“Planning for the long-term safety of a community is one of the most meaningful things a municipal government can do. I am determined to make road safety a priority, now and for the future.”

– Stephen Mandel, Mayor of Edmonton

University of Alberta

DISCOVERING SPATIAL CO-CLUSTERING PATTERNS IN COLLISION DATA

by

Dapeng Li

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

Master of Science

Department of Computing Science

Fall 2013
Edmonton, Alberta

Dedication

The thesis is dedicated to my dear father Dacheng Li, to my loving mother Xiaoai Niu, and to my beloved brother Min Li. Their support, encouragement, and constant love have sustained me throughout my life.

Abstract

Identifying spatial patterns of collisions is critical for improving the efficiency and effectiveness of traffic enforcement deployment and road safety. Recently, many studies have centred on finding locations with high collision concentration, so-called hotspots. However, most of them only focus on the location information of the collision data, without integrating the non-spatial attributes into analysis. Taking non-spatial attributes into account opens opportunities to reveal attribute-related hotspots that otherwise goes undetected, and can add valuable indicators for explaining those hotspots. In this thesis, we address this problem. We propose a method for identifying the sets of non-spatial attribute-value pairs (AVPs) that together contribute significantly to the spatial clustering of the corresponding collisions. We call such AVP sets Spatial Co-Clustering Patterns (SCCPs). By applying our method on Edmontons collision data, we discovered larger numbers of meaningful hotspot patterns than traditional methods did, and revealed the relevant non-spatial indicators for explaining those hotspots.

Acknowledgements

First of all, I would like to sincerely thank my supervisors Dr. Joerg Sander and Dr. Mario A. Nascimento for their kindness and professionalism. Without their guidance and persistent support, this thesis would not have been possible. My appreciation also goes to Dr. Dae-Won Kwon and Dr. Stevanus Tjandra at the City of Edmonton's Office of Traffic Safety (OTS), who provided me with the collision data, suggests and feedback. I would also like to express my appreciation to my fellow friends who gave me kind help when I was seeking ideas or possible solutions: Charlene Nielsen, Haiming Wang, Feng Jiang, Ailin Zhou, Dan Han and Yang Zhao. Finally, the most special thanks go to my family who have loved and supported me for all my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

A traffic collision, also known as a road collision, motor vehicle collision, car accident or automobile accident, is an unintentional collision between a moving vehicle and any other vehicle, pedestrian, animal, or stationary obstruction, such as a hedge or building. Traffic collisions have been a great threat to human life and impose huge economic burden on the society. Although governments have been improving traffic infrastructure and taking preventative measures, traffic collisions remain a big problem worldwide. A report by the World Health Organization in 2004 [2] stated that, globally, the number of people killed in road traffic crashes each year was estimated at almost 1.2 million, while the number injured could be up to 50 million. Without increased efforts and new initiatives, these figures would increase by some 65% between 2000 and 2020 [3, 4]. In 2009, the City of Edmonton's Office of Traffic Safety reported 490 serious and 4,713 minor injuries [5]. Using the cost components described in a collision cost study by Paul De Leur, Laura Thue and Brian Ladd [6], collisions in the city of Edmonton cost over 0.5 billion each year.

Identifying the spatial patterns of collisions is critical for improving road safety, as well as improving the efficiency and effectiveness of traffic enforcement deployment. In traffic safety applications, geographic information system (GIS) technology has been widely used to geo-code collision locations, leading to large amounts of spatial data about collisions. Apart from location information, GIS software also

enables us to store non-spatial attributes related to each collision occurrence, such as the cause, severity, road surface condition, etc. And each attribute has a set of possible values. Taking road surface condition as an example, the possible values can be 'Dry', 'Wet', 'Snowy/Icy', etc. All this plethora of information in the collision data provides great potential for identifying patterns.

In recent years, many studies in traffic safety area have focused on finding collision "hotspots". Although there is an abundance of literature addressing hotspot analysis, no universal definition of hotspots exists [7]. The concrete definition of hotspots usually depends on the goal of our analysis and the specific method we use. In traffic area, hotspots basically refer to locations with unusually high concentration of collision occurrences. The general purposes of collision hotspot analysis may include: (1) finding clusters of collisions, and (2) identifying problematic locations for safety improvement. Up to now, a number of methods have been put forward for identifying collision hotspots by advanced statistical analytics, for instance, using density estimation [8] or spatial autocorrelation measures [9]. However, most of them only focus on the location information of the collision data, leaving the abundant non-spatial attributes unutilized. Taking non-spatial attributes into account, on the other hand, opens opportunities to reveal a larger number of hotspot patterns, and can add valuable indicators for explaining certain collision hotspots.

In this thesis, we address this problem. We propose a method which integrates the non-spatial attributes of the collision data into hotspot analysis. In particular, we investigate which non-spatial attribute-value pairs (AVPs) contribute together to the spatial clustering of corresponding collisions. We call such an AVP set a Spatial Co-Clustering Pattern (SCCP). The proposed method can be applied to historical collision data. The detected SCCPs can: (1) lead to the discovery of a large number of attribute-related hotspots, which allow the government to deploy its traffic enforcement resources more efficiently and effectively; (2) reveal the relevant AVPs that contribute to those hotspots, which can help explain the frequent collision occurrence at those locations. All these results found have the potential to allow more effecive and efficient strategies for deploying resources for traffic enforcement and

road safety.

1.2 Thesis Contributions

In this thesis, we propose a method to identify the sets of non-spatial attribute-value pairs (AVPs) that together contribute significantly to the spatial clustering of the corresponding collisions. Here we list four main contributions of this thesis.

Our first contribution is that we find a way to integrate non-spatial attributes of the collision data into hotspot analysis, which leads to the full utilization of the collision data and the discovery of larger numbers of hotspot patterns. Compared to traditional hotspot analysis methods, which regard collisions as varying only on one dimension, location, our method treats collisions as varying on both location and non-spatial attributes, such as the cause, severity, road surface condition, etc. Therefore, our method can discover larger numbers of hotspot patterns in the collision data than the traditional methods.

Another contribution of this thesis is that we put forward two objective criteria for the interestingness of AVP sets for hotspot analysis. The first one is that the AVP set leads to a spatial clustering of the corresponding collisions. And the second one is that all subsets of the AVP set make a significant contribution to that spatial clustering. Compared to the subjective criteria used in traditional hotspot analysis methods which are based on domain knowledge, these criteria guarantee that (1) the corresponding collisions of the AVP set are inherently clustered in space, and (2) there is no subset of the AVP set that doesn't make a significant contribution. Therefore, the interestingness of the AVP set selected with these criteria is objectively defined.

Following those criteria, we then introduce our concept of Spatial Co-Clustering Pattern (SCCP), which represents an interesting AVP set. Each SCCP will lead to the discovery of corresponding interesting "attribute-related" hotspots. We then propose a method for efficiently discovering all SCCPs among all possible AVP sets, with multiple techniques for computation reduction applied. This is our third main contribution.

Moreover, we also applied our method on Edmonton's historical collision data to discover the interesting hotspot patterns inside. The SCCPs discovered by our method: (1) lead to the discovery of a large number of attribute-related hotspots, which allow the government to deploy its traffic enforcement resources more efficiently and effectively; (2) reveal the relevant AVPs that contribute to those hotspots, which can help explain the frequent collision occurrence at those locations. All the results found have the potential to allow more effective strategies for deploying resources for traffic enforcement and road safety.

1.3 Thesis Outline

The rest of the thesis is organized as follows. We discuss traditional hotspot analysis methods and their limitation in Section 2. Our concept of Spatial Co-Clustering Pattern (SCCP) and our SCCP discovery method are presented in Section 3. In section 4, we present our experimental setup and the results on Edmonton's historical collision data. Section 5 concludes with a summary and some directions for future research.

Chapter 2

Related Work

2.1 Traditional hotspot analysis methods

In the earliest studies, traditional collision hotspot analysis centred on road segments or specific junctions [10, 11], and considered collisions as individual points on a map. Based on that, some researchers ranked locations according to accident rate (accidents per driven vehicle kilometre), while others used accident frequencies (accidents per road kilometre) [12]. Since the late 1970s, more sophisticated statistical models have been applied to road collision analysis, such as Poisson regression models [13, 14] and and negative binomial regression models [15, 16]. However, all of these methods above assume the occurrences of collisions on different road sections are independent [17]. The spatial dependence of collisions is therefore ignored.

In recent years, GIS has been widely used to geo-code collision locations and integrate disparate databases. Meanwhile, there are also increasing interest in the spatial dependence of collisions in a similar area, and the association of spatial factors to road collisions. According to Anderson [12], this dependence of collisions is argued to be the result of a shared common cause(s) among the collisions, albeit of varying intensity [18, 19]. Under the circumstances, new methods that assume the spatial dependence, such as kernel density estimation [8] and spatial autocorrelation [9], have been put forward and used.

More detailed introductions of the traditional hotspot analysis methods are given in the following three subsections.

2.1.1 Hotspot analysis with statistical models

Count data consist of non-negative integer values and are encountered frequently in modeling traffic-related phenomena. Examples of count data variables in traffic include the number of accidents observed on road sections during a period of one year, the amount of vehicle exposure on road sections per years, etc. A common mistake in traffic-related analysis is to model count data as continuous data by applying standard least squares regression(*e.g.*,multiple linear regression), which yield predicted values that are non-integers or negative [20]. These results are inconsistent with the real data and thus make standard regression analysis inappropriate for modeling count data.

The unsatisfactory property of standard regression models has led to the investigation of a number of more sophisticated statistical models. The most popular ones are Poisson and negative binomial regression models. Poisson and negative binomial regression models are usually used to establish the relationship between road collisions and geometric conditions of road sections, signalization, pavement types, and so on. Maximum likelihood estimation is used to estimate the unknown parameters of these models. When the relationship is well established, we obtain a corresponding collision prediction model. To target collision hotspots, road sections are ranked according to a measure computed with the collision prediction model, for instance, the expected number of collisions at each road section. Then the top-ranked road sections are considered as hotspots.

About Poisson regression model and negative binomial model, more detailed introductions are given in the following subsections.

Poisson regression model

The Poisson regression model is used to approximate the count data of events whose occurrences are rare, such as road accident occurrences, failures in manufacturing or processing, etc. To help illustrate the principal elements of a Poisson regression model, consider a set of n intersections in a city. Let Y_i be a random variable representing the number of collisions on intersection i per year, where $i = 1, 2, ..., n$. In a Poisson regression model, the probability of intersection i having y_i accidents

per year (where y_i is a non-negative integer) is given by

$$P(Y_i = y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \tag{2.1}$$

This model assumes that $Y_i, i = 1, 2, ..., n$, are independently and Poisson distributed with Poisson parameter λ_i, which is equal to the expected number of accidents per year at intersection i, $E(Y_i)$. We estimate Poisson regression models by specifying the Poisson parameter λ_i as a function of explanatory variables. In our intersection accident example, the explanatory variables may include the geometric conditions of the intersections, signalization, pavement types, *etc*. The relationship between explanatory variables and the Poisson parameter is often approximated with the log-linear model, given by

$$\lambda_i = \exp(\boldsymbol{\beta}\boldsymbol{\chi_i}) \tag{2.2}$$

where $\boldsymbol{\chi_i}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of parameters. Then the expected number of accidents per year at intersection i is given by $E(Y_i) = \lambda_i = \exp(\boldsymbol{\beta}\boldsymbol{\chi_i})$. The regression parameters $\boldsymbol{\beta}$ of this model can be estimated using the maximum likelihood estimation [21]. The likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp[-\exp(\boldsymbol{\beta}\boldsymbol{\chi_i})][\exp(\boldsymbol{\beta}\boldsymbol{\chi_i})]^{y_i}}{y_i!} \tag{2.3}$$

As with most statistical models, the estimated parameters can be used to make inferences about the unknown population characteristics thought to influence the count process [20].

Negative binomial regression model

One property of the Poisson distribution is that the mean of the count data equals its variance. However, in many studies, the variance of the count data is significantly larger than its mean, which is known as the overdispersion problem. One major reason of the overdispersion problem in many studies is that certain variables impacting the Poisson parameters across observations have been omitted from the regression model [20]. For overdispersed count data, using Poisson regression models will lead to a biased parameter vector, unless corrective measures are taken. In such cases, a negative binomial model is often used instead.

The negative binomial model is derived by rewriting Equation 2.2 such that, for each observation i,

$$\lambda_i = \exp(\boldsymbol{\beta}\boldsymbol{\chi_i} + \epsilon_i) \tag{2.4}$$

where $\exp(\epsilon_i)$ is a gamma-distributed disturbance term with mean 1 and variance α [20]. The addition of this term allows the variance to differ from the mean as below [20]:

$$Var(Y_i) = E(Y_i)[1 + \alpha E(Y_i)] = E(Y_i) + \alpha E(Y_i)^2 \tag{2.5}$$

where $\alpha \geq 0$ and is usually referred to as dispersion parameter. From Equation 2.5 we can see that this model allows the variance to exceed the mean. Also, the Poisson regression model can be regarded as a limiting model of the negative binomial regression model as α approaches 0. The negative binomial distribution has the form:

$$P(Y_i = y_i) = \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha)y_i!}(\frac{1/\alpha}{(1/\alpha) + \lambda_i})^{1/\alpha}(\frac{\lambda_i}{(1/\alpha) + \lambda_i})^{y_i} \tag{2.6}$$

where $\Gamma(.)$ is a gamma function. This results in the likelihood function:

$$L(\lambda_i) = \prod_{i=1}^{n} \frac{\Gamma((1/\alpha) + y_i)}{\Gamma(1/\alpha)y_i!}(\frac{1/\alpha}{(1/\alpha) + \lambda_i})^{1/\alpha}(\frac{\lambda_i}{(1/\alpha) + \lambda_i})^{y_i} \tag{2.7}$$

Compared to the Possion regress model, the negative binomial regression model is more general. But it requires more extensive computation to estimate model parameters and to generate inferential statistics than the Poisson regression model [17].

2.1.2 Hotspot analysis with kernel density estimation

The kernel density estimation (KDE) method is based on the assumption that road collisions occurring in a similar area are spatially dependent. According to Tessa [12], this dependence is argued to be the result of a shared common cause(s) among the collisions, albeit of varying intensity [18, 19].

Compared to the above methods with statistical models, the main advantage of KDE method lies in determining the spread of risk of an accident. The spread of risk can be defined as the area around a defined cluster in which there is an increased likelihood for an accident to occur based on spatial dependency.

KDE is an interpolation technique, which generalizes collision locations to the entire study region. To do the interpolation, a symmetrical surface, which is determined by the kernel, is placed over each collision location point. For each location point in the study region, we sum the values of all individual surfaces over it, as its density estimate for the distribution of accident points. This procedure is repeated for all the successive points. A more detailed introduction of KDE can be found in Silverman's book [22].

The density estimate at the location (x, y) is given as:

$$f(x,y) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{d_i}{h}) \tag{2.8}$$

where n is the number of collisions, h is the bandwidth or kernel size, K is the kernel function, and d_i is the distance between the location (x, y) and the location of the ith collision ($i = 1, 2, ..., n$).

The effect of placing these kernels over the points is to create a smooth and continuous surface of density estimates. Around each collision location point, a circular area (the kernel) of pre-specified bandwidth is created, and the risk of the collision occurrence at that point is spread into it according to the kernel function (See Figure 2.1). Summing all of these values at all places, including those at which no collisions were observed, gives a smooth and continuous surface of density estimates. Then the locations with high density estimates will be the detected hotspots.

In the KDE method, a range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. Although there are various kernel functions to choose from, most agree that the kernel function will not significantly impact results [22, 1]. However, the bandwidth h will impact the resulting density map greatly. A very small bandwidth, for example, will lead to inadequate smoothing and simply highlight individual points[23]. If the bandwidth is increased, there is a possibility that the circular neighbourhood would include more collision location points, which finally results in a smoother density surface [22]. It often takes trial and error on the bandwidth in order to produce an appropriate density surface.

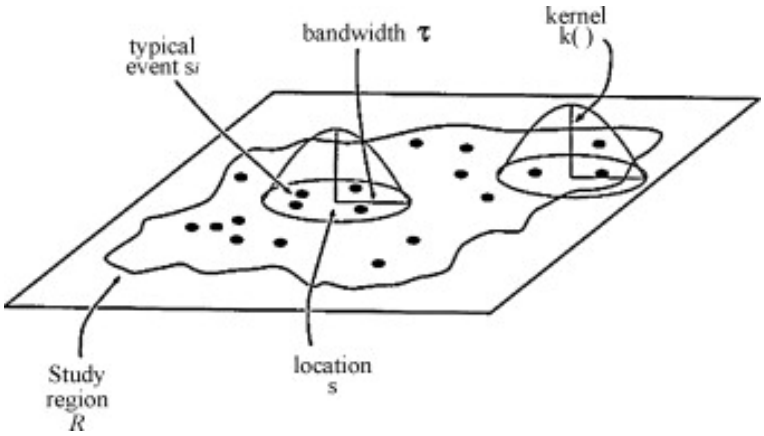

Figure 2.1: Diagram of how KDE method works, reproduced from Bailey and Gatrell's work [1]

2.1.3 Hotspot analysis with spatial autocorrelation

In its most general sense, spatial autocorrelation is concerned with the degree to which objects or activities at some place on the earth's surface are similar to other objects or activities located nearby [24]. Its existence is reflected by Waldo Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things."[25]

The assessment of spatial autocorrelation involves measuring the degree to which the value of a variable for each location co-varies with values of that variable at nearby locations [18]. When the level of co-variation is higher than expected, contiguous locations have similar values and the spatial autocorrelation is positive. When the level of co-variation observed are negative, high values of the variable are surrounded by low values and the spatial autocorrelation is negative. The lack of significant positive or negative co-variation suggests the absence of spatial autocorrelation [26].

To quantify the spatial autocorrelation, Geary's Ratio and Moran's I are two popular indices that are generally used. Geary's Ratio and Moran's I combine the measures of both attribute similarity and location proximity into a single index. They measure and test if patterns of point distributions are clustered or dispersed in space with respect to their attribute values. In traffic accident application, the point can refer to the collision location point, and the attribute value is the number

of collisions at each location point.

Geary's Ratio and Moran's I are used to assess the global spatial autocorrelation over the entire study region. So they are also called global statistics. For instance, Moran's I index is often called Global Moran's I index. Most analysts favor Global Moran's I as its distributional characteristics are more desirable and this index has greater general stability and flexibility [26]. Global Moran's I index is defined as:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \bar{X})(x_j - \bar{X})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \tag{2.9}$$

where x_i is the number of collisions at location point i, $\bar{X}$ is the mean of all x_i, $w_{i,j}$ is the spatial weight between point i and j (e.g., inverse of the distance), n is the total number of x_i, and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \tag{2.10}$$

Given the number of collisions at each location, the Global Moran's I index evaluates whether the distribution of collisions at these locations is clustered, dispersed or random in space. The distribution of collisions presents "a clustered pattern" when high numbers and low numbers in the space are more spatially clustered than would be expected if underlying spatial processes were random. In contrast, it presents "a dispersed pattern" when high numbers and low numbers in the space are more spatially dispersed than would be expected if underlying spatial processes were random.

Possible values of Global Moran's I range from -1 to 1. A positive value indicates a clustered pattern and a negative value indicates a dispersed pattern. An associated z-score [24] is calculated to evaluate the statistical significance of the Global Moran's I index. For statistically significant positive z-scores, the higher the z-score is, the more pronounced the clustered pattern is, *i.e.*, the less likely the clustered pattern is generated by some random spatial process. For statistically significant negative z-scores, the lower the z-score is, the more pronounced the dispersed pattern is, *i.e.*, the less likely the dispersed pattern is generated by some random spatial process.

As stated above, Geary's C Ratio and Moran's I Index are global statistics since they are measures of the entire study area. To investigate the accurate collision hotspot locations, it is necessary to use local measures such as the local Moran's I index [27] and Getis-Ord Gi* statistic [28, 29]. These statistics are used to quantitatively measure the level of spatial autocorrelation at the local scale. Particularly, the Getis-Ord Gi* statistic is useful to identify cold/hot spots where low/high values cluster spatially. The Getis-Ord Gi* statistic is defined as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}} \tag{2.11}$$

where x_i is the number of collisions at location point i, $\bar{X}$ is the mean of all x_i, $w_{i,j}$ is the spatial weight between point i and j (e.g., inverse of the distance), n is the total number of x_i, and S_0 is the variance of all x_i:

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \tag{2.12}$$

Similar to the Global Moran's I index, the Getis-Ord Gi* statistic also has an associated z-score to indicate its statistical significance. Actually, since the Getis-Ord Gi* statistic itself is a z-score, its associated z-score is itself. For statistically significant positive z-scores, the higher the z-score is, the more pronounced the clustering of high numbers is, *i.e.*, the less likely the clustering of high numbers is generated by some random spatial process. For statistically significant negative z-scores, the lower the z-score is, the more pronounced the clustering of low numbers is, *i.e.*, the less likely the clustering of low numbers is generated by some random spatial process. Then the locations with z-scores higher than a pre-specified threshold (*e.g.*, 2.0) are targeted as the hotspots.

2.2 Limitation of traditional hotspot analysis methods

While the above traditional methods are commonly used in collision hotspot analysis, none of them utilize the abundant non-spatial attributes of collision data, such as the cause and severity. As a result, they don't differentiate collisions on their

various attribute values, but treat all collisions as if they are the same, except for location. In other words, they regard collisions as varying only on one dimension, location (or two dimensions x, y coordinates).

Following is an example of using the traditional spatial autocorrelation method to identify hotspots of Edmonton's collisions in 2011:

Figure 2.2 displays the raw collision data of Edmonton in 2011. On the map, each point represents a location where at least one collision happened in 2011. Multiple collisions can be geo-coded at the same location.

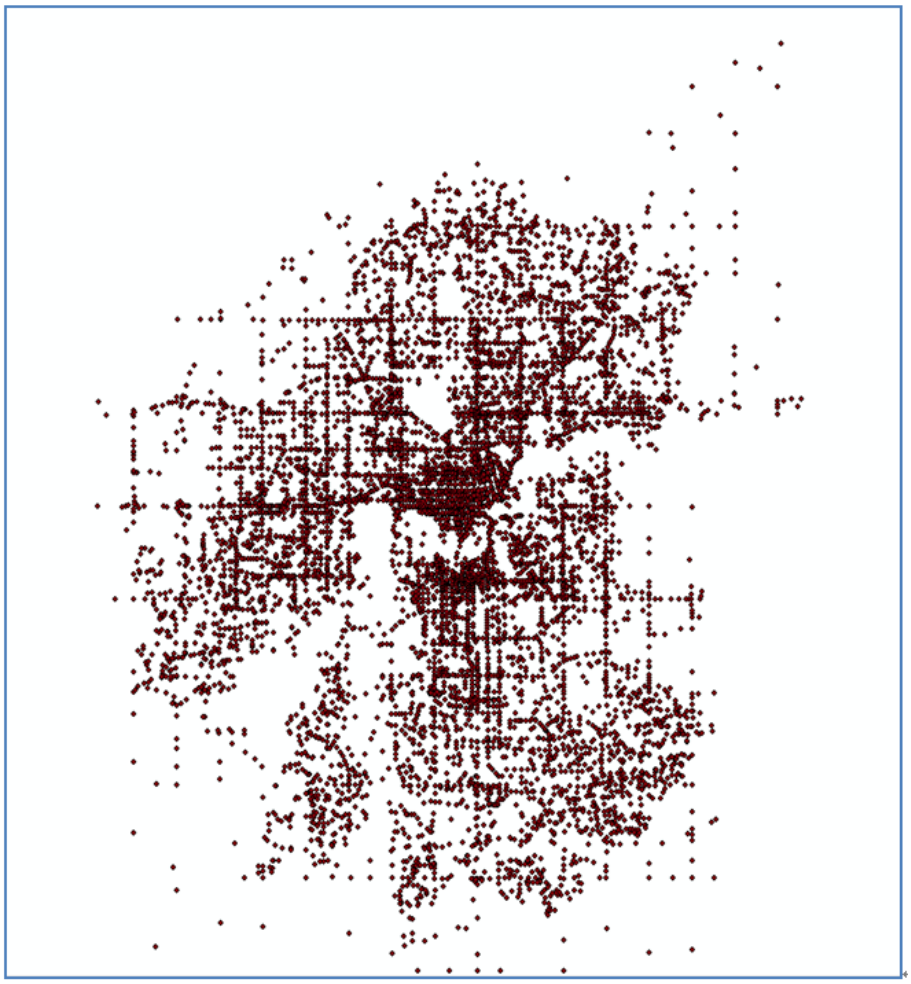

Figure 2.2: Display of the raw collision data of Edmonton in 2011

In a traditional spatial autocorrelation method, we first get the count of collisions at each point, disregarding their various non-spatial attributes. With these counts, we further calculate for each point a spatial autocorrelation statistic, e.g., Getis-Ord

Gi* statistic [28, 29]. After that, we label each point in different colours according to its calculated statistic. If the statistic is higher than a pre-specified threshold (e.g., 2.0), the point will be labeled in red and be considered as a hotspot. We can further overlay the resultant hotspot map with other background layers to help identify the concrete locations of the detected hotspots, as illustrated in Figure 2.3.

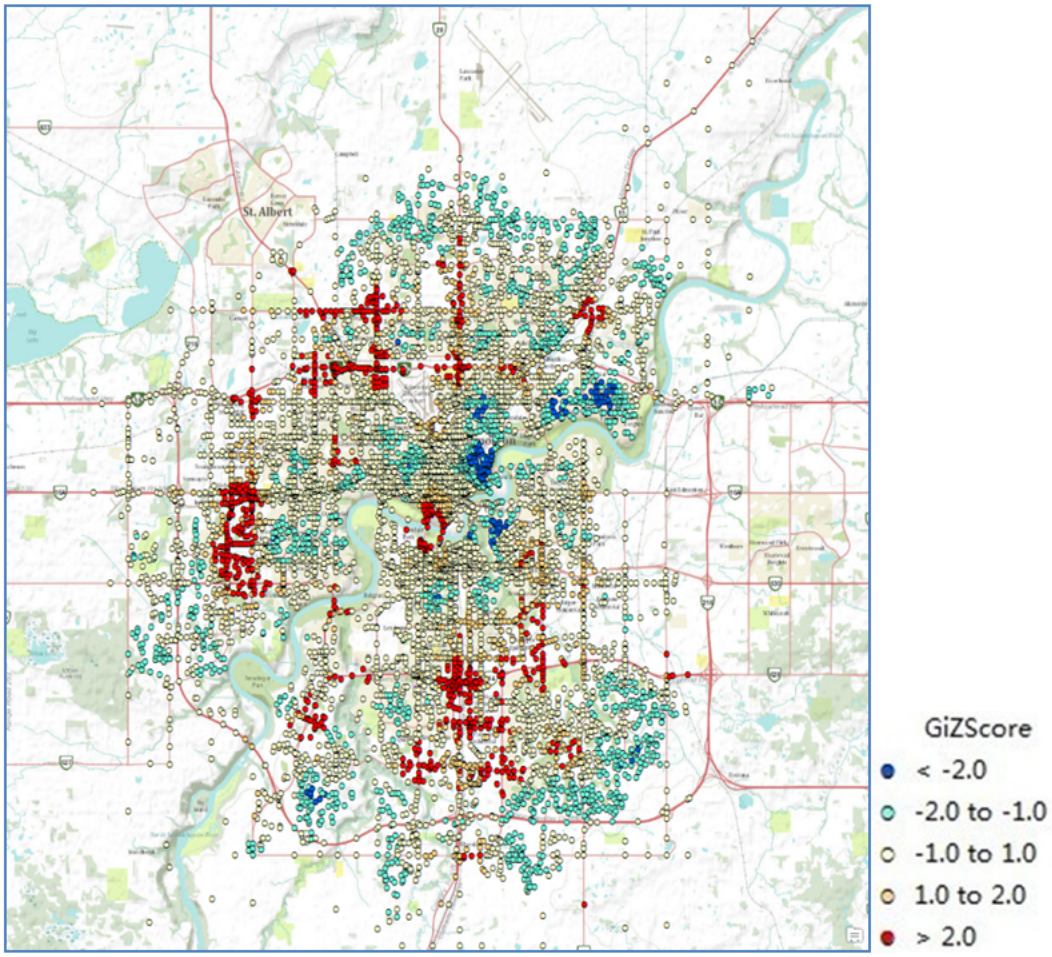

Figure 2.3: Hotspots of the overall collisions

From the above example, we see that traditional hotspot methods only focus on the location information of collision data, and treat collisions with various non-spatial information as the same. Therefore, they can only detect the "overall" hotspots of all the collisions, regardless of their various attribute values. On the other hand, if we investigate on certain specific types of collisions in terms of AVPs (e.g., collisions with *Day of the week* = '*Saturday*' *and Road surface condition* =

'Dry') and extract the corresponding collisions from the overall data, the hotspots detected with the same traditional method as above (shown in Figure 2.4) would vary greatly from the results in Figure 2.3. Such "attribute-related" hotspots are also potentially interesting, because they reveal the concentration locations of some specific types of collisions. In addition, the relevant AVPs of these hotspots can also help explain the frequent collision occurrence at these locations.

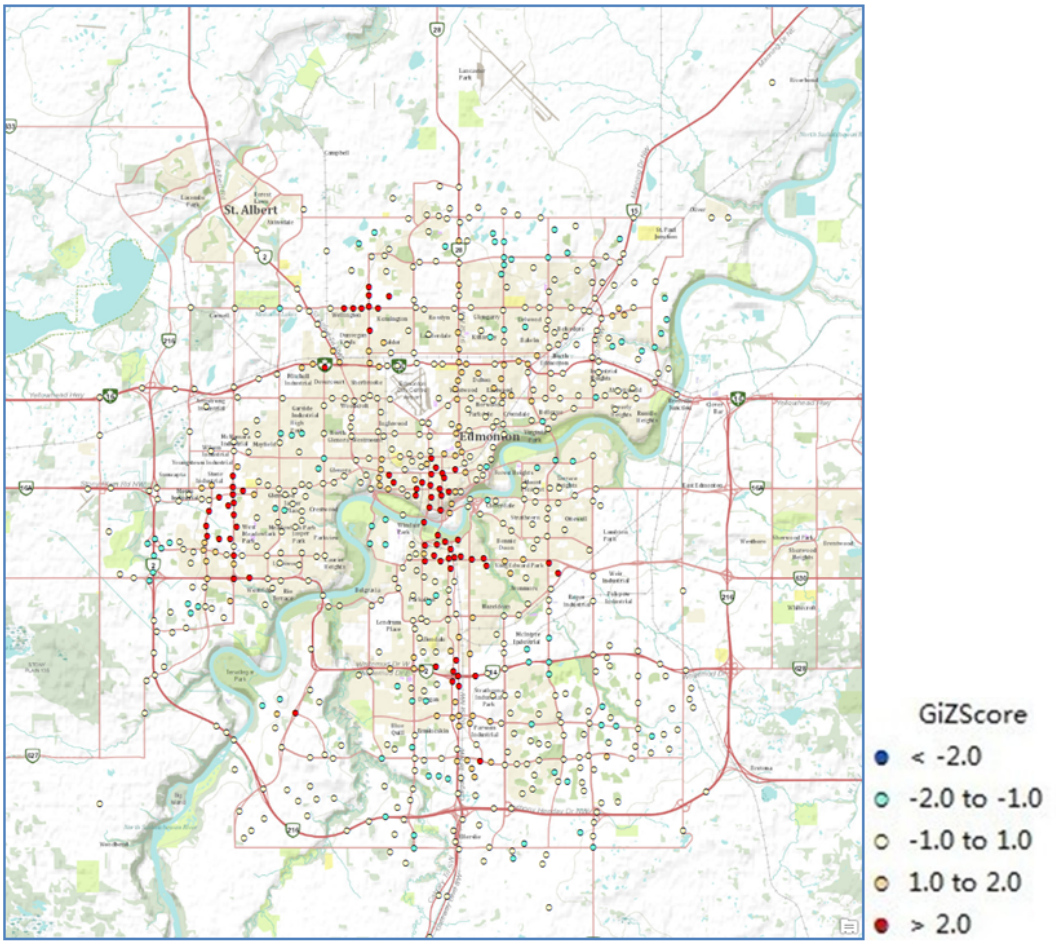

Figure 2.4: Hotspots of the collisions with *Day of the week = 'Saturday' and Road surface condition = 'Dry'*

Therefore, integrating non-spatial attributes into hotspot analysis is a promising and rewarding avenue to improve traditional methods.

Chapter 3

Our Methodology

3.1 Spatial Co-Clustering Pattern

In Section 2, we argued that some "attribute-related" hotspots as displayed in Figure 2.4 are also potentially interesting. To discover such "attribute-related" hotspots, a simple approach is:

1. Select a combination of attribute-value pairs (AVPs) that is "interesting" to us (probably based on domain knowledge), *e.g.*, *Day of the week = 'Saturday' and Road surface condition = 'Dry'*. We call such a combination of AVPs an "interesting" AVP set;

2. Obtain the corresponding collisions that satisfy the AVP set;

3. Detect the hotspots of those collisions with the traditional methods.

However, this approach has two obvious problems: (1) the interestingness of the selected AVP set is subjective; (2) by only considering the selected AVP set as interesting, we may miss many other interesting AVP sets.

To address these problems, in this section, we first set objective criteria for the interestingness of AVP sets for hotspot analysis. Based on those criteria, we then define a special pattern which represents an interesting AVP set, called Spatial Co-Clustering Pattern (SCCP). Each SCCP will lead to the discovery of corresponding interesting "attribute-related" hotspots. Then later, in the following section, we will propose a method to discover all SCCPs among all possible AVP sets.

3.1.1 The criteria for the interestingness of AVP sets

In the traffic safety area, hotspots refer to locations with high collision concentration. If the collisions are inherently not clustered in space, it is unlikely that we can find some interesting hotspots inside.

For example, assume the distributions of two different types of collisions are as shown in Figure 3.1. The two types of collisions correspond to two AVP sets, denoted as AVP Set 1 and AVP Set 2. In Figure 3.1, each point represents a location where one collision happened. We can see that collisions of AVP Set 1 are dispersed in space, while collisions of AVP Set 2 are clustered spatially. In this case, it is obvious that AVP Set 2 is more interesting for hotspot analysis.

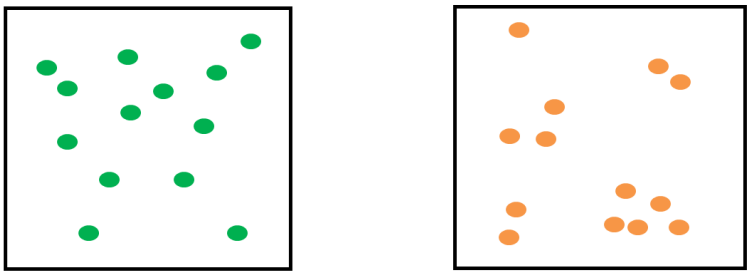

Figure 3.1: distributions of two different types of collisions

Thus we set the following criteria for the interestingness of AVP sets:

Criterion 1: the AVP set must lead to a spatial clustering of the corresponding collisions.

Even if an AVP set leads to a spatial clustering of the corresponding collisions, we cannot guarantee that each of its subset makes a significant contribution to that spatial clustering. If any of its subsets does not make a significant contribution, we should not include this subset to this AVP set and report the whole AVP set as interesting. For example, we can argue that the spatial clustering of collisions with *Day of the week = 'Saturday' and Road surface condition = 'Dry'* may be only attributed to the first constituent AVP: *Day of the week = 'Saturday'*, and the second constituent AVP does not make a significant contribution. If this is true, we should only report {*Day of the week = 'Saturday'*} as an interesting pattern, and not include the AVP set {*Day of the week = 'Saturday' and Road surface condition = 'Dry'*}.

Therefore the second criterion for the interestingness of AVP sets is the following:

Criterion 2: all subsets of the AVP set must make a significant contribution to the spatial clustering of the corresponding collisions.

3.1.2 The definition of Spatial Co-Clustering Pattern

Let $F : \{f_1, f_2, ..., f_t\}$ be a set of instances of spatial feature f. Each spatial feature f has a set of attributes $A : \{A_1, A_2, ..., A_n\}$. Each attribute A_i has a set of possible values $v_i : \{v_{i1}, v_{i2}, ..., v_{im_i}\}$.

For an AVP set $s : \{A_i = v_{ip}, A_j = v_{jq}, ..., A_k = v_{kr}\}$, its corresponding instance set, denoted by $F(s)$, is defined by $F(s) = \{f | f.A_i = v_{ip}, f.A_j = v_{jq}, ..., f.A_k = v_{kr}\}$. $F(s)$ is a subset of F. We call the number of instances in $F(s)$ its *size*.

Let s' be a subset of s. Its corresponding instance set is denoted as $F(s')$. Because s' is a subset of s, we can derive that $F(s')$ is a superset of $F(s)$.

For a certain s, if the instances of $F(s)$ are spatially clustered [1], we say "$F(s)$ presents a clustered pattern" for short.

We call an AVP set s a Spatial Co-Clustering Pattern (SCCP) if and only if it satisfies:

1. $F(s)$ presents a clustered pattern;

2. For any subset s' of s, the probability of obtaining a random subset from $F(s')$, with the same size as $F(s)$ and a clustered pattern more pronounced than $F(s)$, is low.

Above is the general definition of a SCCP. If an AVP set s satisfies Property (2), we say the $F(s)$ clusters "significantly better" than a random selection from $F(s')$.

In our application, the spatial feature f refers to "collision" specifically. Correspondingly, s and s' are two AVP sets of the collision data, and $F(s)$ and $F(s')$ are

[1] In our application, this means: after aggregating instances geo-coded on the same location and obtaining the number of instances at each aggregated location point, high numbers and low numbers are respectively more spatially clustered than would be expected if underlying spatial processes were random.

their corresponding collision sets.

The two properties of SCCP are consistent with the two criteria given in the previous subsection. Property (1) ensures Criterion 1 is satisfied. Property (2) ensures that for any subset of s' of s, $F(s)$ is unlikely a random selection from $F(s')$. This ensures that all subsets of the AVP set make a significant contribution to the spatial clustering of the corresponding collisions, thus Criterion 2 is ensured.

In the next section, we will provide statistical methods to quantitatively test whether an AVP set satisfies these two properties.

The following is an example of SCCP:

For the AVP set s: {*Day of the week ='Saturday', Road surface condition ='Dry'*}, the corresponding collisions of $F(s)$ are spatially clustered. Moreover, for its subsets s_1: {*Day of the week ='Saturday'*}, s_2: {*Road surface condition ='Dry'*}, and s_3: $\emptyset$ *(empty set)*, $F(s)$ clusters significantly better than a random selection from any of $F(s_1)$, $F(s_2)$, $F(s_3)$, as shown in Figure 3.2. In this case, s is a SCCP.

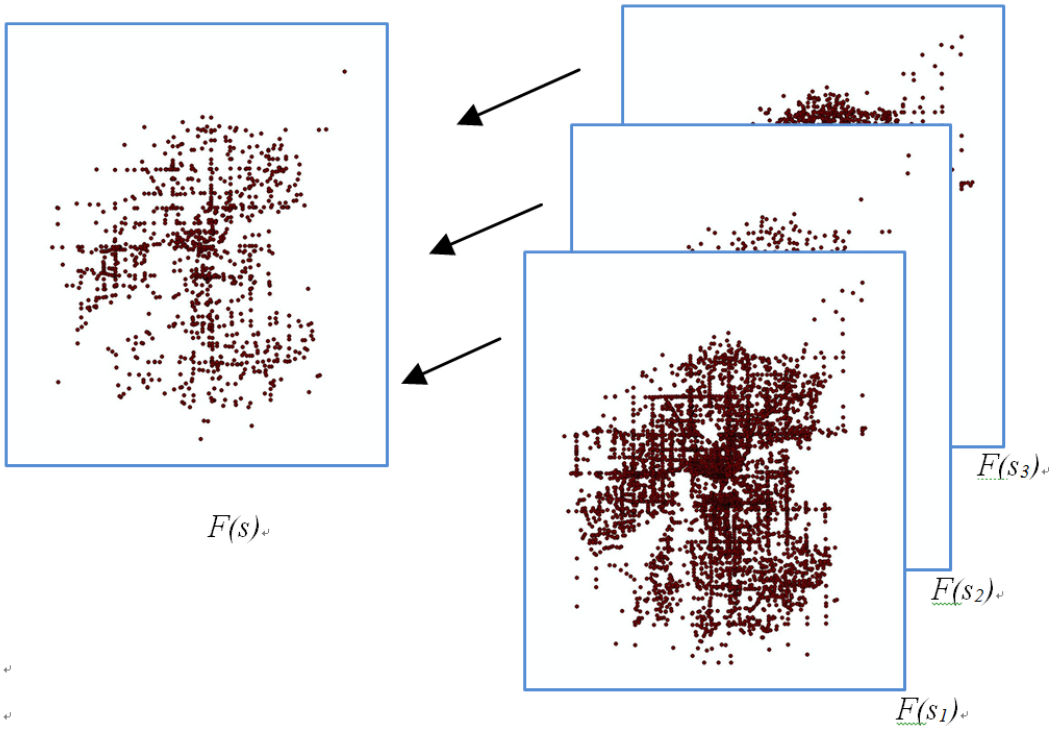

Figure 3.2: $F(s)$ clusters significantly better than a random selection from any of $F(s_1), F(s_2), F(s_3)$.

3.2 Spatial Co-Clustering Pattern Discovery Method

In this section, we propose a method to discover SCCPs among all possible AVP sets. Consistent with the two properties of SCCPs (see Section 3.1.2), our SCCP discovery method consists of two steps:

- *Step 1*: find all AVP sets s whose corresponding collisions of $F(s)$ are spatially clustered;
- *Step 2*: for each s, test whether there exists a subset of s, s', that satisfies: $F(s)$ does not cluster significantly better than a random selection from $F(s')$;

For Step 1, we use Global Moran's I index and the associated z-score [24] (see Section 2.1.3) to evaluate whether the overall distribution of collisions in $F(s)$ is clustered. Only when it is clustered, the corresponding s will be recognized as a candidate for SCCP, and continues to undergo Step 2.

For Step 2, we use randomization tests [30] to check s against each of its subsets s', to determine whether $F(s)$ clusters significantly better than a random selection from $F(s')$.

More details are given in the following two subsections.

3.2.1 Step 1: find all AVP sets that lead to spatial clustering of collisions

In Step 1, we use the Global Moran's I index stated in Section 2.1.3 to evaluate the overall distribution of collisions. Given the collision counts at each location, the Global Moran's I index evaluates whether the distribution of collisions at these locations is clustered, dispersed or random in space. An associated z-score [24] is calculated to evaluate the statistical significance of the Global Moran's I index: a positive z-score indicates a clustered pattern; and the higher the z-score is, the less likely the expressed clustered pattern is generated by some random process, *i.e.*, the more pronounced the clustered pattern is.

In Step 1, we enumerate all possible AVP sets s, and only retain those whose corresponding z-score of $F(s)$ is higher than a pre-specified threshold γ (*e.g.*, 5.0 in our experiment) as candidates for SCCP.

Due to the combinatorial explosion, the number of possible AVP sets is very large. So computing the Global Moran's I index and the associated z-score for each one is computationally expensive. To reduce the amount of computation, we ignore those AVP sets that have only a small number of (*e.g.*, less than 50) corresponding collisions, and don't calculate their Global Moran's I indices and the associated z-scores. We call such AVP sets "infrequent". Since these AVP sets correspond to collisions whose occurrences are so rare, they deserve less attention. To determine whether an AVP set is infrequent or not, we pre-specify a threshold, called frequency threshold. Only for a "frequent" AVP set whose corresponding collisions are more than the frequency threshold, we need to obtain its $F(s)$, and calculate its Global Moran's I index and associated z-score.

To judge whether an AVP set is frequent or not, we need to query its corresponding collisions over the whole collision data, which is also computationally expensive. When searching for frequent AVP sets, in order to reduce the queries over the collision data, we adopt a technique which is similar to that in the Apriori algorithm [31] for finding frequent itemsets. The basic intuition is that any subset of a frequent AVP set must be frequent. Therefore, frequent AVP sets having k AVPs can be generated by joining frequent AVP sets having $k-1$ AVPs, and deleting those that contain any subset that is not frequent. In this way, we can discard a large number of infrequent AVP sets without querying their corresponding collisions over the data. Therefore, the computation for querying is largely reduced. More details about this technique will be given in the pseudocode of our algorithm at the end of this chapter.

In summary, the Step 1 can be decomposed into the following two tasks:

(1) Find all frequent AVP sets with the above technique;

(2) For each frequent AVP set s, calculate the Global Moran's I index and associated z-score for its $F(s)$. If the z-score is higher than the pre-specified threshold, s continues to undergo Step 2; else, it is discarded.

3.2.2 Step 2: test the contribution of each subset to the clustering of collisions

In Step 2, the randomization test we use is based on the probability model that the collision counts (x_i) are randomly assigned to the location points (i), which is called a randomization model according to Lehmann [32]. When using randomization to check s against one of its subset s', the null hypothesis of interest is

H_0: the z-score of $F(s)$ is as the same as that of an equally sized random subset from $F(s')$;

Now we test whether the above null hypothesis H_0 should be accepted or rejected, given the pre-specified significance level α. Let the size of $F(s)$ be m and the size of $F(s')$ be n, then based on the randomization model stated above, there are $\binom{n}{m}$ ("n choose m") ways to randomly get a subset of size m from $F(s')$. In other words, there are $\binom{n}{m}$ possible random subsets of size m generated from $F(s')$.

Assume we can calculate the z-scores of all these random subsets. To test whether $F(s)$ is a random selection from $F(s')$, we compare the z-score of $F(s)$ (denoted as $z_{F(s)}$) with those of all these random sets to see whether it is *unusually* large. If the probability of the z-scores of random subsets being as large as or larger than $z_{F(s)}$ (which is called the p-value), is smaller than the pre-specified significance level α, then the null hypothesis H_0 is rejected; else, it is accepted.

For example, assume the histogram in Figure 3.3 shows the distribution of the z-scores of all random subsets. The value of $z_{F(s)}$ is close to the right end of the histogram, and only 1% of all z-scores are as large as or larger than $z_{F(s)}$, *i.e.*, the p-value is equal to 0.01. If we set the significance level α as 0.05, then the p-value is smaller than α, and the null hypothesis H_0 is rejected. In this case, $F(s)$ is considered to cluster significantly better than a random selection from $F(s')$;

In practice, calculating the z-scores of all these random subsets is computationally not feasible: we need to first generate all the $\binom{n}{m}$ random subsets from $F(s')$, which is quite computational expensive; and then calculate the global Morans I index and associated z-score for each of them. One easy and very practical solution to this problem is to use Monte Carlo sampling [30] to estimate the p-value. To do so, we repeatedly and randomly select m collisions from the n collisions in $F(s')$

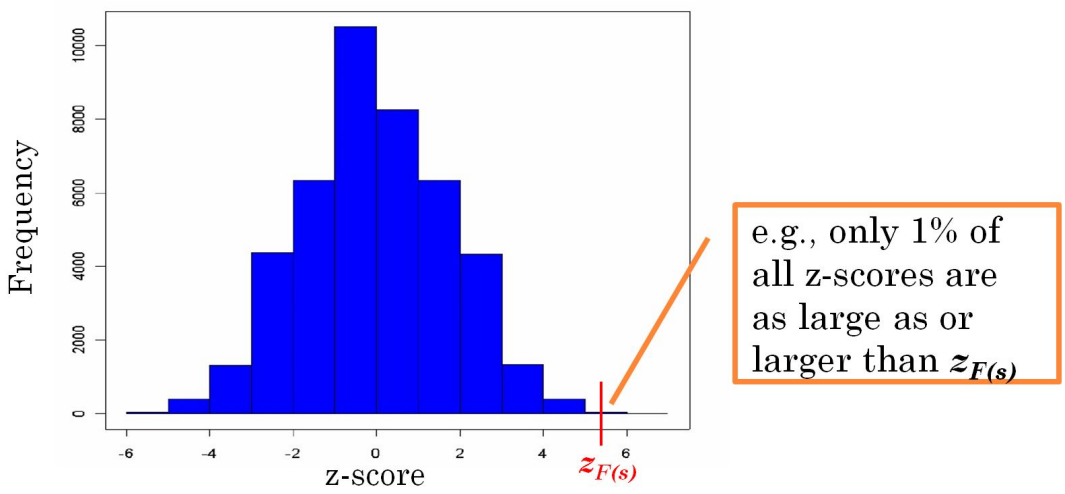

Figure 3.3: the distribution of z-scores of all random subsets

as a random subset. The z-scores of a few thousand random subsets are usually sufficient to get an accurate estimate of the p-value, and sampling can be done with or without replacement.

In Step 2, we check each s retained from Step 1 with the above mechanism. If there exists a subset of s, s', that satisfies: $F(s)$ does not cluster significantly better than a random selection from $F(s')$, then s is discarded; or else, it is a SCCP.

After finding all the SCCPs, we can further identify their respect hotspots with the traditional methods stated in Section 2.1, *e.g.*, using Getis-Ord Gi* statistic [28, 29].

The pseudocode for our SCCP discovery algorithm is given below. Usual set theoretic notation is employed. k-AVPS refers to a AVP set consisting of k AVPs. L_k is the set of frequent k-AVPS. C_k is the set of candidate k-AVPS, which are potentially frequent k-AVPS.

Algorithm 1 SCCP discovery algorithm

Input: collision dataset F, frequency threshold ϵ, z-score threshold γ, significance level α, number of Monte Carlo sampling runs N

Output: set of SCCPs

$L_1 \leftarrow$ {frequent 1-AVPS in F}
$k \leftarrow 2$
while $L_{k-1} \neq \emptyset$ **do**
    $L_k \leftarrow \emptyset$
    $C_k \leftarrow \{a \cup \{b\} | a \in L_{k-1} \wedge b \in \cup L_{k-1} \wedge b \notin a\}$
    for all k-AVPS $c \in C_k$ **do**
        $isEligible \leftarrow True$
        for all $(k-1)$-subsets s of c **do**
            if $s \notin L_{k-1}$ **then**
                $isEligible \leftarrow False$
                break
            end if
        end for
        if $isEligible == True$ **then**
            $frequency \leftarrow$ number of corresponding collisions of c in F
            if $frequency \geq \epsilon$ **then**
                add c into L_k
            end if
        end if
    end for
    $k \leftarrow k + 1$
end while
$L \leftarrow L_1 \cup L_2 \cup ... \cup L_k$
for all AVP set $s \in L$ **do**
    $z_{F(s)} \leftarrow$ z-score of $F(s)$
    if $z_{F(s)} < \gamma$ **then**
        delete s from L
    end if
end for
for all AVP set $s \in L$ **do**
    $z_{F(s)} \leftarrow$ z-score of $F(s)$
    for all subsets s' of s **do**
        generate N random subsets with the same size as $F(s)$ from $F(s')$
        $n \leftarrow$ number of random subsets whose z-scores $\geq z_{F(s)}$
        $p \leftarrow n/N$
        if $p > \alpha$ **then**
            delete s from L
            break
        end if
    end for
end for
Return L

Chapter 4

Experiments and Discussions

To evaluate our proposal, we implemented our method using the Python programming language and ArcGIS software, and applied it to Edmonton's historical collision data. In the following sections, we will first introduce our experimental setup, including the collision data we used, data pre-processing and parameter settings. Following that, we will present our experimental results and discuss on them.

4.1 Experiment Setup

4.1.1 Data used

The collision data we used are obtained from the Motor Vehicle Collision Information System (MVCIS) maintained by the Office of Traffic Safety, City of Edmonton. From this system we extracted the 6 years of collision data available, from 2006 to 2011. This dataset is stored as a table. Each row is a record of a collision and each column is an attribute of collisions. This table contains both spatial and non-spatial information about collisions, including the geographic location information, cause, severity, *etc*. Table 4.1 shows the column headings of this table and their descriptions.

For each collision, its location is geo-coded with GIS technology and the geographic location information is stored in the *X Coordinate* and *Y Coordinate* attributes. With GIS software, the spatial distribution of collisions can be displayed on a map, where each collision can be represented by a point with a particular location. For example, the spatial distribution of Edmonton's collisions in 2011 is

Column Headings	Description	
RPT_NO	**Report No. (YYYY-4~6 digit No.)**	
AV_NUM	Avenue Code (only number)	001-299: Regular 300-899: Named Road 900: Mapped location & bridge
AV_LOC	Avenue Code (w/ number & sub-code)	
ST_NUM	Street Code (only number)	
ST_LOC	Street Code (w/ number & sub-code)	
QUAD	Quadrant (SW, NE), need to fill in NW	
AV_POR	Avenue Portion Code	
ST_POR	Street Portion Code	
PORT	Portion Code (##)	
NA1_FUL	Full Name of named road	
NA2_FUL	Full Name of named road	
NA1_TY	Road type code (S-Road, M-Midblock, B-Bridge)	
NA2_TY		
N12_FUL	Full Name of Two Roads	
LOC_CODE	**Location Code (########-$$##-#)**	
EPS_VIS	EPS Visit to scene	
DATE	Collision Date (YYYYMMDD)	
DAY_WK	Day of week	
TIME	Time	
SEV	Severity	
TOT_PD	Total Property Damage	
CAUSE	Collision Cause	
SUR_CON	Surface condition	
TR_CONT	Traffic Control	
COL_ID	**Collision ID (######)**	
X_COORD	X Coordinate	
Y_COORD	Y Coordinate	
PORTION_DESC	Portion Description Code	
DUP_FLAG	Description of how location was selected	

Table 4.1: column headings of the collision table

displayed in Figure 4.1.

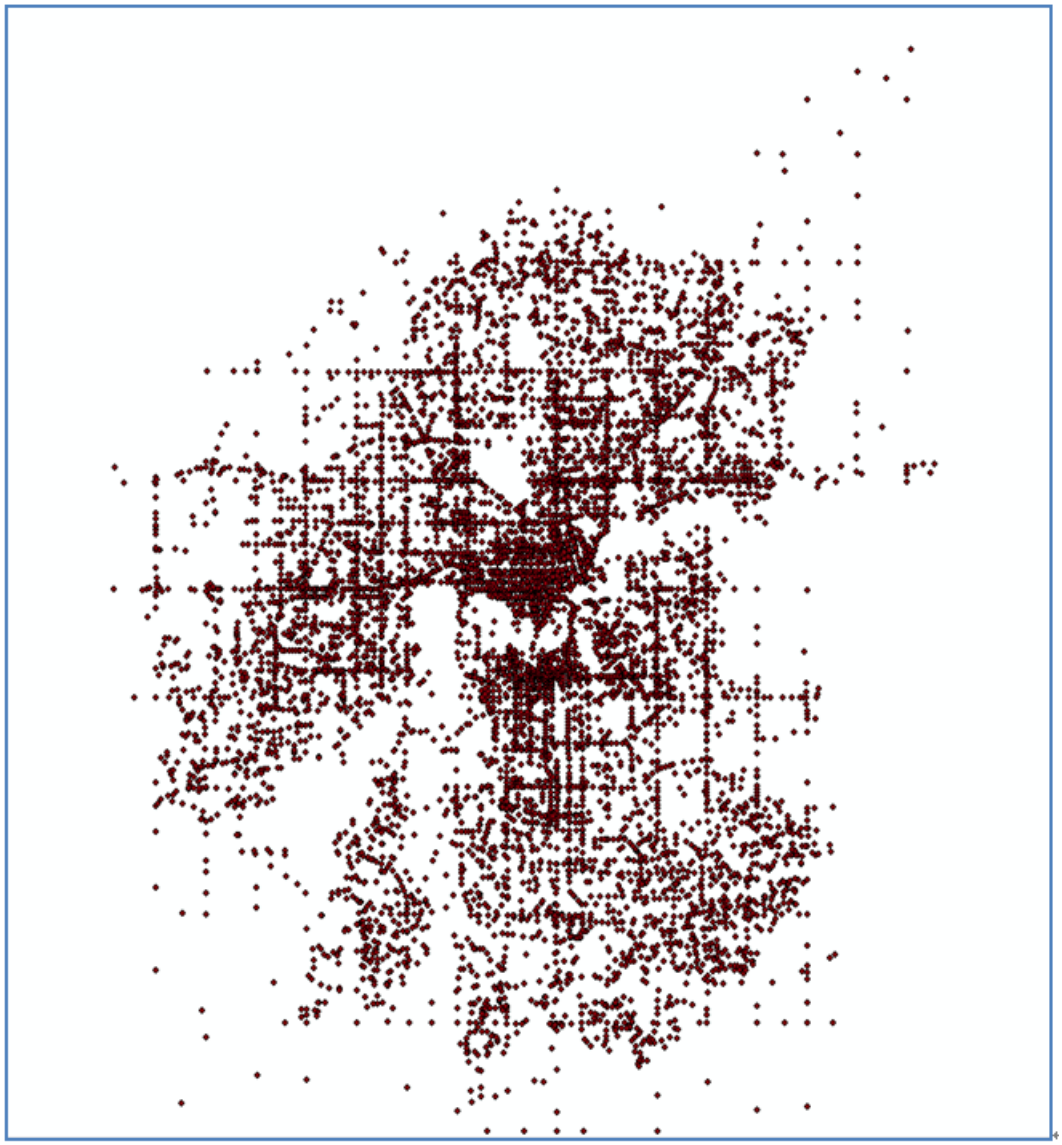

Figure 4.1: The spatial distribution of Edmonton's collisions in 2011

Apart from the geographic location information, the following non-spatial attributes are available for analysis:

- *Severity*
- *Cause*
- *Road surface condition*

- *Traffic control*
- *Time*
- *Day of the week*
- *Date*

Each attribute above has a set of possible values. Taking the attribute *Road Surface Condition* as an example, its possible values are 'Dry', 'Wet', 'Loose Sand/Dirt/Gravel', 'Snowy/Icy', and 'Other'.

4.1.2 Data pre-processing

Cleaning collision data with missing information

In the data collection process of MVCIS, some information of collisions was missing. For some collisions, the location information was not reported. For some others, certain non-spatial information was not collected, such as *Cause, Time, etc*. Therefore, in our dataset extracted from MVCIS, a considerable amount of collision records are incomplete.

For our Spatial Co-Clustering Pattern (SCCP) discovery method, we should treat collisions with missing location information and missing non-spatial information differently:

For collisions with missing non-spatial information, the missing values of certain non-spatial attributes will not affect the analysis of AVP sets which only contains other non-spatial attributes. For example, if a collision's *Cause* attribute value is unknown but *Day of the week* and *Road surface condition* attributes are known, then it can still be included for the analysis of AVP set *Day of the week = 'Saturday' and Road surface condition = 'Dry'*. However, the missing values of non-spatial attributes will significantly affect the analysis of AVP sets that contains the same non-spatial attributes. For example, if a collision's *Cause* attribute value is unknown, then it cannot be included for the analysis of AVP set *Cause= 'Followed too close' and Day of the week = 'Saturday'*. In order to produce reliable analysis results, it should be excluded from the analysis for that AVP set. Based on the above

analysis, for missing values of attributes, we assign a special value 'Unknown' to them. This special value is different from any other known values of this attribute. Thus such collisions will be excluded from the analysis for AVP sets that contains the same non-spatial attribute.

For collisions with missing location information, the treatment is different. They cannot be used by our SCCP discovery method. The reason is that when we evaluate the distribution of collisions with Global Moran's I index, we must know every collision's location in order to calculate the distance between collisions. So we directly deleted these data from our dataset. Taking the collision data in 2011 as an example, 1705 collision records with missing location information are deleted from the overall 23442 records. The percentage is 7.27%.

Categorizing values for certain attributes

Since our SCCP discovery method need to enumerate all possible AVP sets and evaluate each of them, the computation amount highly depends on the number of possible AVP sets. Assume the collision table has a set of attributes $A : A_1, A_2, ..., A_n$, the number of possible values (except the special value 'Unknown') for A_i is $m_i (i = 1, 2, ..., n)$, then the number of all possible AVP sets is: $m_1 \times m_2 \times ... \times m_n$. For our original collision data, the number of possible values of each non-spatial attribute mentioned above is listed in Table 4.2

Attribute	The number of possible values
Severity	3
Cause	26
Road surface condition	4
Traffic control	14
Time	1440
Day of the week	7
Date	2191

Table 4.2: the number of possible values of non-spatial attributes

From Table 4.2, we can calculate the number of all possible AVP sets of the original data is: $3 \times 26 \times 4 \times 14 \times 1440 \times 7 \times 2191 = 96,468,503,040$, which

is very large. In order to reduce the computation amount of the SCCP discovery process, we need to decrease the number of possible AVP sets effectively.

Among the 7 non-spatial attributes listed in Table 4.2, the *Time* and *Date* attributes have a particularly large number of possible values. For the *Time* attribute, its values consist of 4 digits. The first two digits represent "hour" (ranging from '00' to '23') and the latter two represent "minute" (ranging from '00' to '59'). For example, a possible values of *Time* attribute is '0730', which means the collision happened at "7:30 am". So the number of all possible values of *Time* attribute is: $24 \times 60 = 1440$. For the *Date* attribute, its values consist of 8 digits. The first four digits represent "year" (ranging from '2006' to '2011'), the middle two digits represent "month" (ranging from '01' to '12'), and the last two digits represent "day" (ranging from '01' to '31'). For example, a possible values of *Date* attribute is '20110322', which means the collision happened on "March 22, 2011". The number of possible values of *Date* attribute is 365 for the years 2006, 2007, 2009, 2010, and 2011, and 366 for the year 2008. So the total number is: $365 \times 5 + 366 = 2191$.

In order to decrease the number of possible AVP sets, we categorized the values of the above two attributes into a much smaller number of classes. For the *Time* attribute, its values were categorized into the following 5 ranges : '0000-0659' (Before Dawn), '0700-0959' (AM Peak), '1000-1559' (Working Hours), '1600-1859' (PM Peak), and '1900-2359' (Evening). We created a new attribute called *Time segment*, which has the above 5 values, and derived the values for the *Time segment* attribute from the *Time* attribute. In this way, the 1440 possible values of *Time* attribute were categorized into the 5 values of *Time segment* attribute. For the *Date* attribute, we categorized its values according to the month it belonged to. In light of this, we created a new attribute called *Month*, whose values range from 01 to 12, and transferred the values of the *Date* attribute into the corresponding values of the *Month* attribute. In this way, the 2191 possible values of *Date* attribute were categorized into 12 classes.

After categorizing the values of these two attributes, the number of all possible AVP sets is decreased from $96,468,503,040$ to: $3 \times 26 \times 4 \times 14 \times 5 \times 7 \times 12 = 1,834,560$. In this way, the computation amount of of SCCP discovery process can

be reduced effectively.

After replacing the *Time* and *Date* attributes with the corresponding *Time segment* and *Month* attributes, the list of non-spatial attributes available for analysis is:

- *Severity*
- *Cause*
- *Road surface condition*
- *Traffic control*
- *Time segment (derived from the original Time attributes)*
- *Day of the week*
- *Month (transferred from the original Date attributes)*

4.1.3 Parameter settings

Setting frequency threshold

As mentioned in Section 3.2.1, in Step 1 of our SCCP discovery method, the first task is to find all "frequent" AVP sets. In order to reduce the computation amount of the SCCP discovery process, we ignore those "infrequent" AVP sets whose corresponding set of collisions is too small, and don't calculate their Global Moran's I indices and the associated z-scores.

To determine whether an AVP set is infrequent or not, we need to pre-specify a threshold, called frequency threshold. If the numbers of corresponding collisions of an AVP set is smaller than the frequency threshold, this AVP set is infrequent; else, it is frequent.

To set a proper frequency threshold, we need to consider the total number of collisions as well as the computation amount. On the one hand, we want the threshold to be reasonably small compared to the total number of collisions, so that we may not miss some potentially interesting AVP sets. On the other hand, we also want the threshold to be large, so that we can ignore a larger number of infrequent AVP sets

and reduce the computation amount more effectively. Therefore, we need to make a trade-off between the risk of missing interesting patterns and the computation time when setting the frequency threshold.

Taking Edmonton's collisions in 2011 as an example, after cleaning collisions with missing location information, the total number of remaining collisions is $21,737$. After discussing with the Office of Traffic Safety (OTS) staff, we think it is reasonable to set the frequency threshold below 500 (2.3% of the total number). AVP sets that correspond to fewer than 500 collisions deserve less attention, and thus we directly ignore them.

To investigate how different frequency thresholds will change the computation amount and the results of the SCCP discovery process, we used 200 and 500 as two sample frequency thresholds, and compared the computation time for finding all frequent AVP sets as well as the number of the frequent AVP sets found when using different numbers of attributes. We first only used the following four attributes: *Severity, Cause, Road surface condition, Day of the week*, and got the experimental results in Table 4.3. Then we added the attribute *Traffic control*, and got the experimental results in Table 4.4. Furthermore, we added the attribute *Month*, and got the experimental results in Table 4.5. At last, we added another attribute *Time segment*, and got the experimental results in Table 4.6.

Frequency threshold	Computation time (hour)	The number of frequent AVP sets
500	0.22	122
200	0.49	253

Table 4.3: Experimental results on sample frequency thresholds when using 4 attributes

Figure 4.2 and Figure 4.3 illustrate how the computation time for finding all frequent AVP sets and the number of the frequent AVP sets found increase with the number of attributes used when using 200 and 500 as the frequency thresholds. From these figures we can see that when using the same number of attributes, a larger frequency threshold leads to smaller computation time and a smaller number of frequent AVP sets. Besides, when the number of attributes used increases, the

Frequency threshold	Computation time (hour)	The number of frequent AVP sets
500	1.33	252
200	4.77	581

Table 4.4: Experimental results on sample frequency thresholds when using 5 attributes

Frequency threshold	Computation time (hours)	The number of frequent AVP sets
500	3.27	382
200	29.22	1080

Table 4.5: Experimental results on sample frequency thresholds when using 6 attributes

Frequency threshold	Computation time (hours)	The number of frequent AVP sets
500	7.03	603
200	224.79	1947

Table 4.6: Experimental results on sample frequency thresholds when using 7 attributes

computation time and number of frequent AVP sets for a larger frequency threshold increase much slower than those for a smaller frequency threshold.

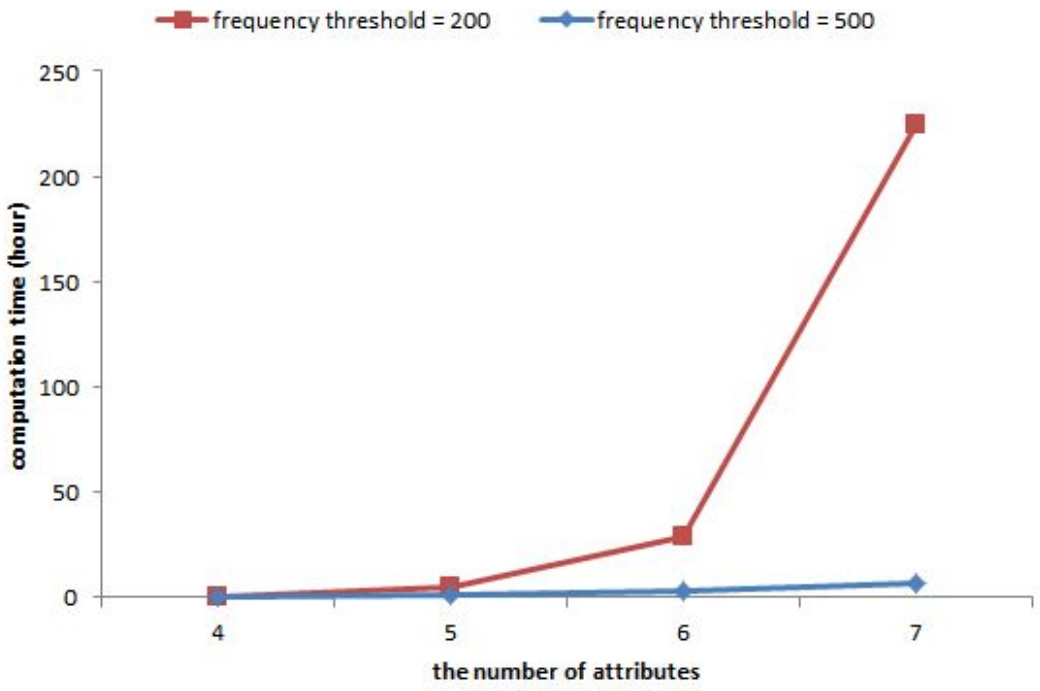

Figure 4.2: Computation time for sample frequency thesholds when using different number of attributes

In our following experiment on SCCP discovery, we use all the 7 available attributes mentioned in Section 4.1.2. In order to keep the computation time at a feasible level, we set the frequency threshold to 500. With this setting, the computation time is 7.03 hours and the number of the frequent AVP sets found is 603. In the next step, we calculate the Global Moran's I index and the associated z-score for each of the 603 AVP sets, to evaluate whether its corresponding collisions cluster well in space.

Choosing the distance function for spatial autocorrelation measures

In our Spatial Co-Clustering Pattern (SCCP) discovery method, we use Global Moran's I index to evaluate the clustering degree of collisions in Step 1. After finding all the SCCPs, we further identify their respective hotspots with the traditional spatial autocorrelation methods stated in Section 2.1.3, *e.g.*, using Getis-Ord Gi* statistic [28, 29]. Both Global Moran's I index and Getis-Ord Gi* statistic in-

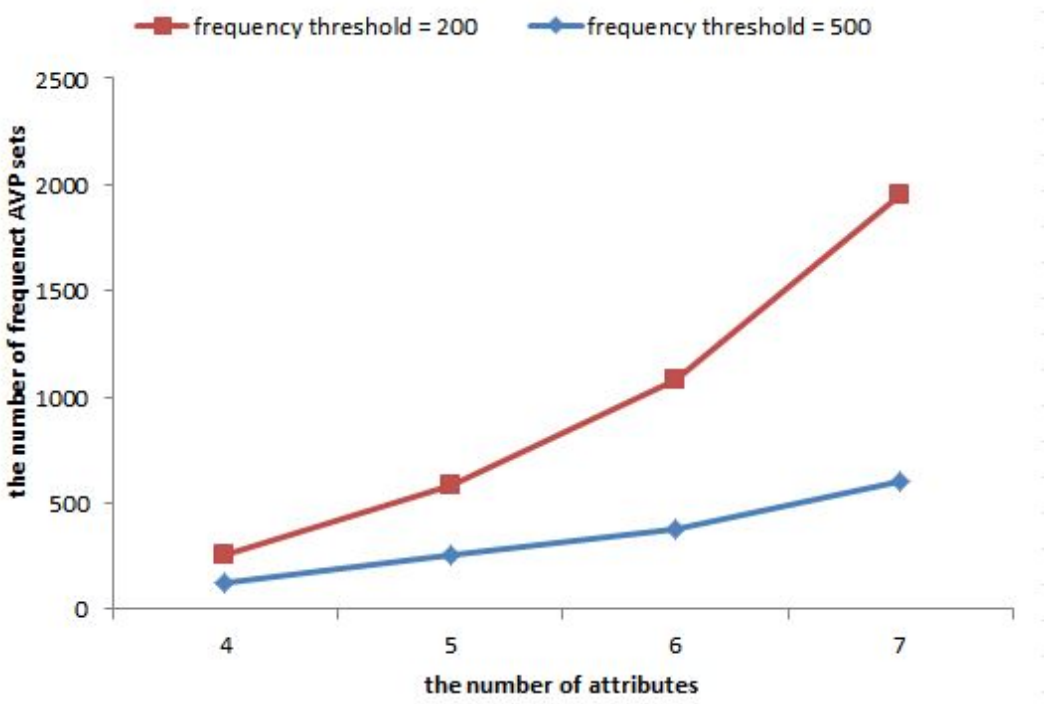

Figure 4.3: The number of frequent AVP sets for sample frequency thesholds when using different number of attributes

volve distances in their calculations. For example, the expression of the Getis-Ord Gi* statistic is restated here:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{[n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2]}{n-1}}}$$

In this expression, $w_{i,j}$ is the spatial weight between location point i and j. It is usually a function of the spatial distance $d_{i,j}$ between point i and j, *i.e*, $w_{i,j} = \frac{1}{d_{i,j}}$. (This is discussed in detail in the following subsection.) Therefore, before calculating Global Moran's I index or Getis-Ord Gi* statistic, we must first decide how to define the spatial distance $d_{i,j}$ between point i and j.

Let the geographic coordinates of point i be (x_i, y_i), the geographic coordinates of point j be (x_j, y_j). Then the spatial distance $d_{i,j}$ is a function of x_i, y_i, x_j, y_j. The two most common distance functions used in spatial statistics are Euclidean distance and Manhattan distance.

Euclidean distance is calculated as:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4.1}$$

Manhattan distance is calculated as

$$d_{i,j} = |x_i - x_j| + |y_i - y_j| \tag{4.2}$$

Generally speaking, Manhattan distance is more appropriate than Euclidean distance when travel is restricted to a street network, which is the case in our application on collision data analysis. In our application, the travel distance between two collision locations is more appropriately the Manhattan distance between them. So in our following experiment, we use Manhattan distance to calculate the spatial weight between two location points.

Modeling spatial weights

In this section, we discuss how to model the spatial weight $w_{i,j}$ between point i and j based on the spatial distance $d_{i,j}$ between them. In the traffic area, the most common models include inverse distance, fixed distance band and zone of indifference. A brief introduction of these three models is given as following:

(1) Inverse distance model:

This model uses the inverse distance as the spatial weight. The expression of $w_{i,j}$ is given as:

$$w_{i,j} = \frac{1}{d_{i,j}^{\delta}} \tag{4.3}$$

where δ can be assigned any appropriate value such as 0.5, 1.0, 1.5, 2.0.

In this case, the farther away a neighbor is from a point, the smaller the impact it has on this point. The relationship between the spatial weight and the spatial distance under this model (with $\delta = 1.0$) is illustrated in Figure 4.4.

(2) Fixed distance band model:

In this model, neighbors within a specified distance band d_0 (*e.g.*, 1 km) are weighted equally, while features outside the specified distance don't influence calculations (their weight is zero). The expression of $w_{i,j}$ is given as:

$$w_{i,j} = \begin{cases} 1 & \text{if } d_{i,j} \leq d_0 \\ 0 & \text{if } d_{i,j} > d_0 \end{cases} \tag{4.4}$$

The relationship between the spatial weight and the spatial distance under this model is illustrated in Figure 4.5.

Figure 4.4: The relationship between the spatial weight and the distance under inverse distance model (with $\delta = 1.0$)

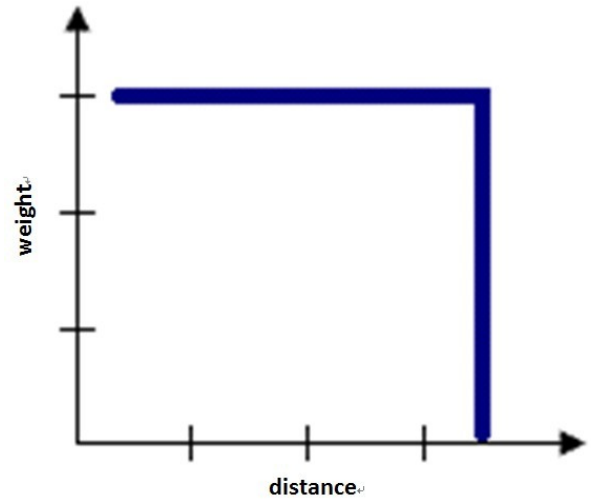

Figure 4.5: The relationship between the spatial weight and the distance under fixed distance band model

(3) Zone of indifference model:

This model is a combination of the above two models. Neighbors within the specified distance band d_0 (*e.g.*, 1 km) are weighted equally. Once the specified distance is exceeded, the level of influence (the weight) drops off quickly. The expression of $w_{i,j}$ is given as:

$$w_{i,j} = \begin{cases} 1 & \text{if } d_{i,j} \leq d_0 \\ \frac{1}{(d_{i,j}-d_0+1)^{\delta}} & \text{if } d_{i,j} > d_0 \end{cases} \tag{4.5}$$

where δ can be assigned any appropriate value such as 0.5, 1.0, 1.5, 2.0.

The relationship between the spatial weight and the spatial distance under this model (with $\delta = 1.0$) is illustrated in Figure 4.6.

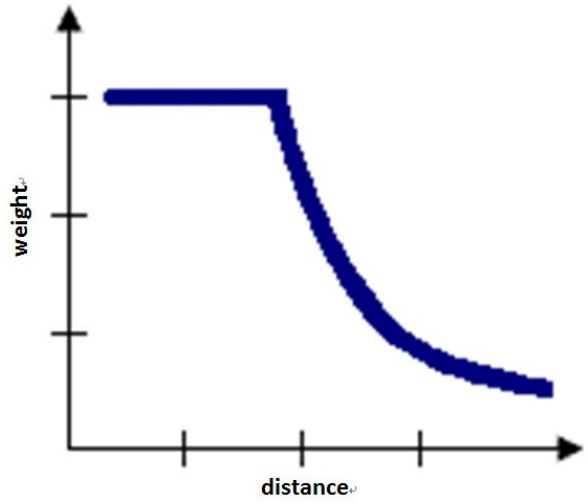

Figure 4.6: The relationship between the spatial weight and the distance under zone of indifference model (with $\delta = 1.0$)

All the three models are implemented in the ArcGIS software we used. For a specific application, the more realistically we model how features interact with each other in space, the more accurate our results will be. In practice, we need to choose the model that best fit our goal of analysis.

To better understand the difference between these three models, we did some experiments on our collision data to compare their respective influences on the output hotspot analysis results. Taking the collisions on Saturdays (*Day of the week = 'Saturday'*) in 2011 as an example, when setting δ as 1.0 and d_0 as 1 km, we obtained different hotspot analysis results with the three models, presented in Figure 4.7.

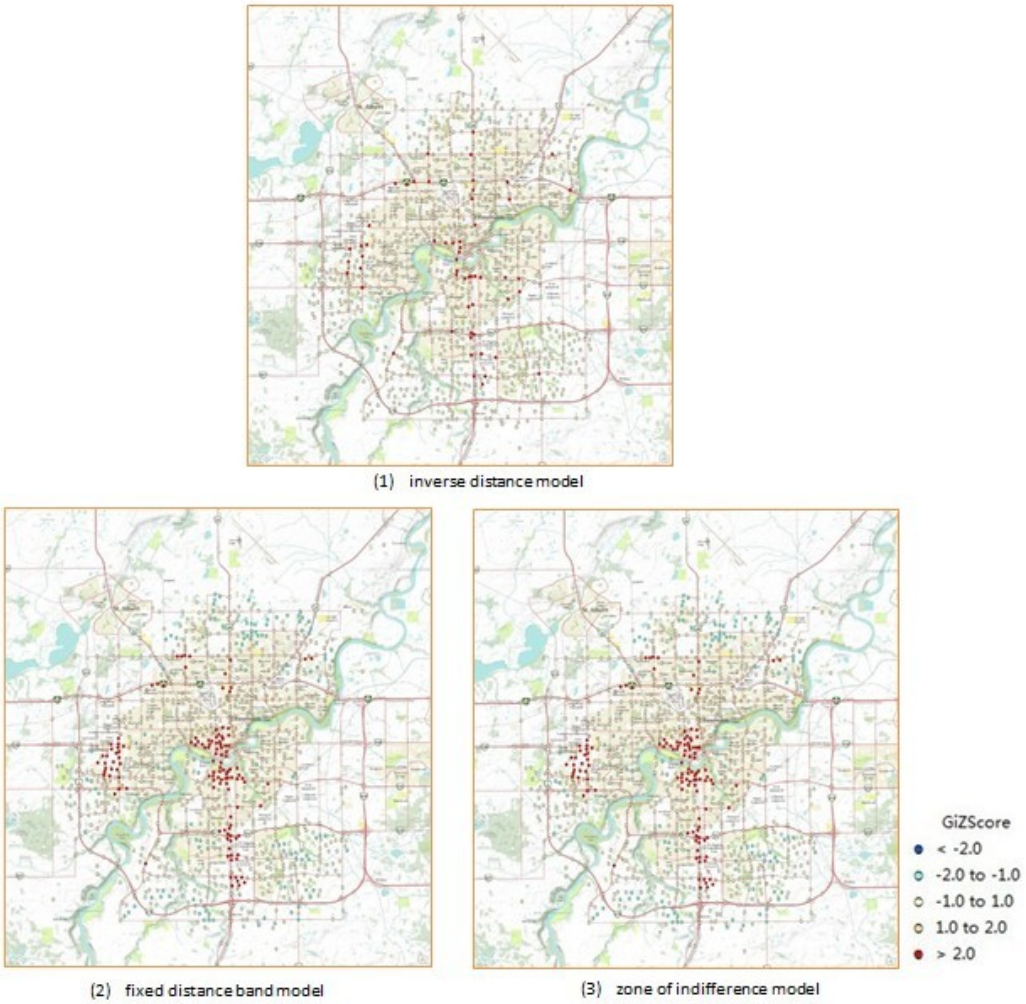

Figure 4.7: hotspot analysis results on collsions that happened on Saturdays in 2011, with three spatial weight models used

From the hotspot analysis results in Figure 4.7, we notice that the fixed distance band model and the zone of indifference model lead to almost the same hotspot analysis results on our collision data. Besides, we can see that the inverse distance model lead to *"pin-point"* hotspots, while the fixed distance band model and the zone of distance model leads to hotspot *regions*.

From the above discussion, we draw the following conclusion:

If we want to differentiate individual location points and get precise hotspot *points*, then the inverse distance model is a better choice for us. However, if we want to consider points within a neighborhood (*e.g.*, 1 km) as a whole and not differentiate between them, or if we want to find hotspot *regions* instead of points, then the fixed distance band model or the zone of indifference model is a better choice.

After discussion with the Office of Traffic Safety (OTS) staff in Edmonton, we decided to use the fixed distance band model, based on the following reasons. First, the OTS staff are more interested in the hotspots regions of various types of collisions than hotspot points. Second, considering our large computation amount, the fixed distance band model is preferable to the zone of indifference model because it is simpler and involves less computation. So in our following experiment, we use the fixed distance band model to calculate the spatial weight between two location points.

Setting distance band

After determining the spatial weight model, we still need to set the distance band d_0. The distance band we choose determines the scale of our analysis. Its value depends on the goal of a specific application. In collision hotspot analysis as an example, we may be interested in hotspot patterns at different levels, such as neighborhood patterns, regional patterns, city wide patterns, *etc*.

Figure 4.8 displays the hotspot analysis results on the same collision data with various distance bands. The collision data used here is the set of collisions that happened on Saturdays (*Day of the week ='Saturday'*) in 2011. The spatial weight model is the fixed distance band model.

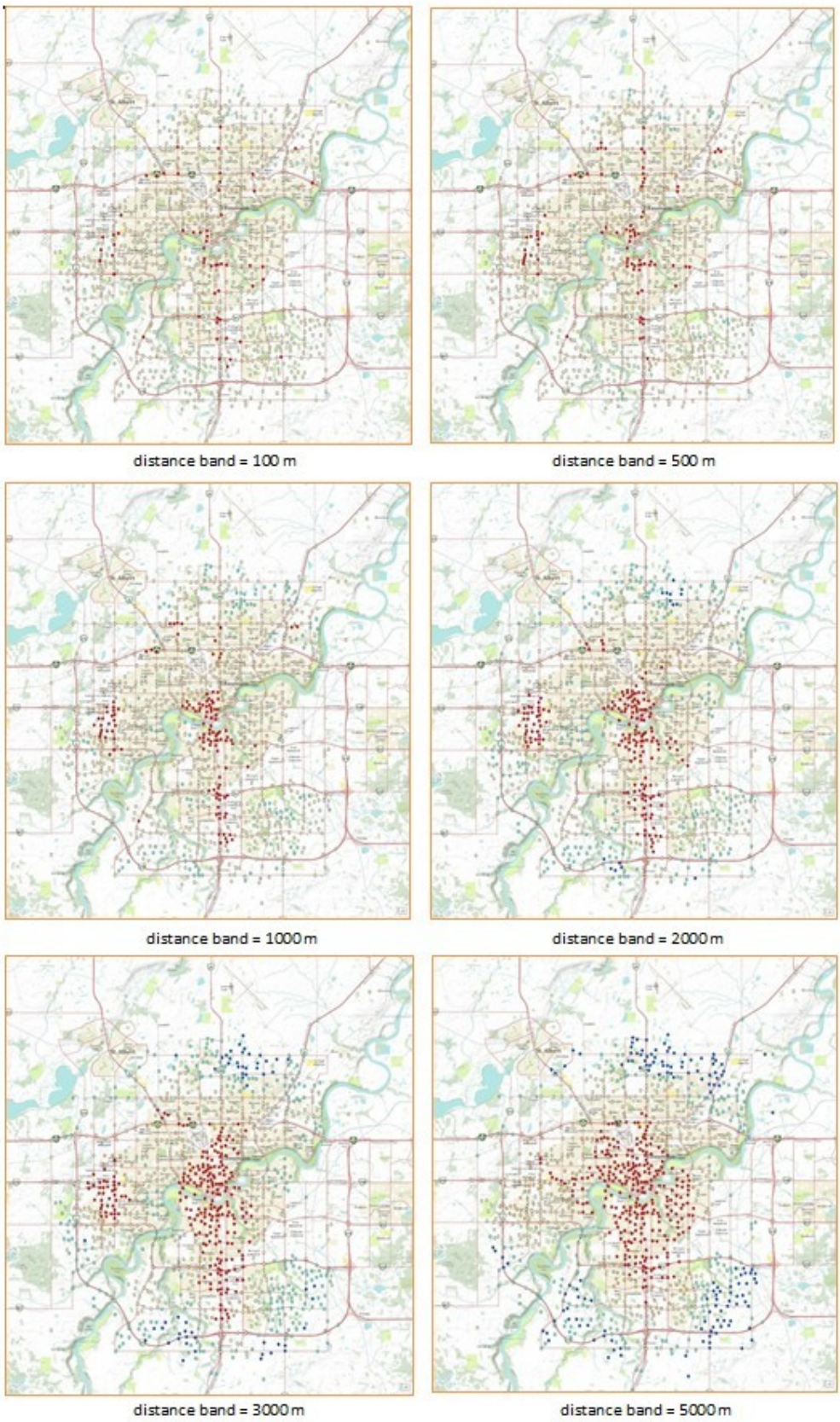

Figure 4.8: hotspot analysis results on collisions that happened on Saturdays in 2011, with various distance bands used

When the distance band increases from 100m to 5000m, the scale of our analysis increases correspondingly. When the distance band is as small as 100m, only neighbors within this small distance are involved in the evaluation for hotspots. As a result, the output hotspot patterns are at a small neighborhood level. In contrast, when the distance band is as large as 5000, all neighbors within this large distance are involved in the evaluation of hotspots, and they are all equally weighted. As a result, the output hotspot patterns are at a large regional level. In this extreme case, only one large hotspot region is recognized, which is located in the middle of the city.

After discussion with the Office of Traffic Safety (OTS) staff in Edmonton, we think a distance band between 1 km and 1.5 km is most informative for our analysis. In our following experiment, we use 1 km as the distance band.

4.2 Experimental Results and Discussions

With the experimental setup we described in Section 4.1, we applied our SCCP discovery method to Edmonton's historical collision data in 2011. In Step 1, with the frequency threshold $\epsilon = 500$ and the z-score threshold $\gamma = 5.0$, we found 52 frequent AVP sets that lead to spatial clustering of collisions. Then in step 2, with the significance level $\alpha = 0.1$, 14 of those 52 AVP sets passed the test to be SCCPs. These 14 SCCPs are listed in Table 4.7.

These SCCPs revealed that collisions with certain attribute values clustered particularly well in space. In other words, certain specific types of collisions happened particularly frequently at some particular locations (hotspots). As all subsets of a SCCP were verified to contribute to the collision clustering, we also obtained valuable information for explaining the frequent collision occurrences at certain hotspots. After finding all the SCCPs, we then further identified their respective hotspots with the traditional spatial autocorrelation method, using Getis-Ord Gi* statistic.

In the following we present a few sample results of the detected SCCPs and their corresponding hotspots from Edmonton's collision data in 2011.

The detected SCCPs from Edmonton's collision data in 2011
{*Day of the week = 'Saturday'*}
{*Severity = 'Property Damage Only'*}
{*Traffic control = 'No Control Present'*}
{*Time segment= '1000-1559' (Working Hours)*}
{*Cause= 'Failed to Observe Traffic Signal'*}
{*Cause= 'Struck Parked Vehicle'*}
{*Day of the week = 'Thursday'*}
{*Day of the week = 'Saturday', Road surface condition = 'Dry'*}
{*Severity = 'Property Damage Only', Time segment= '1000-1559' (Working Hours)*}
{*Road surface condition = 'Dry', Traffic control = 'No Control Present'*}
{*Cause= 'Followed too Close', Traffic control = 'No Control Present'*}
{*Day of the week = 'Saturday', Time segment= '1600-1859' (PM Peak)*}
{*Road surface condition = 'Dry', Cause= 'Struck Parked Vehicle'*}
{*Cause= 'Followed too Close', Traffic control = 'No Control Present', Time segment= '0700-0959' (AM Peak)*}

Table 4.7: the detected SCCPs from Edmonton's collision data in 2011

Example 1: the detected SCCP is {*Day of the week = 'Saturday', Road surface condition = 'Dry'*}. Its corresponding hotspots are displayed in Figure 4.9. The major hotspot regions are tagged with geographic information.

From the analysis result in Figure 4.9, we see that the hotspots of collisions under the scenario "Day of the week = 'Saturday', Road surface condition = 'Dry'" are mainly located at major shopping areas in Edmonton, such as West Edmonton Mall, Southgate Mall and downtown area. In addition, the SCCP indicates that the day of Saturday and the dry road surface are two significant contributing factors to the clustering of such collisions, which can help explain the formation of those hotspots. With the knowledge discovered above, the City of Edmonton's Office of Traffic Safety can, for example, assign more traffic police to the above five shopping areas on a sunny Saturday to patrol and implement traffic enforcement, in order to decrease the collisions in these hotspot regions.

Example 2: the detected SCCP is {*Road surface condition = 'Dry', Cause= 'Struck Parked Vehicle'*}. Its corresponding hotspots are displayed in Figure 4.10. The major hotspot regions are tagged with geographic information.

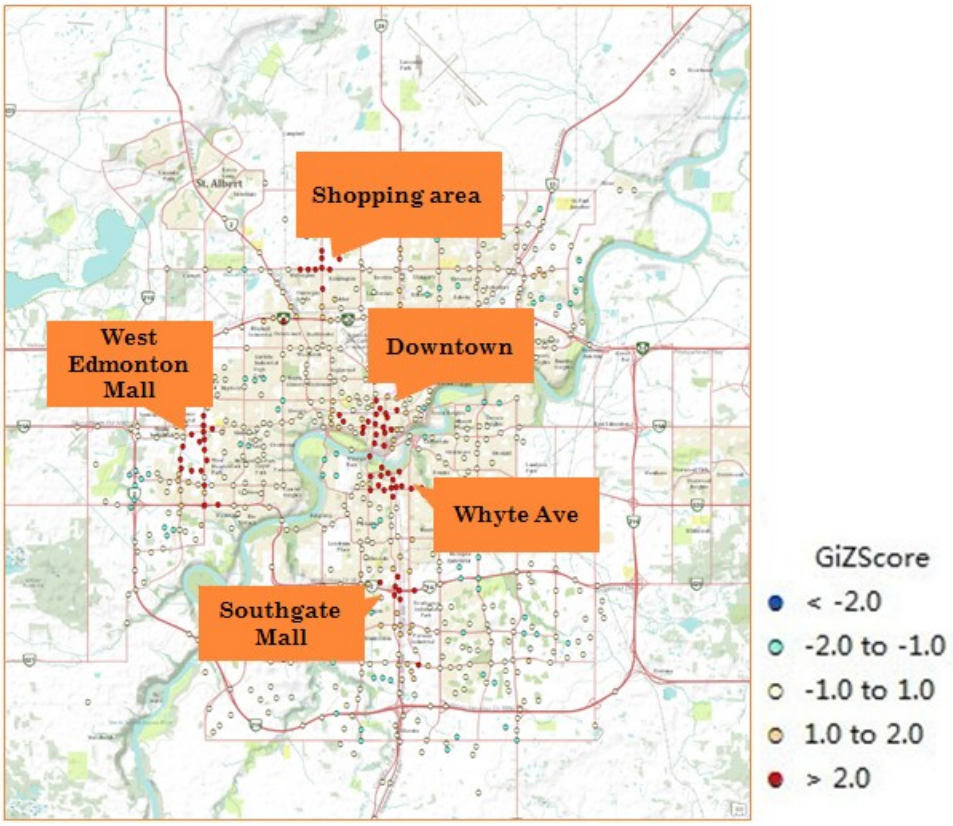

Figure 4.9: the hotspot analysis result for collisions with {*Day of the week = 'Saturday', Road surface condition = 'Dry'*}

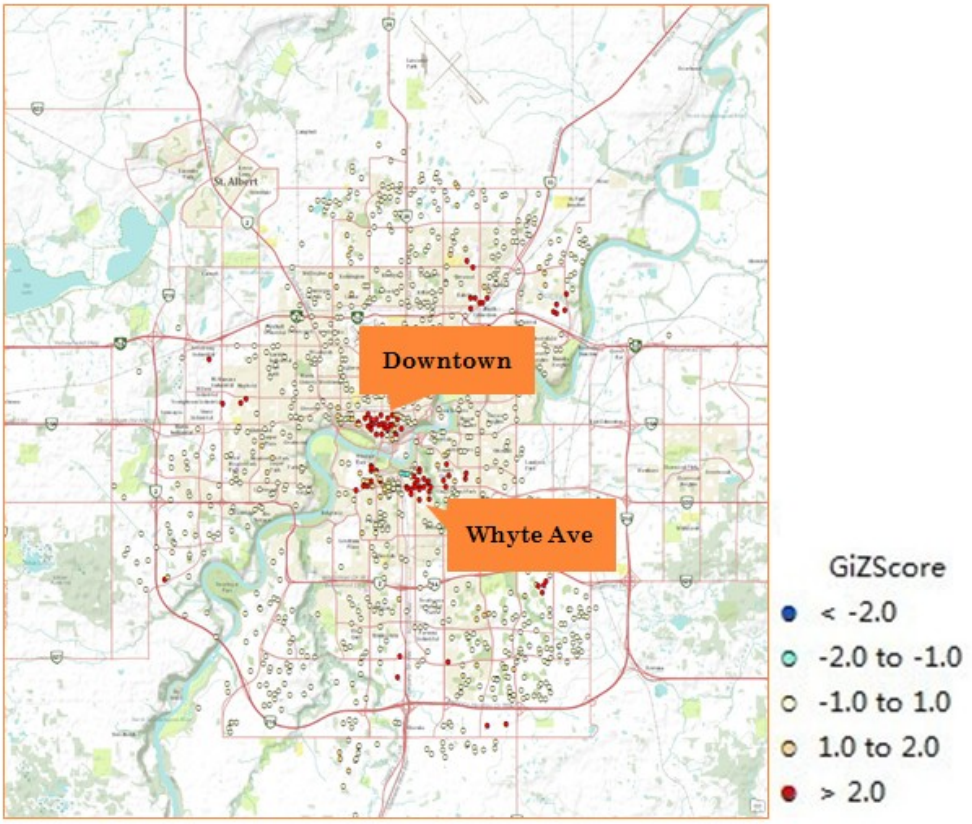

Figure 4.10: the hotspot analysis result for collisions with {*Road surface condition ='Dry', Cause='Struck Parked Vehicle'*}

From the analysis result in Figure 4.10, we see that the hotspots of collisions under the scenario "Road surface condition = 'Dry', Cause='Struck Parked Vehicle'" are mainly located in the downtown area and on Whyte Avenue in Edmonton. In addition, the SCCP indicates that both the dry road surface and the cause 'Struck Parked Vehicle' contribute to the clustering of those collisions significantly, which can guide us to investigate the formation of those hotspots. With the knowledge discovered above, the City of Edmonton's Office of Traffic Safety can further survey the parking areas of these two regions and take proper measures to decrease this type of collisions in future.

Example 3: the detected SCCP is {*Cause='Followed too Close', Traffic control='No Control Present', Time segment='0700-0959' (AM Peak)*}. Its corresponding hotspots are displayed in Figure 4.11. The major hotspot regions are tagged with geographic information.

From the analysis result in Figure 4.11, we see that the hotspots of collisions under the scenario "Cause='Followed too Close', Traffic control='No Control Present', Time segment='0700-0959' (AM Peak)" are mainly located at the two high-level bridges and two intersections along the Yellowhead Trail. In addition, the SCCP indicates that each of the three AVPs makes a significant contribution to the clustering of those collisions, which can help explain the formation of those hotspots. With the knowledge discovered above, the City of Edmonton's Office of Traffic Safety can, for example, improve the traffic control condition at these four hotspot locations or implement traffic enforcement there during the AM peak hours, in order to decrease this type of collisions at these four locations.

In addition, by comparing the analysis results from each single year's data, we can also investigate the trend of the hotspots of a particular type of collisions. Taking the type of collisions in *Example 3* as an example, its trend of hotspots is discribed in *Example 4*.

Example 4: hotspot analysis results for collisions with "*Cause='Followed too Close', Traffic control='No Control Present', Time segment='0700-0959' (AM Peak)*" in the past years; results are displayed in Figure 4.12.

From Figure 4.12, we can see how the hotspots of collisions under the sce-

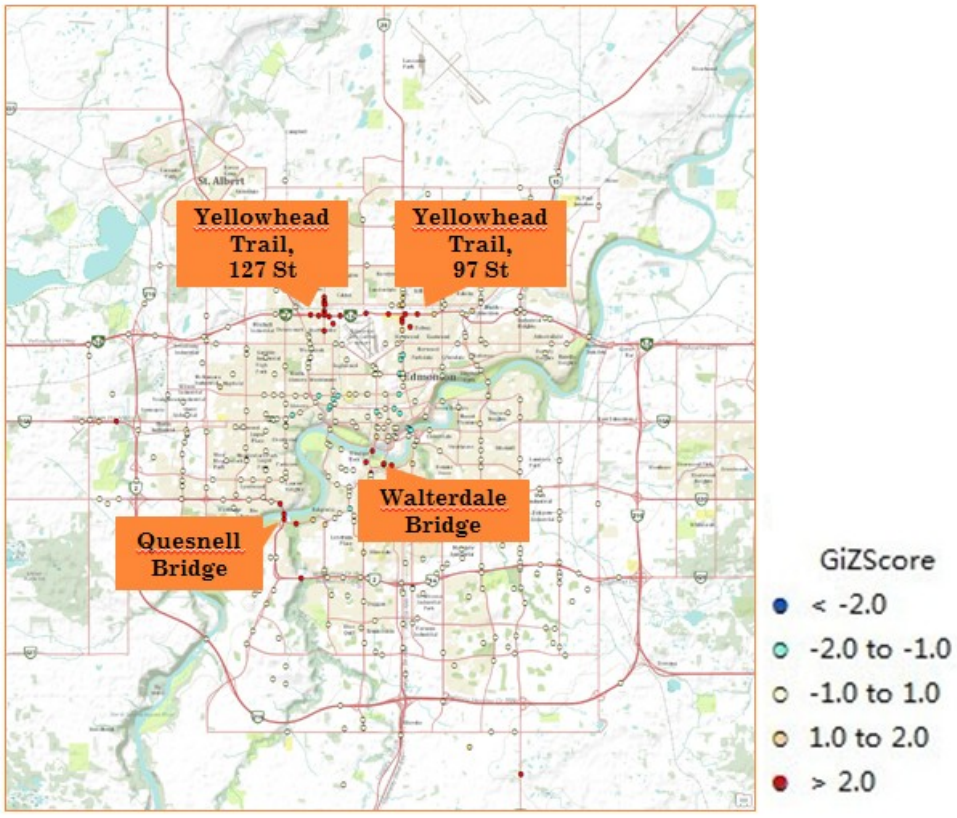

Figure 4.11: the hotspot analysis result for collisions with {*Cause='Followed too Close', Traffic control='No Control Present', Time segment='0700-0959' (AM Peak)*}

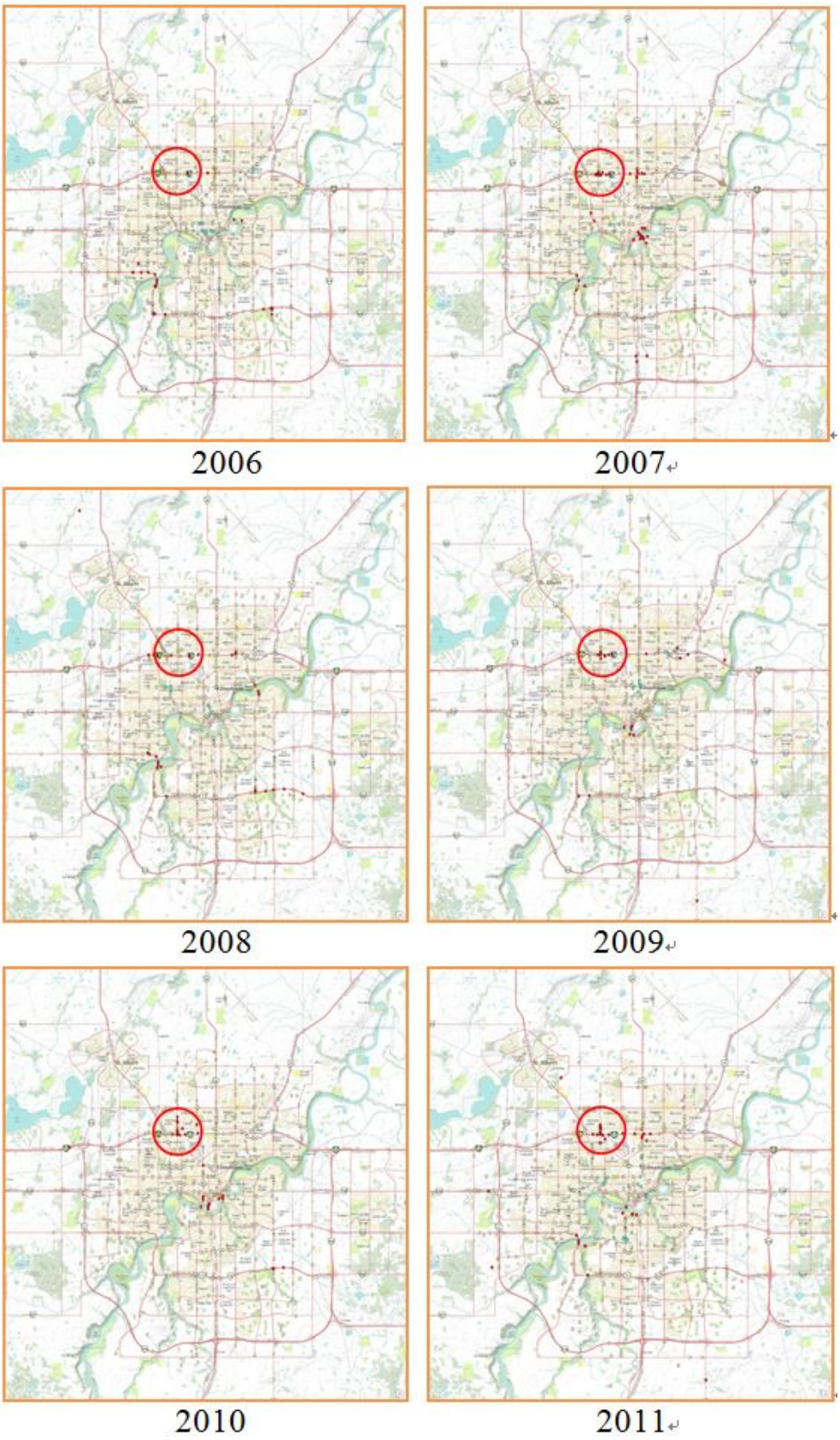

Figure 4.12: the hotspot analysis results for collisions with "*Cause='Followed too Close', Traffic control='No Control Present', Time segment='0700-0959' (AM Peak)*" in the past years

nario “Cause=‘Followed too Close’, Traffic control=‘No Control Present’, Time segment=‘0700-0959’ (AM Peak)” changed in the past several years. For example, for the hotspot region within the red circle in Figure 4.12, it seems to expand in the recent years.

In summary, the results of our experiments show that our method can discover a larger number of meaningful hotspot patterns in Edmonton’s historical collision data than the traditional methods. Compared to the traditional methods which only detect the hotspots of the overall collisions, our method discovers the hotspot patterns of various specific types of collisions, which are most interesting according to the criteria given in Section 3.1.1. In addition, our method also reveals the relevant AVPs that contribute to the clustering of those collisions, adding valuable indicators for explaining certain hotspots.

Chapter 5

Conclusions and Future Work

Hotspot analysis methods are attractive for understanding the spatial patterns of collisions, which is critical for improving the efficiency and effectiveness of traffic enforcement deployment as well as road safety. However, most of the traditional hotspot analysis methods only focus on the location information of the collision data, without integrating the abundant non-spatial attributes into the analysis. Taking non-spatial attributes into account, however, opens opportunities to reveal attribute-related hotspots that may otherwise go undetected, and adds valuable indicators for explaining certain hotspots.

In this thesis, we introduced the concept of a Spatial Co-Clustering Pattern (SCCP), which is a set of non-spatial AVPs that together contribute significantly to the spatial clustering of the corresponding collisions. Then we presented our SCCP discovery method. By applying our method on Edmonton's historical collision data, we discovered a number of SCCPs. We then further identified their respective hotspots with the traditional spatial autocorrelation method. The experimental results showed that our method can discover a larger number of meaningful hotspot patterns in Edmonton's historical collision data than the traditional methods, and reveal relevant non-spatial indicators for explaining certain hotspots. These discoveries have the potential to allow more effective and efficient deployment of resources for traffic enforcement and road safety.

Though our method is implemented to analyze collision data, it can also be applied to other spatial data with non-spatial attributes, such as crime data. Future research will consider applying this method in other application areas. In addition,

though we already applied some techniques to reduce the amount of computation, the current method is still computationally expensive. In future, new techniques for computation reduction should be investigated. Finally, our method can currently only analyze static historical data. A future challenge is to design a mechanism for incremental update of SCCPs.

Bibliography

[1] Trevor Bailey and Anthony Gatrell. *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex, 1995.

[2] Margie Peden, Richard Scurfield, David Sleet, Dinesh Mohan, Adnan Hyder, Eva Jarawan, and Colin Mathers. World report on road traffic injury prevention. Technical report, World Health Organization, Geneva, Switzerland, 2004.

[3] Elizabeth Kopits and Maureen Cropper. Traffic fatalities and economic growth. *Accident Analysis & Prevention*, 37(1):169 – 178, 2005.

[4] Christopher Murray and Alan Lopez. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, volume 1 of *Global Burden of Disease and Injury Series*. Harvard School of Public Health, 1996.

[5] Chris Neuman. Motor vehicle collisions 2009. Technical report, The City of Edmontons Office of Traffic Safety, Edmonton, AB, Canada, 2010.

[6] Paul De Leur, Laura Thue, and Brian Ladd. Collision cost study final report. Technical report, Capital Region Intersection Safety Partnership, 2010.

[7] Tessa Anderson. Comparison of spatial methods for measuring road accident 'hotspots': a case study of london. *Journal of Maps*, 3(1):55 – 63, 2007.

[8] Clive Sabel, Phil Bartie, Simon Kingham, and Alan Nicholson. Kernel density estimation as a spatial-temporal data mining tool: exploring road traffic accident trends. Geographical Information Science Research UK 2006, pages 191 – 196, University of Nottingham, Nottingham, UK, April 2006.

[9] Keith Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286 – 306, 1995.

[10] Isabelle Thomas. Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention*, 28(2):251 – 264, 1996.

[11] Wade Cook, Alex Kazakov, and B. N. Persaud. Prioritising highway accident sites: a data envelopment analysis model. *The Journal of the Operational Research Society*, 52(3):303 – 309, March 2001.

[12] Tessa Anderson. Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3):359 – 364, 2009.

[13] Sarath Joshua and Nicholas Garber. Estimating truck accident rate and involvements using linear and poisson regression models. *Transportation Planning and Technology*, 15(1):41–58, 1990.

[14] Paul Jovanis and Hsin-Li Chang. Modeling the relationship of accidents to miles traveled. *Transportation Research Record*, 1068:42–51, 1986.

[15] Michael Maher and Ian Summersgill. A comprehensive methodology for the fitting of predictive accident models. *Accident Analysis & Prevention*, 28(3):281 – 296, 1996.

[16] Mohamed Abdel-Aty and A.Essam Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633 – 642, 2000.

[17] Shaw-Pin Miaou. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4):471 – 482, 1994.

[18] Benot Flahaut, Michel Mouchart, Ernesto San Martin, and Isabelle Thomas. The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach. *Accident Analysis & Prevention*, 35(6):991 – 1004, 2003.

[19] Benot Flahaut. Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation. *Accident Analysis & Prevention*, 36(6):1055 – 1066, 2004.

[20] Simon Washington, Matthew Karlaftis, and Fred Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press, 2011.

[21] Jan Salomon Cramer. *Econometric Applications of Maximum Likelihood Methods*. Cambridge University Press, 1989.

[22] Bernard Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1986.

[23] David O'Sullivan and David Unwin. *Geographic Information Analysis*. Science/Geography. Wiley, 2003.

[24] Michael Goodchild. *Spatial autocorrelation*. CATMOG Series. Geo Books, 1986.

[25] Waldo Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:pp. 234–240, 1970.

[26] Sudeshna Mitra. Enhancing road traffic safety: A gis based methodology to identify potential areas of improvement. *Leonard Transportation Center, California State University, San Bernardino*, 2008.

[27] Luc Anselin. Local indicators of spatial associationlisa. *Geographical Analysis*, 27(2):93–115, 1995.

[28] Arthur Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189 – 206, 1992.

[29] J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4):286 – 306, 1995.

[30] Michael Ernst. Permutation methods: a basis for exact inference. *Statistical Science*, 19(4):676 – 685, 2004.

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

[32] Erich Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. HoldenDay, 1975.