VizAR system to a particular application problem. Under these three levels various visualization tools can be developed. However, the

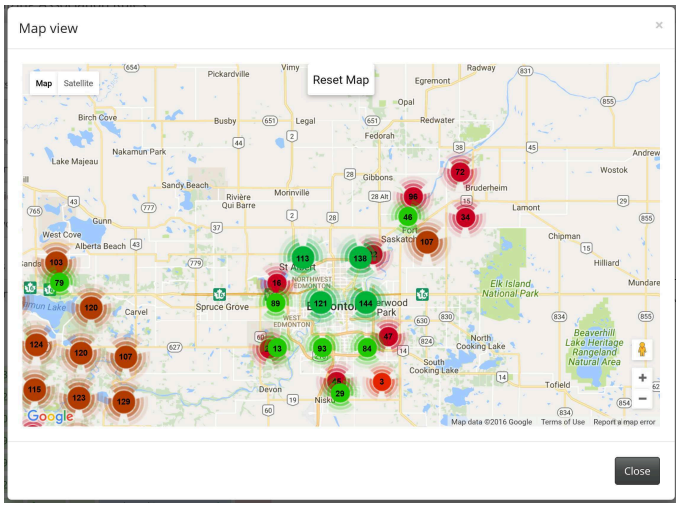


Figure 5.7: A prototype of an interactive map to be used in the instance level

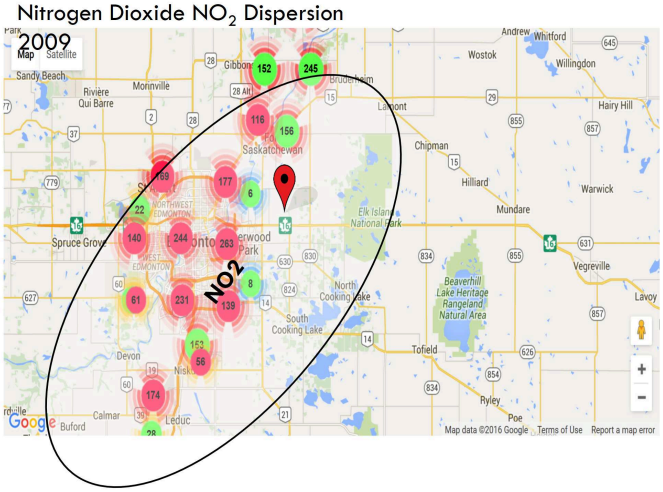


Figure 5.8: A chemical dispersion information to the instance level prototypes

usability of such tools depends on the users. We demonstrated the VizAR system to the DoMiNO team. Experts in the DoMiNO team found that VizAR is a very effective system and can be helpful in discovering answers to the research questions they have. The feedback we received from the user community could be useful to implement novel visualization tools and to improve/extend the existing functions.

Chapter 6

Conclusions and Future Work

In Chapter 1 we put forward three theses which can be summarized as, *statistically significant dependencies could be efficiently and effectively used to find valid spatial co-location patterns, contrast sets and common sets, and visualization tools can be devised to effectively explore such large space of spatial patterns*. In light of the work done, as presented in Chapter 3, 4 and 5, we revisit these theses in this chapter to arrive at conclusions and to understand the future avenues of research.

6.1 Conclusions

In this work we addressed the problem of using statistically significant dependencies to find various spatial patterns and devising visualization schemes to discover knowledge using them. In particular, we are motivated by an important problem in the environmental health domain to find spatial associations between industrial air pollutant and adverse birth outcomes. This particular problem can be transformed into a spatial pattern mining problem or more specifically to a co-location pattern mining problem. Co-location pattern mining and its counterpart association rule mining are two well-studied problems in the data mining community. However, most of the existing techniques to find such association patterns, heavily depend on the frequency based thresholds which are hard to determine and pose several drawbacks. In such methods, if a low threshold is defined, the final result would include a very large number of noisy rules whereas a stringent threshold means that rare rules could be lost. In recent years association rule mining community has shown

interest in adapting statistical significance tests to measure the quality of an association rule in terms of the strength of the dependency between the antecedents and the consequent. Subsequently, a few very recent works in co-location pattern mining also adapted statistical test based approaches to find strong co-location patterns. On the other hand, rather than developing dedicated co-location pattern mining techniques, a spatial dataset can be transformed into a general transaction dataset and an association rule mining technique could be applied to find co-location patterns. The advantages presented by this approach lead the co-location pattern mining community to adapt similar approaches when devising better co-location pattern mining algorithms. Moreover this opened up a new avenue of research to use statistically significant dependency (i.e. association) rule analysis techniques with a suitable transactionization approach to find statistically significant co-location patterns.

The increasing complexity of the spatial datasets and spatial data analysis tasks have introduced novel analytical problems to the community. For instance, in our application problem important research questions such as "what kind of air-pollutant sets are commonly co-located with a particular adverse birth outcome in major Canadian cities?" or "what kind of air-pollutant sets can distinguish the occurrence of a particular adverse birth outcome in a particular city from the others?" motivate to devise new spatial pattern mining techniques since the traditional co-location patterns alone cannot answer such questions. Contrast set mining techniques target a very similar problem to find discriminating association patterns which can differentiate data instances belonging to different classes. Despite this, spatial variants for contrast set mining techniques remains an under explored area.

Although in emerging pattern mining—a research area which shares some of the objectives with contrast set mining—techniques have been developed to consider spatial information as well, they neither attempt to address the same contrast set mining problem we are interested in nor include any statistically significant tests in their approaches. On the other hand finding common association patterns in multiple groups also remains an area that is not well explored.

Addressing the limitations in existing analytical methods to find required spatial patterns in our motivating problem, we propose a set of novel analytical methods

and tools to use statistically significant dependencies to discover statistically significant spatial co-location rules, contrast and common sets. In doing so we focused on the following four major challenges: 1) Transactionizing a spatial dataset; 2) Discovering statistically significant co-location patterns; 3) Comparing and contrasting spatial groups; and 4) Visualizing co-location patterns. We proposed a novel grid transactionization method, **AGT** (Aggregated Grid Transactionization), to address some of the limitations posed in existing spatial data transactionization approaches. A spatial dataset transactionized using AGT method can be used with Fisher's exact test to find statistically significant co-location rules. To do so we proposed **AGT-Fisher** approach. AGT-Fisher uses a constrained version of an efficient algorithm called Kingfisher to mine statistically significant dependency rules using Fisher's exact test. To compare and contrast spatial groups, we introduced two novel spatial pattern types called *spatial contrast sets* and *spatial common sets*. We defined that spatial contrast sets and spatial common sets are two special cases of the general co-location patterns. Hence, we proposed two new algorithms, **DiSConS** and **DiSComS**, to efficiently mine those patterns based on the output of the AGT-Fisher approach. DiSConS algorithm is designed to find co-location patterns which can uniquely characterize and contrast a particular spatial group from the others, whereas the DiSComS algorithm is intended to find co-location patterns that are significantly common in many spatial groups. To let the knowledge users meaningfully explore the resulting patterns we devised a novel co-location pattern visualization system called **VizAR**. All these methods and tools can be combined into a single framework, Di3SP (Discovering Statistically Significant Spatial Patterns), to use statistically significant dependencies to discover statistically significant spatial co-location rules, contrast and common sets. System architecture of the Di3SP framework is provided in Figure 6.1. Di3SP framework is designed to address four major challenges: Modules 1,2, 3 and 4, and 5 in Figure 6.1 represent the analytical methods and tools we devised to address these challenges respectively.

Our experiments revealed that our proposed AGT method can indeed aggregate a subset of the intended transactions based on a reference feature to address the "combined exposure effect". Moreover, when compared with a previous grid trans-

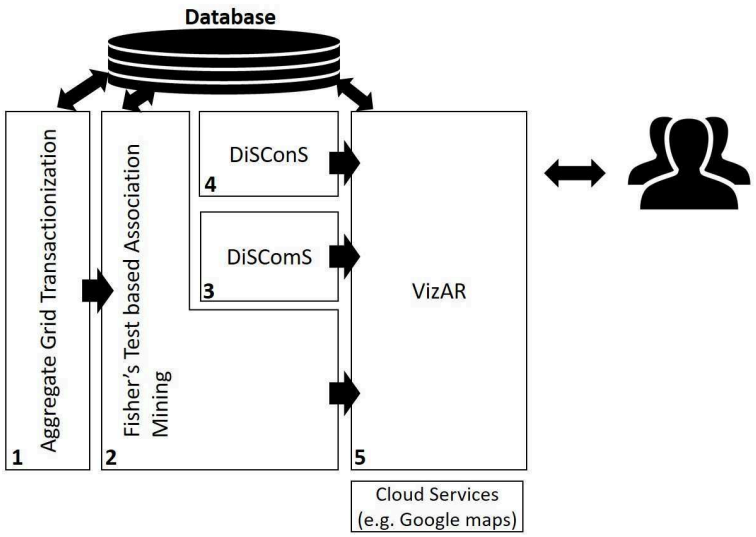


Figure 6.1: System Design of the Di3SP Framework

actionization approach AGT was able to obtain better average lift indicating the statistical dependency of the rules it finds. To evaluate whether the statistically significant dependencies can actually capture rare but statistically significant patterns which will also hold for unseen data, we performed experimentation on a set of frequent itemset mining datasets. This experimentation revealed that the rules found with Fisher's exact test can maintain a very stable lift and leverage (i.e. RMSE is very close to zero) in multiple train and test datasets. Moreover, most of the rules have a very low support threshold which would not have been feasible to retrieve if a frequency based threshold is used. These results verify our thesis 1, that statistically significant dependencies indeed can find rare, stable and significant rules which also hold true in unseen data. Furthermore, we constructed a rudimentary associative classifier based on the contrast sets we found using statistically significant dependencies to validate the effectiveness of the discriminating patterns found. When this associative classifier, CS^2, is applied to several datasets from UCI ML repository and compared against few standard rule based classifiers, it is revealed that it maintains a very close classification accuracy as other standard classifiers. We emphasize that our goal is not to build a classifier but to find patterns which can discriminate classes or groups. This result and our previous results with AGT-Fisher indeed verify our thesis 2, that statistically significant dependencies could be

used to find statistically sound contrast/common sets.

We applied our analytical methods on the provincial ABO datasets from APHP and Canada wide cities' ABO datasets from CNN to address the research questions posed by our motivating application problem. In those experiments, we discovered a number of potential and interesting air pollutant(s) associations with adverse birth outcomes. We found that air pollutants such as NO_2, Particulate Matter (PM), CO and heavy metals such as Lead, Cadmium and Arsenic are commonly associated with adverse birth outcomes in many spatial regions. This conforms to the existing knowledge in environmental health regarding the involvement of these chemicals in causing adverse health effects. In addition, the discriminating patterns we found reveal signature patterns which can uniquely characterize a specific spatial group and contrast it from the others. These patterns could be of great interest for further research as well as making local and global policies to mitigate these adverse conditions. We used a prototype of our VizAR framework during a full team meeting of the researchers of the DoMiNO project to allow the domain experts and knowledge users to browse through the discovered patterns. Their positive feedback indicates that VizAR is indeed successful in transferring the knowledge we intended. Furthermore the researchers agree that most of the patterns we found are interesting and worth further investigations. This demonstrates our third thesis which states that visualization tools can be devised to effectively explore a large number of various types of co-location patterns.

6.2 Future Research

The work we did in collaboration with the DoMiNO team is currently an on-going work. Some of the methods we introduced, tools we devised and knowledge we discovered need further investigations. Following, we list some of the current and future research avenues which we are working on and expect to investigate.

1. AGT-Fisher approach currently works with only binary transactions and choose to ignore that recurrent items might exist. On the other hand in probabilistic datasets an item might be associated with an existential probability. Al-

though there are association rule analysis methods developed to handle recurrent items and uncertain datasets, further research has to be carried out to use statistical significance tests such as Fisher's test for that purpose.

2. In AGT-Fisher we have used a level of significance of 0.05 with Fisher's test as a rule of thumb. However, a discussion need to be carried out with domain experts for what level of significance with which other threshold conditions (e.q. minimum support, lift or leverage) should be used to find spatial patterns which would be of relevance for them.

3. The contrast/common set mining algorithms we introduced can be easily generalized to include other type of rich information such as time. This would effectively allow a user to find contrast sets or common sets among more complex type of groups such as spatio-temporal groups. This has to be further investigated with a suitable application problem and datasets.

4. Although contrast sets can be evaluated with the help of an associative classifier, no recognized evaluation criterion exists for common sets. Further research has to be done to come up with an evaluation method for common sets and conduct a more comprehensive evaluation on our proposed common set discovery program.

5. The patterns we discovered should be further evaluated and investigated closely for their quality with the help of domain experts. Currently we are collaborating with researchers from the DoMiNO team who follow different GIS or epidemiological approaches to address the same problem to identify overlaps between the knowledge discovered by each different method.

6. When demonstrated a prototype of the VizAR program to the domain experts several feedback and comments were received, which could be helpful in improving the prototype. For instance, adding chemical distribution, varying the significance of the patterns according to the time, add additional filters such as item exclusions are some of them. While some of these comments

can be immediately addressed, some may require further investigations prior to execution to be carried out.

Bibliography

[1] Frequent Itemset Mining Dataset Repository. http://fimi.ua.ac.be/data/.

[2] LUCS-KDD Software Library. https://cgi.csc.liv.ac.uk/~frans/KDD/Software/.

[3] UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/.

[4] Aibek Adilmagambetov, Osmar R Zaïane, and Alvaro Osornio-Vargas. Discovering co-location patterns in datasets with extended spatial objects. In *Proceedings of the 15th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 84–96, 2013.

[5] Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

[6] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD*, pages 207–216, 1993.

[7] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference Very Large Data Bases (VLDB)*, pages 487–499, 1994.

[8] Luiza Antonie, Osmar R Zaïane, and Robert 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.

[9] Sajib Barua and Jörg Sander. Sscp: mining statistically significant co-location patterns. In *Proceedings of the 12th International Symposium on Spatial and Temporal Databases (SSTD)*, pages 2–20, 2011.

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

[11] Michael Brauer, Cornel Lencar, Lillian Tamburic, Mieke Koehoorn, Paul Demers, and Catherine Karr. A cohort study of traffic-related air pollution impacts on birth outcomes. *Environmental Health Perspectives*, 116(5):680–686, 2008.

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

[13] Michelangelo Ceci, Annalisa Appice, and Donato Malerba. Discovering emerging patterns in spatial databases: A multi-relational approach. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 390–397. Springer, 2007.

[14] Elise Desmier, Frédéric Flouvat, Dominique Gay, and Nazha Selmaoui-Folcher. A clustering-based visualization of colocation patterns. In *Proceedings of the 15th Symposium on international database engineering & applications*, pages 70–78. ACM, 2011.

[15] Wei Ding, Tomasz F Stepinski, and Josue Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*, pages 425–436. SIAM, 2009.

[16] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 43–52. ACM, 1999.

[17] Sandie Ha, Hui Hu, Dikea Roussos-Ross, Kan Haidong, Jeffrey Roth, and Xiaohui Xu. The effects of air pollution on adverse birth outcomes. *Environmental research*, 134:198–204, 2014.

[18] Wilhelmiina Hamalainen. Efficient discovery of the top-k optimal dependency rules with fisher's exact test of significance. In *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pages 196–205. IEEE, 2010.

[19] Wilhelmiina Hämäläinen. Statapriori: an efficient algorithm for searching statistically significant association rules. *Knowledge and Information Systems*, 23(3):373–399, 2010.

[20] Wilhelmiina 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.

[21] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 1–12, 2000.

[22] Robert J Hilderman and Terry Peckham. A statistically sound alternative approach to mining contrast sets. In *Proceedings of the 4th Australia Data Mining Conference (AusDM-05)*, pages 157–172, 2005.

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

[24] Lars Järup. Hazards of heavy metal contamination. *British medical bulletin*, 68(1):167–182, 2003.

[25] Reena Kandyala, Sumanth Phani C Raghavendra, and Saraswathi T Rajasekharan. Xylene: An overview of its health hazards and preventive measures. *Journal of oral and maxillofacial pathology*, 14(1):1, 2010.

[26] Petra Kralj, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. Contrast set mining for distinguishing between similar diseases. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, pages 109–118. Springer, 2007.

[27] Michael S Kramer, Robert W Platt, Shi Wu Wen, KS Joseph, Alexander Allen, Michal Abrahamowicz, Béatrice Blondel, Gérard Bréart, et al. A new and improved population-based canadian reference for birth weight for gestational age. *Pediatrics*, 108(2):e35–e35, 2001.

[28] Eric Lavigne, Abdool S Yasseen, David M Stieb, Perry Hystad, Aaron van Donkelaar, Randall V Martin, Jeffrey R Brook, Daniel L Crouse, Richard T Burnett, Hong Chen, et al. Ambient air pollution and adverse birth outcomes: Differences by maternal comorbidities. *Environmental research*, 148:457–466, 2016.

[29] Jundong Li, Aibek Adilmagambetov, Mohomed Shazan Mohomed Jabbar, Osmar R. Zaïane, Alvaro Osornio-Vargas, and Osnat Wine. On discovering co-location patterns in datasets: a case study of pollutants and child cancers. *GeoInformatica*, 20(4):651–692, 2016.

[30] Jundong Li and Osmar Zaiane. Associative classification with statistically significant positive and negative rules. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 633–642. ACM, 2015.

[31] Jundong Li, Osmar R Zaïane, and Alvaro Osornio-Vargas. Discovering statistically significant co-location rules in datasets with extended spatial objects. In *Proceedings of the 16th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 124–135. Springer, 2014.

[32] Bing Liu Wynne Hsu Yiming Ma. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*, 1998.

[33] Mohomed Shazan Mohomed Jabbar and Osmar R. Zaïane. Learning statistically significant contrast sets. In *Proceedings of the 29th Canadian Conference on Artificial Intelligence (Canadian AI 2016)*, pages 237–242. Springer, 2016.

[34] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(Feb):377–403, 2009.

[35] J Ross Quinlan. C 4.5: Programs for machine learning. *The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann,— c1993*, 1, 1993.

[36] Amit Satsangi and Osmar R Zaïane. Contrasting the contrast sets: An alternative approach. In *Proceedings of the 11th International Database Engineering and Applications Symposium (IDEAS)*, pages 114–119. IEEE, 2007.

[37] Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns: A summary of results. In *Proceedings of the 7th International Symposium on Spatial and Temporal Databases (SSTD)*, pages 236–256, 2001.

[38] Shashi Shekhar, Pusheng Zhang, and Yan Huang. Spatial data mining. In *Data mining and knowledge discovery handbook*, pages 837–854. Springer, 2009.

[39] Paul B Tchounwou, Clement G Yedjou, Anita K Patlolla, and Dwayne J Sutton. Heavy metal toxicity and the environment. In *Molecular, clinical and environmental toxicology*, pages 133–164. Springer, 2012.

[40] Geoffrey I Webb. Discovering significant rules. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 434–443, 2006.

[41] Geoffrey I Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

[42] Geoffrey I Webb, Shane Butler, and Douglas Newlands. On detecting differences between groups. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 256–265. ACM, 2003.

[43] Geoffrey I Webb and Songmao Zhang. K-optimal rule discovery. *Data Mining and Knowledge Discovery*, 10(1):39–79, 2005.

[44] Hui Xiong, Shashi Shekhar, Yan Huang, Vipin Kumar, Xiaobin Ma, and Jin Soung Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, pages 78–89, 2004.

[45] Xiaoxin Yin and Jiawei Han. Cpar: Classification based on predictive association rules. pages 331–335, 2003.

[46] Osmar R Zaiane, Jiawei Han, and Hua Zhu. Mining recurrent items in multimedia with progressive resolution refinement. In *Proceedings of the 16th International Conference on Data Engineering (ICDE)*, pages 461–470. IEEE, 2000.

[47] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.