Modifying a Local Measure of Spatial Association to Account for Non-Stationary Spatial Processes

by

Ian Kenneth Mackenzie
Bachelor of Arts, Anthropology
University of Victoria 2003

A Thesis Submitted in Partial fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Geography

Supervisory Committee:

_____
Dr. Trisalyn Nelson, Supervisor (Department of Geography)

_____
Dr. Barry Boots, Departmental Member (Department of Geography)

_____
Dr. Michael Wulder, Departmental Member (Department of Geography)

_____
Dr. Hannah Wilson, Outside Member (Geography Department, Malapsina University-College)

Supervisor: Dr. Trisalyn A. Nelson

Abstract

With an increasing number of large area data sets, many study areas exhibit spatial non-stationarity or spatial variation in mean and variance of observed phenomena. This poses issues for a number of spatial analysis methods which assume data are stationary. The Getis and Ord's G_i^* statistic is a popular measure that, like many others, is impacted by non stationarity. The G_i^* is used for locating hot and cold spots in marked data through the detection of spatial autocorrelation in values that are extreme relative to the global mean value, or the mean entire study area. This thesis describes modifications of the Getis and Ord's G_i^* local measure of spatial association, in part to account for regional differences (spatial non-stationarity) in a dataset. Instead of using data from the entire study area to calculate the mean parameter, as is done for the standard G_i^*, I capture points for calculation of the mean using a circular distance band centred on the pivot location, which I call the local region (similar to the Ord and Getis O_i statistic). This approach can be applied to a single instance of a local region or to multiple spatial scales of the local region. I explore both in this paper using simulated datasets and a case study on mountain pine beetle infestation data. I find that the local region, when of a similar size to a true region (homogeneous section of the study area where the mean is approximately the same across locations), obtains similar results to the standard G_i^* calculated separately on distinct regions (simulated to be distinct), but has the advantage of not needing explicit delineation of regional boundaries or partitioning into separate subareas. The results of a probability score for a multi-scale approach include high and low scores that are more evenly distributed across the study area and that are thus able to pick out more subtle variations within different regions. Through the case study I demonstrate how the multi-scale approach may be applied to a real dataset.

Examiners:

_____

Dr. Trisalyn Nelson, Supervisor (Department of Geography)

_____

Dr. Barry Boots, Departmental Member (Department of Geography)

_____

Dr. Michael Wulder, Departmental Member (Department of Geography)

_____

Dr. Hannah Wilson, Outside Member (Geography Department, Malapsina University-College)

Table of Contents

List of Tables

Table

List of Figures

Figure

Acknowledgements

Thank you to Dr. Trisalyn Nelson for her continual assistance on all things spatial and her editorial prowess; Dr. Barry Boots and Dr. Michael Wulder for their feedback on several drafts of my thesis manuscript; and lastly, thank you to friends, family, and colleagues in the SPAR laboratory for their support.

Chapter 1 – Introduction

At a landscape scale, the spatial pattern of a phenomenon can differ markedly from one area to another. I refer to places where the spatial pattern of a phenomenon is homogeneous as a region and recognize that large study areas typically include several regions. Within regions, global parameters (e.g., mean, standard deviation) of the spatial process are homogeneous and between regions they differ (Sokal et al. 1993). In spatial analysis, the variation across a study area is the result of first and second order effects. The influence of underlying environmental conditions on a phenomenon is a first order effect and interaction between cases of a phenomenon is a second order effect (O'Sullivan and Unwin 2003, pp 65-66). An example of a first order effect would be the influence of soil-type on tree growth, and a second order effect the influence of interspecies competition on tree height.

The assumption of stationarity is necessary for the accurate use of many spatial statistics. Stationarity refers to a process, or model of a process, having properties independent of absolute location and direction in space, where the parameters of the process or model, including variance and mean, are similar in all sections of the study area and in all directions (Burrough 1987, Haining 1990, Fortin & Dale 2005, pp 11-13). The assessment of stationarity is scale dependent. A spatial process may generate a stationary pattern at one spatial scale, while at a different spatial scale the pattern is heterogeneous.

In spatial statistics, global measures are used to describe the general spatial pattern of a study area and output a single summary measure. Local measures describe the spatial

pattern within a neighbourhood of each phenomenon and are used to identify locations in the study area where cases of the phenomenon are unusual. For this reason, local measures are preferred to global measures for characterizing spatial pattern when there is spatial heterogeneity. However, like global measures, some local analyses are vulnerable to non-stationarity because they require global parameters, such as the mean and standard deviation, to calculate raw scores and to standardize results to *z*-scores. Standardization is important if comparisons are to be made within and between datasets and for determining statistical significance.

An example of a local measure that is calculated and standardized by global parameters is the Getis-Ord G_i^* (Getis and Ord 1992; Ord and Getis 1995). The G_i^* quantifies local positive spatial association in values that are extreme relative to the mean. In the case of the G_i^*, the sum of all values is the divisor in the calculation for raw scores, and the global standard deviation and mean are part of the calculation to standardize the local analysis to *z*-scores. Thus, regional differences in the global parameters (i.e., non-stationarity) across a study area create biases in the results for the raw G_i^* scores and the *z*-scores.

A solution recommended for dealing with non-stationarity in a spatial process is to partition the study area into relatively homogeneous areas with consistent variance, mean and isotropy (isotropy refers to stationarity in the directionality of a spatial process) (Davis et al. 2000, Pélissier and Goreaud 2001, Fortin and Dale 2005, Wagner and Fortin 2005). Fortin and Dale (2005) identify two main approaches to spatial partitioning: 1)

spatial clustering or grouping of adjacent locations that have similar values of the variable under study by generating spatial clusters (e.g., agglomerative clustering - dendrogram, spatial contiguity constraints - Delaunay links, spatial clusters - *k*-means partitioning (Legendre and Fortin 1989), and 2) boundary delineation or dividing areas based on their degree of dissimilarity by delineating boundaries (e.g., lattice wombling and wavelets). In theory the outcome should be the same but in practice there can be differences between the two methods (Fortin & Dale 2005). The main issues with partitions that I am concerned with for this study are partitions: 1) are often difficult to determine and delineate, 2) reduce the sample size, and 3) no longer consider the study area as a whole but as distinct separate pieces (increased edge effects).

The goal of this thesis is to outline a novel approach for dealing with non-stationarity by using a modification of the G_i^* local measure. The approach uses a moving region, centered on each location, to calculate local "global" parameters for the G_i^* and does not require partitioning of large study areas. Throughout this thesis I demonstrate a flexible approach for evaluating the local spatial pattern of association for a non-stationary spatial process that does not require partitions and I outline an extension of this approach that is suitable for multi-scale analysis.

This thesis is organized into six chapters including this introductory chapter (Chapter 1). The main results of this thesis are in Chapters 4 and 5. In Chapter 4, I evaluate my modifications of the G_i^* on simulated datasets in order to account for non-stationarity and in Chapter 5, I evaluate these same modifications on a case study for point data of

mountain pine beetle infested trees in north-central British Columbia. I describe the modifications to the G_i^* statistic in detail in Chapter 3. I begin, however, by providing background on concepts that guide my research (Chapter 2).

CHAPTER 2 – BACKGROUND

Characterizing spatial association is a fundamental concern of spatial data analysis (Boots 2002). Spatial association refers to the relatedness of a set of spatial data and the extent to which nearby data are similar or different (Griffith 1992, Cliff and Ord 1973). Spatial association is sometimes referred to as spatial autocorrelation. I use the more general term of association because spatial dependence can be the result of *i*) true spatial autocorrelation within the phenomenon of interest (i.e., "self-correlation") or *ii*) induced spatial dependence by an underlying environmental condition, or iii) both (Legendre et al 2002). Positive spatial association refers to similarity in nearby data values and negative spatial association refers to neighbourhoods of dissimilar values. A spatial pattern has no spatial association when neighbouring values are neither unusually similar nor unusually different. Measures of spatial association are applied to datasets with marks (attributes other than the spatial coordinates). For instance, for a point representing the location of a tree, a mark could be the tree height.

A number of statistics exist for measuring the degree of spatial association in spatial data. These include global and local measures. Global measures summarize spatial association for an entire area with a single value. Examples of global measures are Moran's I and Geary's c (Fotheringham 1996). For global measures to be used accurately the data must be stationary. The assumption of stationarity is often invalid for datasets with large spatial extents where it is likely that one or more regions will have different properties than the others (Boots 2002). Local measures are useful for measuring spatial association

over large landscapes, and are commonly used to characterize the spatial association of objects that are heterogeneous in spatial distribution and/or attribute value. Local measures are used to calculate spatial dependence values for each location based on its surrounding local neighbourhood. A local neighbourhood can be defined in a number of ways including, but not limited to, *k*-order neighbours (Lee and Drysdale 1981), Delaunay neighbours (Okabe et al. 1992), or metrical distance methods (Hernandez et. al. 1995; Huang and Sevonson 1993). Local measures can be applied to a contiguous raster dataset (Wulder and Boots 1998) or to an irregularly spaced point or area dataset (Getis and Ord 1991, Ord and Getis 2001, Nelson et al. 2005).

There are three popular local measures for detecting spatial association in data having interval or ratio attribute values (i.e., non-categorical). They are local Moran's I_i and local Geary's c_i (Anselin 1995), which are modifications of the global measures Moran's I and Geary's c, and the Getis and Ord G_i local measures (Getis and Ord 1992, Ord and Getis 1995) of which there are two; the first includes the pivot location *i* in the calculation (G_i^*), and the second does not (G_i). The pivot location is the location at the centre of the local neighbourhood for which the local statistic is being calculated. Local Geary's c_i detects positive and negative spatial association and Local Moran's I_i allows identification of positive and negative spatial association in values that are extreme relative to the mean. The Getis and Ord G_i^* statistic identifies only positive spatial association and is unique in its ability to distinguish positive spatial association in extreme high values from positive spatial association in extreme low values. For the G_i^* statistic, a high positive G_i^* *z*-score indicates a spatial grouping of high attribute values,

and a high negative G_i^* *z*-score identifies a spatial grouping of low attribute values (Ord and Getis 1995).

In this study, I focus on modifying the G_i^* statistic. This statistic is particularly valuable for identifying clusters of extreme high and low values. As such, it has applicability to a broad spectrum of research fields including epidemiology (Getis and Ord 1992, Burra 2002), criminology (Eck et al. 2005), and ecology (Fortin and Dale 2005). The utility of this measure has lead to increased use in the last several years and it has become an important method for developing and supporting hypotheses about the spatial patterns of certain phenomena. Addressing problems associated with this measure is therefore critical given its increased popularity and widespread use in several fields of research. In its basic form, the G_i^* statistic is the sum of attribute values in a neighbourhood divided by the sum of all attribute values in the entire study area. The equation is written as follows (Getis & Ord 1992):

$$G_i^* = \sum_j w_{ij}(d)x_j / \sum_j x_j \qquad j \text{ may equal } i \qquad (1)$$

In the context of this study, w_{ij} is a binary spatial weights matrix captured for each location using a circular distance band for the local neighbourhood, *d* is the size of the local neighbourhood (radius of the distance band), and x_j is the attribute value at the j^{th} location. For G_i^* *j* may equal *i* (as indicated by the asterix). In the spatial weight matrix

w_{ij}, a weight of "1" is assigned to all points within distance *d* of the pivot-location *i* and "0" to all points beyond that distance.

Typically the results of the G_i^* are reported as *z*-scores and the focus is often on those *z*-scores greater than two or less than negative two (sometimes referred to as "hotspots" and "coldspots"). The use of z-scores allows comparisons between different datasets. The equation for standardizing the results to *z*-scores is (Ord & Getis 1995):

$$Z(G_i^*) = \frac{\sum_j w_{ij}(d)x_j - W_i^*\bar{x}}{s\left\{\left[\left(n\sum_j w_{ij}^2\right) - W_i^{*2}\right] \Big/ (n-1)\right\}^{1/2}} \qquad \text{all } i \tag{2}$$

Where *s* = the standard deviation for the full dataset, $W_i^*\bar{x}$ is the mean of all attribute values in the dataset, W_i^{*2} is the square of the count of total values in the dataset (*n*), and $n\sum_j w_{ij}^2$ is the square of the count of the values in the local neighbourhood. The expected value of G_i^* is the sum of all values divided by *n*. The equation is written as follows (Ord & Getis 1995):

$$E[G_i^*] = \sum_j w_{ij}(d)/n \tag{3}$$

All local measures require a local neighbourhood to be defined. Getis and Ord (1996) suggest that a maximum size for the local neighbourhood should never exceed half the shorter side of a study area and that the number of neighbours should be at least 30 for large samples and 8 for small samples. One suggestion for defining a neighbourhood is that the statistic G_i^* be evaluated at a series of increasingly larger neighbourhoods until no further spatial autocorrelation is evident (Getis and Griffith 2002). Another suggestion has been to create a semivariogram for each variable and use the distance of the range for *d* (Getis and Griffith 2002). Laffan (2002) recognizes the difficulty in choosing a neighbourhood size and investigates an adaptive neighbourhood based on a process model that can change for different locations. I recognize the difficulties inherent in selecting a neighbourhood; however, this thesis aims to adjust a different part of the G_i^*, a local moving region to calculate global parameters; and so to make comparison easier, the size of the local neighbourhood remains the same for all simulated datasets. For the simulated datasets, a circular distance-band with a radius of 10 metres is used for the local neighbourhood. This distance band captures between 8 and 30 neighbours for each location in all simulated datasets. To note, there are many other non-distance based neighbourhood definitions, for example Voronoi polygons (Okabe et al. 1992), but to remain consistent with the earlier work on G_i^* development (Getis and Ord 1992, Getis and Ord 1996), I restrict the analysis to the use of a distance-band based neighbourhood definition. However, the methods developed in this thesis could easily be modified to include other neighbourhood definitions.

The emergence of Exploratory Data Analysis (Tukey 1977) as an alternative to the classical significance-based approach to statistics has more recently been extended to spatial analysis (thus Exploratory Spatial Data Analysis - ESDA), particularly in combination with the growth of Geographic Information Systems (Haining et al 1998). Exploratory Spatial Data Analysis uses spatial statistics as descriptive methods for detecting patterns, formulating hypotheses, and for pre-testing spatial data to evaluate what standard statistical test will work best (Unwin and Unwin 1998). The methods evaluated in this paper are considered to be in the nature of Exploratory Spatial Data Analysis. Treating the methods as such allows us to temporarily put aside the important matter of evaluating statistical significance, which presents a common problem for all local measures due to multiple testing and a lack of independence in the data (Getis and Ord 1992, Ord and Getis 1995, Anselin 1995).

CHAPTER 3 – DEVELOPMENT OF NEW METHODS

The LR G_i^* (local region G_i^*) is my modification of the original G_i^* statistic (Getis and Ord 1992; Ord and Getis 1995). For the LR G_i^*, as for the standard G_i^*, I use a local neighbourhood with a binary spatial weights matrix and inclusion (weight = 1) based on a circular distance band centred on location x_i. Instead of using all of the data from the entire study region to calculate global parameters, as is done for the standard G_i^*, I capture points for calculation of global parameters using a circular distance band centred on the pivot location, like for the local neighbourhood, but with a radius greater than that of the local neighbourhood. I can think of no immediate reason why any other neighbourhood definition and spatial weights matrix could not be used for both the local neighbourhood and the local region. I rewrite equation 1 to incorporate the local region as follows:

$$G_i^* = \sum_j w_{ij}(d_1)x_j / \sum_j w_{ij}(d_2)x_j \qquad j \text{ may equal } i \tag{4}$$

Where d_1 = the radius of the local neighbourhood; d_2 = the radius of the local region

The standardized form requires a few changes to definitions for the global parameters (mean, standard deviation, and *n* count)

$$Z(G_i^*) = \frac{\sum_j w_{ij}(d_1)x_j - W_i^*\bar{x}}{s\left\{\left[\left(n\sum_j w_{ij}^2\right) - W_i^{*2}\right] \Big/ (n-1)\right\}^{1/2}} \qquad \text{all } j \tag{5}$$

Where, $W_i^* \; n = \sum_j w_{ij}(d_2)$,

$\bar{x} = \sum_j w_{ij}(d_2)x_j / \sum_j w_{ij}(d_2)$, and

$s = \sqrt{\sum_j \left(w_{ij}(d_2)x_j - W_i^* \bar{x}\right)^2}$

The LR G_i^* has similarities with the O_i statistic introduced by Ord and Getis (2001) as a way to detect local spatial association in the presence of global spatial association. The O_i has two parts to its procedure:

1) estimate the global association from the set of all N observations.
2) Analysis based on a region M of the region contained in N. Partition the study area into 'relatively homogeneous' regions.

A criticism of O_i (included in Ord and Getis 2001) is that rejection of the null would point to a cluster of high values. However, rejection of the null also invalidates the variogram or correlogram, used to construct the O_i, which is based on the assumption of spatial stationarity. Ord and Getis' recognize this limitation and suggest that a way to get around this problem is to define the sets M and N and then to compute the association estimates using only the locations in N-M. The calculation needs to be repeated for each location. They see the computational costs of doing this as potentially prohibitive. Computing the association estimates at each location is in fact what I am doing with the LR G_i^* method. Ord and Getis (2001) are also concerned about a masking problem in that a cluster of high values in N-M may bias the estimates of an overlapping neighbouring cluster. They see as a solution a multi-stage approach but, in their own

words, leave this as a challenge for future research. I attempt to address this issue with an extension to the LR G_i^* termed the multi-region G_i^* (MR G_i^*), which involves calculating a LR G_i^* at multiple scales of the local region.

Figure 1 provides a schematic of how the MR G_i^* works. For the MR G_i^* the local neighbourhood distance band is static while the local region (and thus each global parameter) is dynamic. For both the simulated datasets and the case study, the method was set so that the first local region for each location was selected as twice that of the local neighbourhood distance band, the rationale being that if the local neighbourhood is meant to represent the range of influence of a location (beyond which spatial autocorrelation is zero), then doubling its size should safely exceed that range and provide an appropriate initial comparison between the local neighbourhood and its surroundings.

The local region is increased incrementally by a distance equivalent to the radius of the local neighbourhood. For each increment the G_i^* is calculated. The local region is expanded repeatedly by the size of the increment and the statistic calculated until all points in the study area have been captured. The number of iterations of the local region for each point will vary depending on a point's location in the study area. Points located towards the centre of the study area will take less iteration to capture all points than points located towards the edges. The G_i^* result at a single record for the largest local

region (the final increment) of the MR G_i^* is the same as that calculated by the standard, unmodified G_i^* since the largest local region (the final increment) encompasses all data in the dataset.

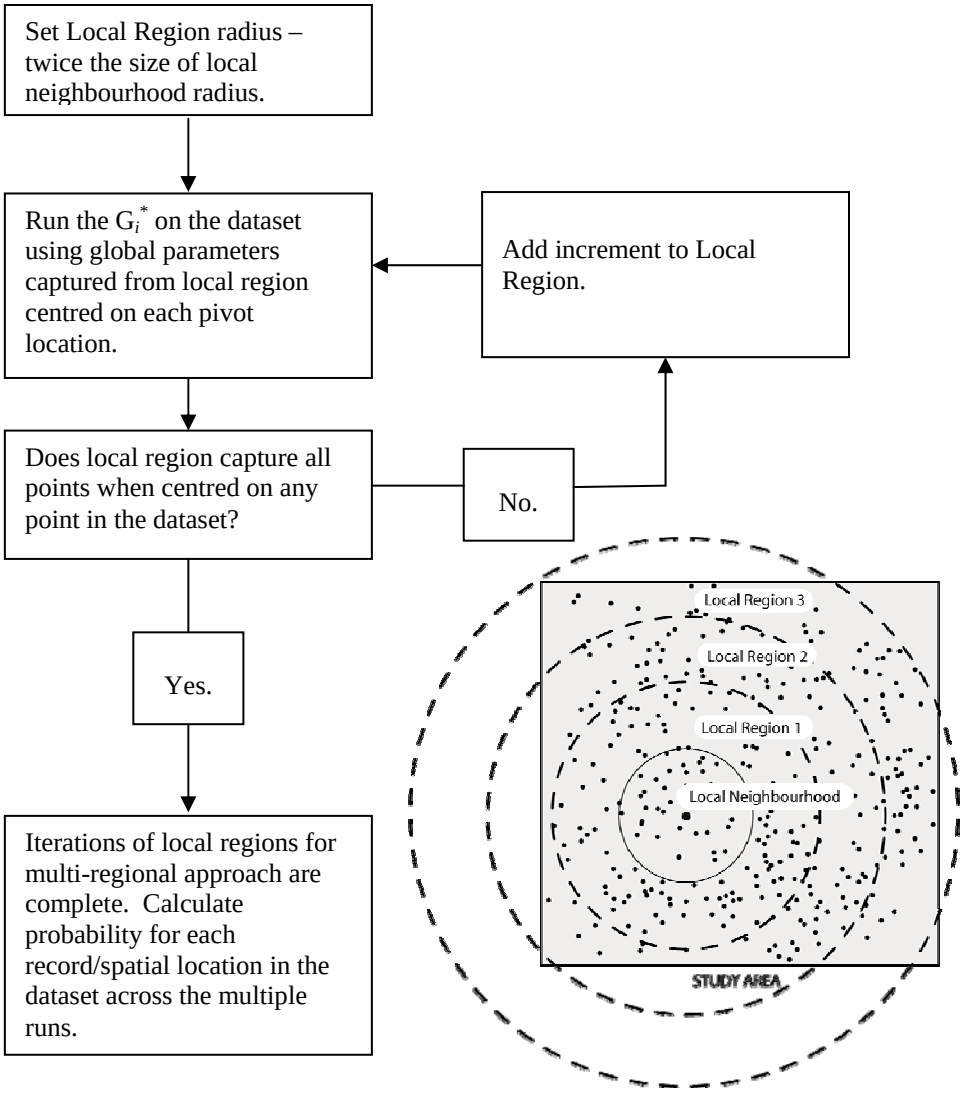


Figure 1 A diagrammatic overview of the MR G_i^* method.

The output of the MR G_i^* is a set of G_i^* *z*-scores, one for each iteration of the local region. I calculated two probability scores one based on the number of times a location receives a *z*-score >=2 and the other on *z*-score <=-2, with each divided by the total number of increments calculated for that location. Unless otherwise stated, a high probability MR G_i^* score refers to a location receiving a MR G_i^* *z*-score >= 2 for fifty percent or greater of the increments of the local region and a low probability MR G_i^* score refers to a location receiving a MR G_i^* *z*-score <=-2 for fifty percent or greater of the increments.

Summary statistics (median and mean) of the MR G_i^* were also calculated across all increments of the MR G_i^* for each location (number of increments will vary between points depending on their location in the study area). However, the mean and even more so the median of the MR G_i^* receive near identical results to the MR Probability scores and because there is some ambiguity to exactly what the mean or median of the MR G_i^* represents, I prefer to use the MR Probability scores for comparisons with other methods (i.e., Standard G_i^* and LR G_i^*). Nevertheless, because MR Median is so similar to MR Probability, it is a useful proxy for the MR Probability where a difficult transformation of probability to *z*-scores would otherwise be required. I use MR Median in place of MR Probability for graphs where I compare the MR G_i^* results to the Standard and LR G_i^* *z*-scores.

Increments for the MR G_i^* can alternately be increased such that they are a linear increase in area rather than a linear increase in distance. There may be theoretical reasons why a researcher might prefer to use an area-based increment. Although theoretically there may be an advantage to using either a distance- or an area-based increment, the trials using both showed no appreciable differences between the results of the two when compared

visually and with a KS test for two samples (p = .05). There may be some advantage to the logarithmic scaling for the linear-distance increments because it gives more weight to the values that are within increments closer to the pivot location.

For the MR G_i^*, it is necessary to standardize results (using Equation 4) in order to compare across a set of local region iterations. Otherwise the result for each iteration of the local region would simply decrease as the distance band increases. Any summary measure of the MR G_i^* would then be skewed to results calculated at the smaller local regions. This is not the case for the LR G_i^* where it would be appropriate to use either equation 2 or 4. I chose to use the standardized method (equation 4) to make it easier to compare the results of the LR G_i^* to the MR G_i^* and standard G_i^* methods, and also because G_i^* results are typically presented as *z*-scores and readers familiar with the G_i^* statistic may be more comfortable interpreting them in this form. Also, *z*-scores provide an indicator of statistical significance, albeit problematic.

The size of the increment used to increase the distance band of the local region for the MR G_i^* needs to be small enough to encompass the range of spatial association inherent to the process(es) of the phenomenon. It makes some sense, at least for the sake of consistency, to use the same or a similar distance to the radius of local neighbourhood since it would likely be chosen to investigate spatial association in the data. The smaller the increments the more precisely the method will capture changes in the results as the distance band of the local region increases. However, if I choose increments that are larger than the range of spatial association this will likely manifest in the changes in *z*-scores across the results of the increments. Generally speaking, so long as a sufficient number of distance bands is achieved with the chosen increment size, then the changes across those distance bands should average out similarly to a range of different increment sizes. A conservative approach would be to use an increment that is less than d_1 to provide more analytical detail at the small cost of lost computational speed.

CHAPTER 4 - SIMULATIONS

INTRODUCTION

In order to assess the approaches in a controlled environment, I applied the LR and MR Gi* methods to simulated data. Simulated data were generated to represent a variety of spatial processes including stationary processes, clustered processes, and processes impacted by large scale trends. This chapter is divided into two main parts: A and B. In Part A, I evaluate the LR Gi* and MR Gi* . In Part B, I demonstrate potential methods for determining an appropriate local neighbourhood size and local region size, and for detecting transitional areas in the study area. First, however, I introduce and describe the simulated datasets.

DATA

This study uses six simulated datasets, which can be categorized into two groups. In the first, the points of the dataset are located randomly and uniformly in the study area and in the second the points are clustered across the study area, although the location of each cluster parent is still random. For each group, clustered and non-clustered, there are three representative datasets: the first representing a stationary spatial process, the second has four distinct regions where the global parameters are different for each region, and the third has a north-south gradient in the attribute values of points, which I refer to as a global trend. All simulation points are contained within a 100 by 100 metre square area. Both point locations and attribute values are simulated.

Figure 2 shows the maps of attributes for all six simulated datasets and Table 1 provides a summary of the simulated datasets. All non-clustered datasets comprise 1000 points each

Clustered datasets comprise 400 points each. I determined that it was necessary for comparative purposes to keep clusters relatively distinct from one another while maintaining a similar density within the clusters as exists for the non-clustered datasets (over the entirety of those datasets). The only way to accomplish this was to use fewer points. I found 400 points to be appropriate in this respect. Quantitative comparisons between datasets were conducted within groups only (clustered or non-clustered) and never between datasets of different groups.

Table 1 – Simulations – description of simulated datasets.

Dataset Name	**Short Name**	**Clustered**	**Spatial Process**
D.1.1	Uniform stationary	No	Stationary
D.1.2	Uniform regional nonstationary	No	Non-stationary – four regions
D.1.3	Uniform global trend	No	Global trend
D.2.1	Clustered stationary	Yes	Stationary
D.2.2	Clustered regional nonstationary	Yes	Non-stationary – four regions
D.2.3	Clustered global trend	Yes	Global trend

Clustered datasets are simulated by way of a compound Poisson process (Diggle 1983). Forty parent locations are assigned random locations in the study area, with coordinate values drawn from a uniform distribution. Each parent location is assigned 10 children and the children are assigned spatial coordinates from a normal distribution with a standard deviation of one and a mean equivalent to the x- and y coordinates of the parent. The number of children could also have been made random, but to maintain some level of control over the spatial pattern I opted to use a fixed number of children. Marks for the dataset are drawn differently for each of the clustered datasets, details below. Only the

children points, not the parents, are included in the final dataset. The final datasets have 400 points.

Dataset 1.1 (D.1.1) – stationary process and no clusters – represents a random spatial pattern, with coordinates drawn from a uniform distribution with limits consistent with the boundaries of the study area, and random marks drawn from a normal distribution with a mean of 100 and a standard deviation of 1.

Dataset 1.2 (D.1.2) – regional non-stationary process and no clusters – represents complete randomness in the spatial location (x and y coordinates drawn from a uniform distribution) and four distinct regions for the marks. The marks for each region are drawn from a normal distribution with a unique mean for each region and a standard deviation of approximately 2. The four unique means were drawn randomly from a normal distribution with a mean of 100 and a standard deviation of 2. The mean of the top-left quadrant mean is 104.15 and the standard deviation is 1.96; the mean of the top-right quadrant is 109.85 and the standard deviation is 1.92; the mean of the bottom-left quadrant mean = 97.97 standard deviation = 1.88; bottom right quadrant = 100.15 and standard deviation = 2.06. All regions are contained within a 50 by 50 metre quadrant of the study area.

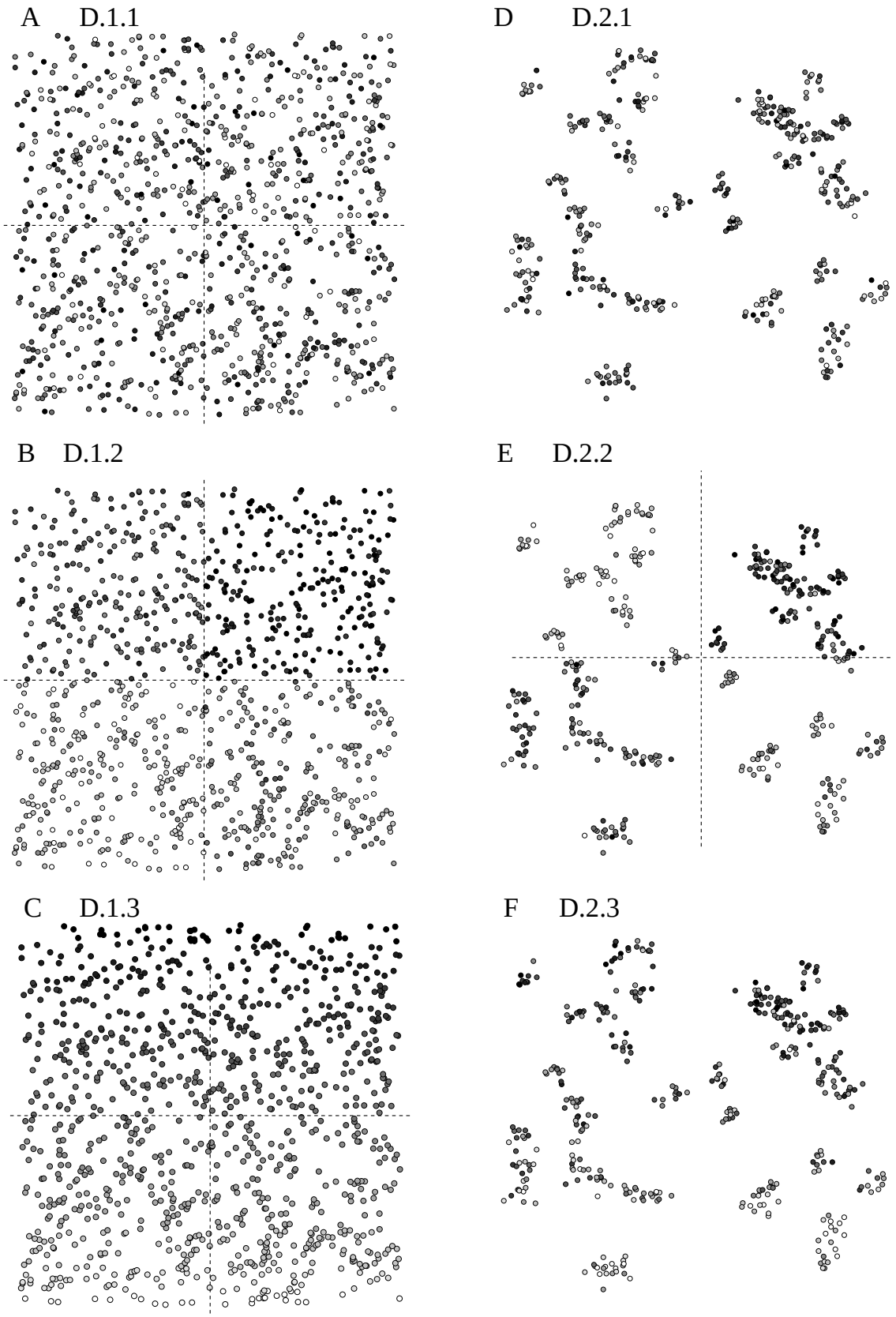


Figure 2 Simulations - Maps of the marks of simulated datasets. **A)** D.1.1, uniform stationary **B)** D.1.2, uniform regional non-stationary **C)** D.1.3, uniform global trend **D)** D.2.1, clustered stationary **E)** D.2.2 clustered regional non-stationary **F)**, D.2.3 clustered global trend.

Dataset 1.3 (D.1.3) – global trend and no clusters – represents complete randomness in the spatial location but a global trend in the attribute values. The exact same marks are used as in D.1.1 but they are assigned to points such that values increase with increased latitude, creating a gradient of increasing mark values.

Dataset 2.1 (D.2.1) – stationary process and clusters – represents a clustered dataset where marks are drawn from a normal distribution with a mean of 100 and a standard deviation of 1. Coordinates are created from a compound Poisson process (describe above) and marks are assigned randomly to points, regardless of point or cluster location.

Dataset 2.2 (D.2.2) – regional non-stationary process and clusters – represents a clustered dataset where there are four distinct regions where marks for a set of adjacent clusters in a quadrant have the same mean. Where a cluster straddles a quadrant boundary the mean marks for that cluster are assigned based on which quadrant the parent location of that cluster is located. Consequently, the regions are not perfectly contained within each quadrant and will exceed a quadrant boundary in some cases. The means for the quadrant were drawn from a normal distribution with a mean of 100 and a standard deviation of 2. The mean of the top-left quadrant mean is 99.99 and the standard deviation is 0.91; the mean of the top-right quadrant mean is 103.08 and the standard deviation is 0.96; the bottom-left quadrant mean is 101.91 and the standard deviation is 0.98; the bottom right quadrant mean is 101.09 and the standard deviation is 0.86.

Dataset 2.3 (D.2.3) – global trend and clusters - represents a clustered dataset where each cluster has a different mean value and those means are distributed across the clusters in a trend running south to north. Forty sets of 10 marks, one set for each cluster, were created as follows: The initial parent values were generated from a normal distribution with a mean of 100 and standard deviation 1. These parent values were then used as the mean value along with a standard deviation of 1 to derive values for each set of marks. Mark sets were then assigned to locations (the same location as for the other two clustered datasets), such that the cluster attribute values with the lowest mean go to the cluster of points with the lowest latitude (based on the location of the original parent location, not included in the final dataset), the cluster with the next lowest mean goes to the cluster of points with the second lowest latitude, and so on, to create a gradient of increasing mean values for the marks of clusters. Within a cluster, the assignment of attribute values is random, that is, it is not determined by location. The mean for the entire dataset is 100 and the standard deviation is 2.71.

For choice of standard deviation, two hundred simulations of the standard Gi* statistic were run on each of five datasets (different than those described above but of the same number of points), all with the same mean of 100, but with different standard deviations, respectively, 1, 3, 10, 100, and 120. D'Agostino-Pearson's and Shapiro-Wilk's normality tests (D'Agostino 1971, D'Agostino et. al 1990; Shapiro and Wilk 1965) found all results sets to fit a Gaussian distribution, percentages of high (>=2) and low (<=-2) Gi* z-scores were of the expected values of ~2.5% in each tail of the distribution, and no important differences were observed between the different results sets when observed in boxplots.

A standard deviation of one was selected for subsequent simulations to maintain distinctiveness of the simulated regions.

The simulated datasets contain only one sample of each type of spatial process, thus effectively counting for a sample size of one. Ideally, multiple simulations of each spatial process would have been conducted to ensure that the results observed were not simply an aberration of that single simulation. Although ideal, to conduct such a simulation would have required large amounts of computing time. An ad hoc sensitivity analysis was conducted by the author, in tandem with the main analysis, to observe what effect randomly changing the parameters of the simulations would do to the results. Changing the parameters had no unexpected effects on the performance of the various Gi* methods, that is, the various methods had the same general effect on the newly simulated datasets as far as concerns the central issues of this thesis (i.e., partitioning, non-stationarity, edge effects, etc.). Nevertheless, if exploring the methods presented in this thesis is a road worthy of further travel, then the author would suggest implementing a larger sample size, using a re-sampling process akin to Monte Carlo analysis.

PART A

In Part A, I evaluate the performance of LR G_i^* and MR G_i^* on the simulated datasets.

METHODS

For the simulated datasets, I first evaluate the performance of the LR G_i^* and second the MR G_i^*. Table 2 is a summary of the different analyses. For all G_i^* methods I use a local neighbourhood of 10 metres. A LR G_i^* is run on the dataset using a local region with a diameter of 50 metres; this is the same size as the smallest side of a regional partition for datasets 1.2 and 2.2. I refer to the results as LR1. For comparative purposes, the LR G_i^* is also run using a local region with a diameter twice the size of the diameter for LR1, thus a diameter of 100 metres. I refer to these results as LR2. An MR G_i^* is run on all datasets with the radius of the first increment of the local region set to 20 metres, twice that of the local neighbourhood radius of 10 metres.

Table 2 – Simulations - description of analysis methods.

Method	Results Short Name	Description
Standard G_i^*		
Standard	Standard	The standard G_i^* is calculated for the entire dataset.
Partition No Edge Correction	Partition	The standard G_i^* is calculated separately for each partition of the dataset. No edge correction solution.
Partition Plus Sampling	Partition$^+$	The standard G_i^* is calculated separately for each partition of the dataset. Plus sampling edge correction solution.
Partition Minus Sampling	Partition$^-$	The standard G_i^* is calculated separately for each partition of the dataset. Inset edge correction solution.
Local Region G_i^* (LR G_i^*)		
Local Region 1	LR1	The LR G_i^* is calculated for the entire dataset with the diameter of the local region equal to the smallest side of a true known simulated region.
Local Region 2	LR2	The LR G_i^* is calculated for the entire dataset with the diameter of twice the local region equal to the smallest side of a true known simulated region.
Multi-Region G_i^* (MR G_i^*)		
Multi-Region Median	MR Median	The median is used to summarize the multiple *z*-scores at each location for the various iterations of the MR G_i^*.
MR Probability *z*-score >=2/<=-2	MR Probability	The probability of a location having unusual high *z*-scores is calculated as the number of times a *z*-score >=2 divided by the total number of iterations of the MR G_i^* for each location. Similarly, for unusual low *z*-scores is calculated as the number of times a *z*-score<=-2 divided by the total number of iterations of the MR G_i^* for each location.

The results are assessed and evaluated using several techniques. The first is by mapping the G_i^* *z*-scores and visually assessing the general pattern of high and low *z*-scores (>=2;<=-2) for each method. The second, is by summarizing and comparing the counts of high and low *z*-scores for each method. The third technique is to calculate the number of coincident high and low *z*-scores between different methods. For instance, a coincident count would occur if at a single location both analyses calculated a G_i^* *z*-score greater or equal to 2 (or alternatively, <=-2). The fourth technique is to chart the percentiles zero to 100 of the G_i^* *z*-scores for each method to create a cumulative frequency graph and to visually compare the distribution of *z*-scores between methods (or more precisely, to report the *z*-scores at different percentiles directly from the spreadsheet used to create the graph). The final technique is to assess the average difference between methods over all classes of *z*-scores (not just high and low). Comparisons in this paper are made primarily in relation to the results of the standard G_i^* on the full dataset and the standard G_i^* on partitioned regions, and the probability and median *z*-scores for the multi-region methods.

EDGE EFFECT CONSIDERATIONS

One of my initial concerns with the local region for the LR G_i^* and the MR G_i^* is that edge effects may be having crucial impacts on the results and therefore the interpretations. While considerable effort has been put into developing edge correction techniques for point pattern analysis on non-marked point processes, for example, for Ripley's K-function (Ripley 1977, Ripley 1982, Haase 1995, Goreaud and Raphael 1999), little has occurred for the spatial analysis of marked data. None of the seminal

papers for the local measures Moran's I_i , Geary's $C_{i,}$ Getis and Ord's G_i and G_i^* (Getis and Ord 1991, Ord and Getis 1995, Anselin 1995) nor more recent papers on the subject provide an edge correction solution beyond minus sampling. Minus sampling is the edge correction technique where locations are removed from final analysis if they are within a distance of the boundary edge such that the consequent local result would be biased. For the methods explored in this study, minus sampling can be used for the local neighbourhood; however, using minus sampling for the local region would quickly become prohibitive as the only points that could be used for the analysis would be those in the centre of the study area beyond a distance from the boundary equivalent to the radius of the local region. To address the concern that edge effects might be having a serious impact on the results, several tests were conducted. The first found that there was no significant difference for a LR G_i^* between the results from the edges of the study area where the local region crosses the study boundary and the results from the centre of the study area, where the local region is completely within the study area, under the condition of complete spatial randomness, as tested with a KS test for two samples (ks = 0.083, p-value = 0.532).

The second test was to introduce an edge correction method into the LR G_i^* method. The edge correction used was along the lines of Ripley's correction method for the k-function (Ripley 1977, Goreaud and Raphael 1999) which weights parameters according to the proportion of the circular neighbourhood that is within the study area. For the LR and MR G_i^*, this method can be used to adjust the counts associated with each local neighbourhood as well as the local region. For the parameters describing the attribute

values for the local region (mean and standard deviation) the assumption is that they are the same in the adjacent area beyond the boundary where I have no data as in the area that falls within the study region. The sum for the local region is adjusted according to the proportion of region that falls inside of the boundary. The conclusion from applying this method to the same CSR dataset used for the test above, was that the edge correction resulted in G_i^* *z*-scores very similar to the non-edge corrected method, for those points where the local neighbourhood was completely inside the study area (but local region could be inside or across the boundary); any difference occurred at three decimal places or greater (thousandths), which is minor in consideration that these results are rarely reported beyond two decimal places. Where the local neighbourhood did cross the study area boundary, a more substantial difference occurred between the G_i^* *z*-scores for the edge corrected versus those of the non-corrected method. However, this edge effect introduces a consistent bias into all analyses that does not affect the results in terms of the broader generalizations that I am looking to make.

I concluded that, for this initial study, edge effects would not compromise results for the more general interpretations that I wished to make and demonstrate, that is, how the method changes when applied to distinctly different spatial processes. However, edge effect and correction methods will continue to be of interest in future research.

An important aspect of evaluating the LR G_i^* is being able to compare the results for datasets with true known regions (those datasets where regions were purposely simulated - D.1.2; D.2.2) to the results of a standard G_i^* calculated independently for each

partitioned region of those same datasets. Initially, three solutions were used for running the standard method independently on the partitioned regions for D.1.2 and D.2.2 (datasets with four distinct regions) to deal with edge effects for the local neighbourhood: plus-sampling (Partition^{+}), minus sampling (Partition^{-}), and no correction (Partition).

From the initial tests calculating the standard G_i^* independently for each region and using the different edge correction solutions, I found several issues regarding the results (Partition^{+}, Partition^{-}) that would make subsequent comparison to other methods (LR G_i^* and MR G_i^*) difficult. For D.1.2, I found upon visual assessment of the mapped *z*-scores that the plus-sampling technique (Partition^{+}) results in distinct "bands" (the width corresponding roughly with the diameter of the local neighbourhood) of high and low G_i^* *z*-scores (>=2;<=-2) along the boundaries dividing the regions (Figure 3). For D.2.2, this is not such a problem because the data are clustered and thus fewer points are near the boundaries between regions. When a point is near the boundary of two or more regions the local neighbourhood captures points from the adjacent region(s), through plus sampling. Because the transition between regions is sharp the local parameters are nearly always highly different than the global parameters that were calculated using data from within the region and consequently a high or low G_i^* *z*-score is nearly always calculated in these boundary areas. Figure 3 includes maps of Partition, Partition^{+}, LR1, and results for D.1.2 and D.2.2. I determined that the resulting "band" of high or low *z*-scores near the borders of regions was an artifact of how the statistic is calculated rather than a true evaluation of the underlying spatial process. Where there are more gradual transitions this would not be such a problem;

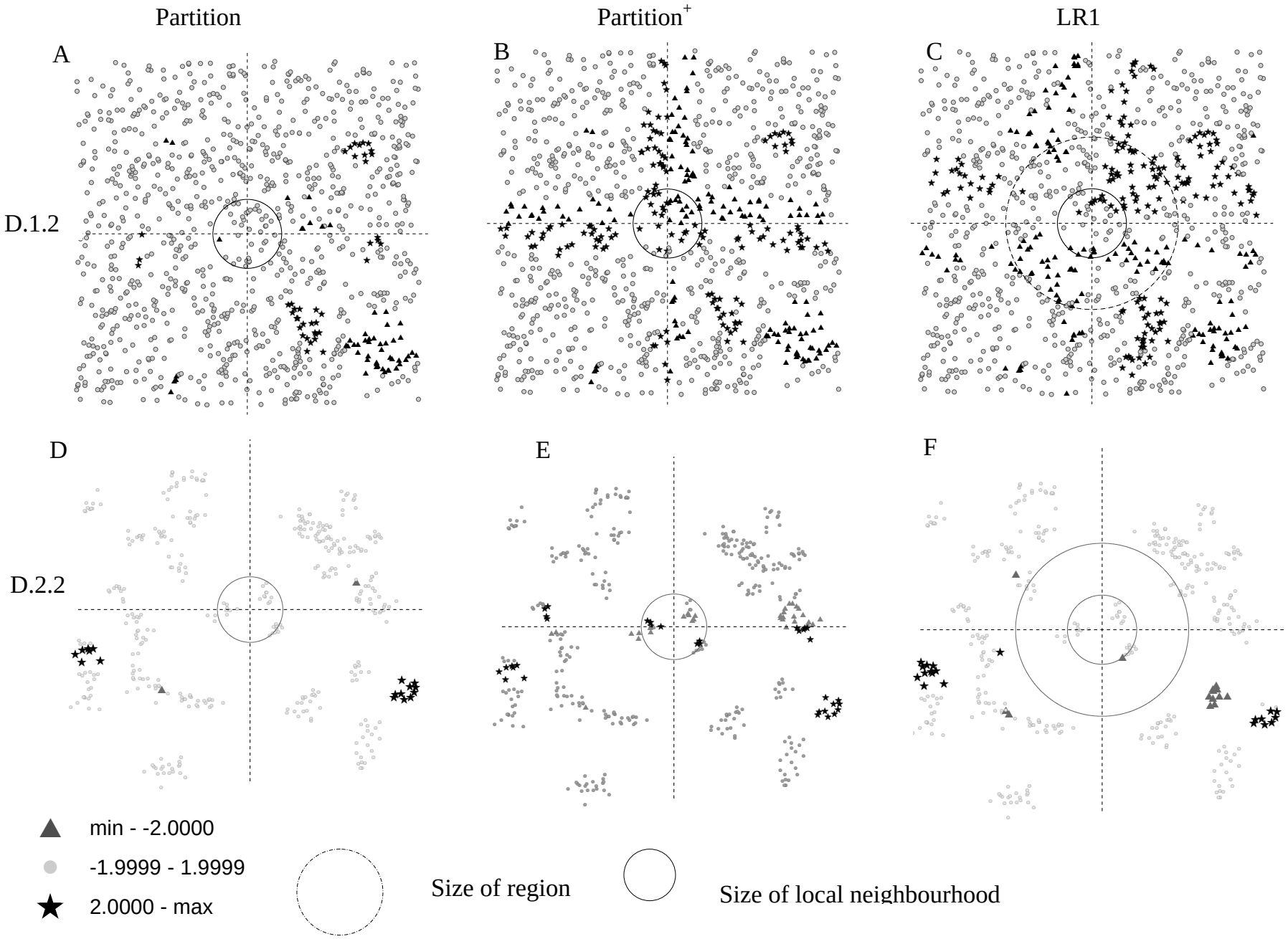


Figure 3 Simulations - D.1.2 and D.2.2 - maps of Partition, Partition^{+}, and LR1 results demonstrating the effects of regional boundaries. A. D.1.2 - Partition, B. D.1.2 - Partition^{+}, C. D.1.2 LR1, D. D.2.2 – Partition, E. D.2.2 – Partition $^{+}$, and F. D.2.2 – LR1.

however, because two of the datasets have sharp transitions, plus-sampling was determined to be inappropriate. Additionally, plus-sampling can only be done for the interior boundaries of the partitioned areas.

The other edge correction solution available to us was the minus sampling technique, however using such a method would reduce the total number of data available for evaluation, as it removes data that fall near region boundaries and as already described above using minus sampling for the LR G_i^* quickly becomes unfeasible. Furthermore, there is evidence that only a small difference between the minus sampling results (Partition$^-$) and the no-edge-correction results (Partition) occurs, as can been seen for D.1.2 in Figure 4, the cumulative frequency graph of *z*-score results for the different methods.

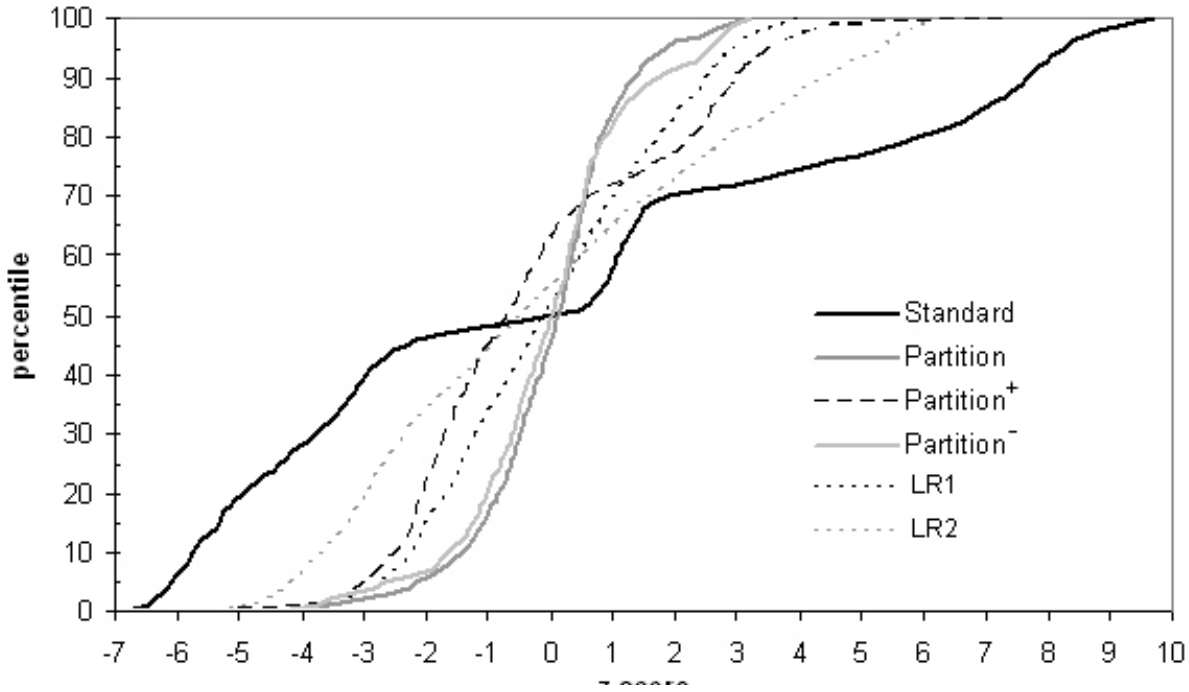


Figure 4 Simulations – D.1.2 – Cumulative frequency graph of results for various methods.

This leads to some initial results regarding the performance of the LR G_i^* methods compared to the standard G_i^* on partitioned regions. In Figure 4, it can be seen that Partition$^-$ and Partition results sets are very similar and for comparative purposes for this study it is probably unnecessary to be concerned with edge effects still present in the Partition method. The graph also illustrates how the LR G_i^* results LR1 lie between Partition and Partition$^+$ results, and that these two results sets (Partition and Partition$^+$) might be considered as "envelopes" of the two possible extremes of regional differences that a local measure such as the LR G_i^* needs to consider (with the exception of where the envelopes necessarily cross over each other near the 50th percentile. The LR2 result (where local region is twice the size of LR1) falls outside of these envelopes. A similar relationship between Partition, Partition$^-$, Partition$^+$, LR1, and LR2 results occurs for D.2.2. The author found the different partition methods produced results that were sufficiently similar to each other compared to the other methods (LR G_i^* and MR G_i^*) that for this study using any particular one would come to the same general conclusions when comparing the partition results to the LR G_i^* method and the MR G_i^* results.

For the reasons outlined above, the presentation and discussion of the remaining results and discussion for the simulated datasets include only Partition results where the standard G_i^* is calculated for regional partitions (D.1.2; D.2.2).

RESULTS

Local Region Versus Partitioning

To evaluate the LR G_i^* I first run the method on datasets where the spatial process is theoretically stationary (D.1.1; D.2.1) and then on datasets where there are regional differences and the boundaries for true regions are known (D.1.2; D.2.2).

First I return to the maps discussed when I was considering edge effects (Figure 3) focusing this time on the similarities and differences of the Partition and LR1 analysis for the non-stationary datasets with four separate regions. The LR1 analysis identifies high and low *z*-scores where none occurred before in the Partition results and reduces some high and low *z*-scores identified in the Partition analysis to moderate *z*-scores (neither high nor low) because groups of points captured by the local neighbourhood near a partition boundary are compared (in the calculation) to a local region that crosses that boundary (Figure 3). The differences between the Partition and LR1 analysis occur near the partition boundaries, while similarities in high and low *z*-scores generally occur away from the transitional boundaries or where the change across boundaries (changing from regional mean to another) is less minor (between the bottom two quadrants of D.2.2 for instance) (Figure 3).

The LR1 local region results have the lowest percentage of counts of high and low *z*-scores compared to the results of the other analyses for non-stationary datasets, with the exception of the Partition results, which are always lower still (Table 3). The LR1 results for all datasets, whether the dataset is comprised of true regions or is a global trend, have the lowest

combined counts (high and low; >=2 and <=-2) compared to the multi-region and the Standard results. For the stationary datasets (1.1; 2.1), the LR1 results neither have the highest or lowest percentage of counts of high and low *z*-scores.

For datasets where true regions are known (D.1.2; D.2.2), high and low *z*-scores (>=2; <=-2) of the LR1 results have the highest coincidence with the high and low *z*-scores of the true Partition results. Recall that LR1 results come from the LR G_i^* method that uses a local region diameter equal to the length of one side of a partitioned region, such that the circular local region has approximately the same coverage as the true region (a little less in fact). For datasets where there are regional differences and the boundaries for true regions are known (D.1.2; D.2.2), the Partition results and LR1 results have the most similar counts of high and low *z*-scores (>=2 and <=-2) compared to any of the other results with the Partition results (Table 3). For the non-clustered true regions dataset, D.1.2, regional non-stationary process and no clusters, 31 of 169 (18%) of the high *z*-scores for LR1 are coincident with the high *z*-scores of Partition results. The next highest count of coincident high z-scores with the Partition results is LR2 results with 5 of 6 (83%) of LR2's high *z*-scores being coincident with those of Partition. LR1 results have the least number of coincident low *z*-scores with the Partition low *z*-scores, 33/148 (22%)(although they account for a high percentage of the total number of LR1 low *z*-scores relative to the other results). Arguably the other results sets have higher coincident counts only because of a much higher count of *z*-scores <=-2 (~ 2 times greater than LR1 results) increasing the chance that some of these *z*-scores will be coincident with those of the Partition results.

Table 3 Simulations - percent of locations with high or low Gi* z-scores (>=2 or <=-2) for various methods.

	D.1.1	D.1.2	D.1.3	D.2.1	D.2.2	D.2.3
Standard	.046	.763	.738	.013	.590	.585
Partition	.036	.095	.785	.007	.048	.160
LR1	.040	.317	.126	.010	.085	.158
LR2	.021	.588	.221	.015	.570	.308
MR Probability (>=50% High/Low)	.046	.693	.371	.013	.618	.400

For D.1.2, regional non-stationary process and no clusters, the average difference of LR1 from other result sets is lowest with Partition at 1.34 standard normal deviates. For clustered D.2.2, regional non-stationary process and clusters, 15/20 (75%) of the high results of LR1 (*z*-score >=2) are coincident with the Partition results and 1/14 (7%) for low results (<=-2) (Table 4). For *z*-scores <=-2 all other non-partition methods besides LR1 have zero coincidence with the Partition results (Table 4). For D.2.2, the average difference of LR1 from Partition is 0.448, and is the closest average difference of Partition from any other results set.

Table 4 Simulations - coincident high and low G_i^* z-scores (>=2;<=-2) between Partition and each of Standard, LR1, LR2 and MR Probability results.

Dataset (Partition count)	D.1.1 (23)	D.1.2 (41)	D.1.3 (391)	D.2.1 (1)	D.2.2 (17)	D.2.3 (92)
>=2						
	Counts					
Standard	3/7	11/297	186/371	0/5	5/127	51/122
LR1	8/12	31/169	96/124	1/4	15/20	27/28
LR2	5/6	11/271	96/237	0/6	7/130	51/66
MR Probability	3/7	11/275	0/0	0/5	7/133	51/92
<=-2	(13)	(54)	(394)	(6)	(2)	(68)
Standard	6/38	45/466	163/367	0/0	0/109	40/112
LR1	5/9	33/148	74/96	0/0	1/14	34/35
LR2	5/15	44/339	83/216	0/0	0/98	40/57
MR Probability	7/39	44/409	154/308	0/0	0/144	40/68

There is a limit where further reducing the size of the local region does not necessarily create results that are more coincident with those of Partition. For D.1.2, the difference between the LR G_i^* results and Partition results lessens as the local region is decreased until at 30 metres, after which the difference remains the same. For stationary datasets, decreasing the size of the local region does not result in a trend towards fewer high and low *z*-score counts, and counts fluctuate across the different scales. There is also a lower limit to which the local region can be sized after which the test statistic cannot be calculated as it results in division by zero. The LR2 results have the second highest level of coincidence with the Partition method after the LR1. This indicates that even if in a real-case scenario one was to miss-specify the true region to twice that of the true region, one would still be obtaining results more similar to a dataset partitioned into true regions than if I were to use the standard or multi-region methods. When I apply the same geographical divisions used to partition the regional non-stationary datasets (1.2 and 2.2) to the stationary datasets (1.1 and 2.1) where there are no true regions, the Partition results are no longer most similar to LR1. For the same partitions for the datasets with global trends (1.3 and 2.3) where there are no true regions, the Partition results are the least similar to the LR1 results.

The essential relationship of the LR1 and Partition results with reference to the Standard results for the different datasets can be seen in Figure 5 which shows cumulative frequency graphs of the LR1 and Partition results for each of the datasets. I use the cumulative frequency graph to observe what percentage of the total dataset falls below the traditional cutoffs of -2 for low *z*-scores and +2 for high *z*-scores. From a statistical perspective, *z*-

scores below -2 and above 2 should account for ~5% of the total results (~2.5% each), given that they are normally distributed (which in the case of the non-stationary/gradient data, is clearly not the case). Using the results of the cumulative distribution graphs, one can observe how conservatively the different methods assign high and low *z*-scores to the data. It is preferred that a method for detecting areas of high and low spatial association assigns high and low *z*-scores more conservatively (of course, it also matters which locations receive these *z*-scores). For the stationary datasets, the LR1, Partition, and Standard results have virtually indistinguishable distributions. The low *z*-scores for all results occur below the fifth percentile and the high *z*-scores all occur above the 95th percentile. For the non-stationary regional datasets, the Partition and LR1 results differ from the Standard results, but are similar to each other, although more so for the clustered dataset than the non-clustered. The Standard results have the greatest range of *z*-scores for the non-stationary datasets. The low z-scores for the LR1 results for D.1.2 fall below the fifteenth percentile and the high z-scores above the 83rd percentile. The low z-scores for the Partition method for D.1.2 fall below the sixth percentile and the high *z*-scores are above the 95th percentile. The low *z*-scores for the Standard results fall below the 47th percentile and the high *z*-scores above the 70th percentile; using the standard G_i^* for this dataset results in such a large number of high and low *z*-scores as to defeat the purpose of using the local measure (to identify unusual clusters of spatial association). For the clustered dataset, D.2.2, both the Partition and LR1 results have low *z*-scores below the fifth percentile and high z-scores above the 95th percentile. The low *z*-scores for the Standard results for D.2.2 fall below the 28th percentile and the high *z*-scores above the 68th percentile. Finally, for the gradient data (global trend), the Partition and LR1 results have a greater disparity than they did for the other datasets (more so for the non-

clustered data than the clustered) and the Standard results again show the greatest range of *z*-scores. For D.1.3, the low *z*-scores of the LR1 results fall below the tenth percentile and the high *z*-scores above the 87th percentile; the low *z*-scores of the Partition results are below the 40th percentile and the high *z*-scores are above the 60th percentile; the low *z*-scores of the Standard results are below the 37th percentile and the high *z*-scores are above the 62nd percentile. In this case the LR1 results have a more conservative count of high and low *z*-scores than the Partition results, which makes some sense given that false partitions were used for the global trend dataset.

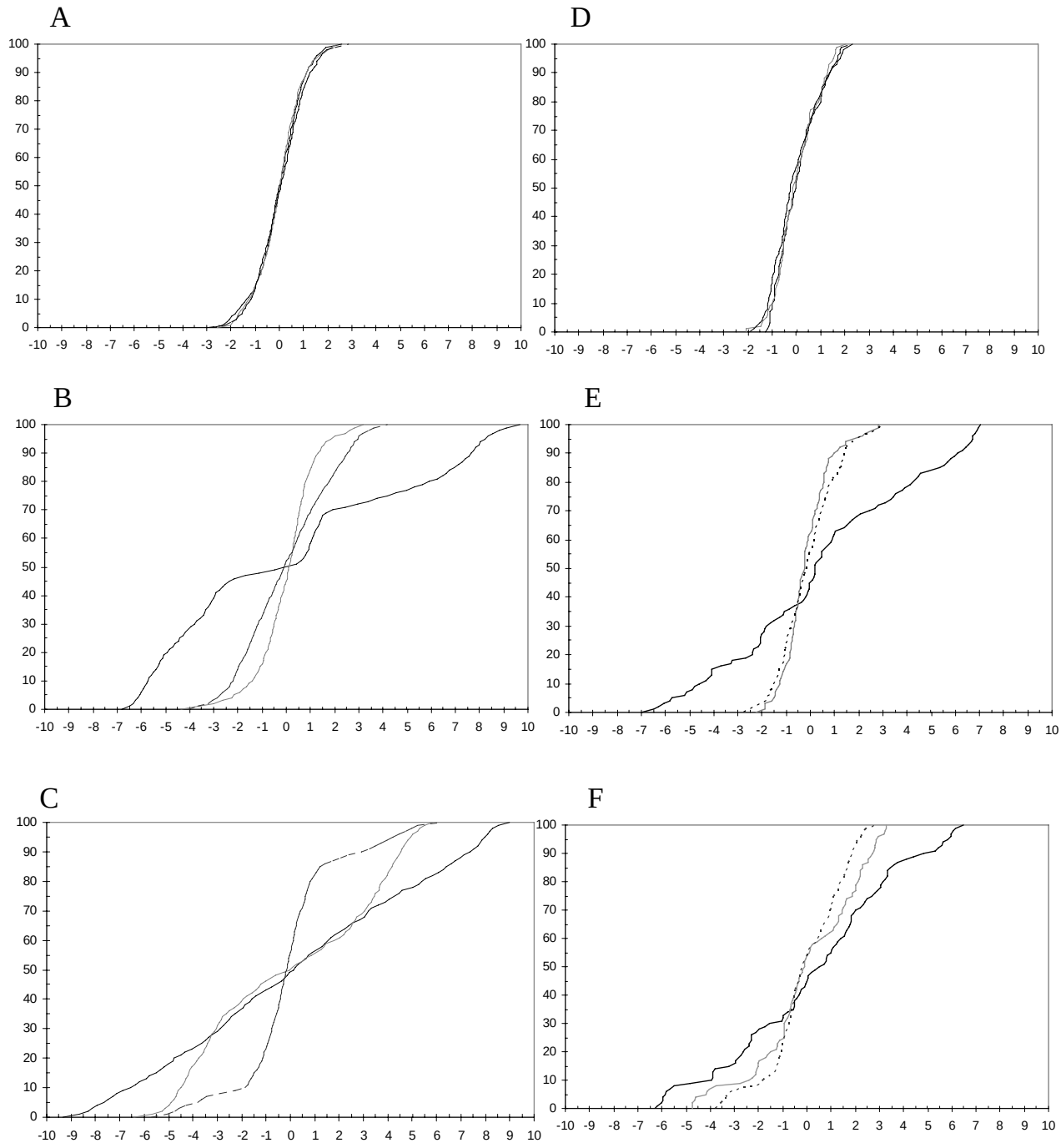


Figure 5 Simulations - Cumulative frequency graphs of LR1, Partition and Standard results for the non-clustered datasets (A. D.1.1, B. D.1.2, and C. D.1.3), left, and the clustered datasets (D. D.2.1, E. D.2.2, and F. D.2.3), right. Standard = black; LR1 = dashed; Partition = grey

Multi-Region Versus Standard

The purpose of this section is to describe and evaluate the performance of the MR G_i^* first for datasets where the spatial process is stationary (D.1.1; D.2.1), second where there are regional differences in the spatial process and the boundaries for true regions are known (D.1.2; D.2.2), and third where the spatial process is a global trend (gradient) (D.1.3; D.2.3).

For all simulated datasets, MR Median and MR Probability (50% probability >=2 or <= -2) obtain similar counts of high and low *z*-scores (>=2 and <=-2) and there is always less than 1 standard normal deviation between the two results set at any location. MR Mean is a close second. As I feel that conceptually the MR Probability is more easily interpreted than the MR Median or MR Mean results, and because these latter methods produce very similar results to MR Probability, I will for the remainder of this thesis only report for the MR Probability results. However, I feel it is important to note that the MR Median results, and the MR Mean to a slightly lesser extent, in all cases receive nearly identical scores to the MR Probability method and that this may give some indication of the strength of the MR Probability method using the 50% threshold to obtain representative scores for a single location across multiple scales of a local region. In other words, using a 50% threshold for MR Probability is for the majority of instances capturing those locations that are on average (and on median) high or low (>=2 or <=-2), whatever the case may be, across the multiple regions.

The MR Probability results consistently have higher counts of high and low z-scores than the local-region results (LR1 and LR2)(Table 5). MR Probability (50% threshold) consistently

has similar counts of high and low *z*-scores to the Standard results and similar *z*-scores across all results. For stationary datasets D.1.1 and D.2.1, MR Probability results and Standard results have near identical numbers of high and low *z*-scores. For D.1.1, stationary process and no clusters, the counts of high and low *z*-scores is low for all MR Probability G_i^* results relative to the total number of points in the dataset and in comparison to the counts of high and low z-scores for the MR Probability results for D.1.3, where there is a global trend, and D.1.2, where there is a regional non-stationary process. For D.2.1, stationary process and clusters, the counts of high and low *z*-scores for all MR Probability (50%) again are low relative to the total number of points in the datasets and in comparison to the counts of high and low *z*-scores for the other clustered datasets (D.2.2; D.2.3). For the datasets with global trends (D.1.3; D.2.3), the Standard results have greater numbers of high and low *z*-scores than the MR Probability (50%) results; however, the results are still highly coincident as I describe below.

The high and low *z*-scores of MR Probability are highly coincident with the high and low *z*-scores of the Standard results. D.1.1, stationary process and no clusters, where 7/7 (100%) and 38/39 (97%) of high and low MR Probability *z*-scores respectively are coincident with high and low *z*-scores for the Standard results; thus, only 3% of all high and low *z*-scores for MR Probability are not coincident with the high and low *z*-scores for the Standard results, and these are all for clusters of low values (Table 5). For D.2.1, the high and low *z*-scores for MR Probability are completely coincident with the Standard high and low G_i^* *z*-scores and the high and low *z*-scores for the Standard methods have complete coincidence with MR Probability. The majority, but not all, of LR1 and LR2 high and low *z*-scores are coincident

with MR Probability and the Standard high and low *z*-scores. For D.1.3, 84% of the high MR Probability results (>=2) and 84% of the low results (<=-2) are coincident with the high and low results of the Standard results; thus, 16% of MR Probability results are >=2 when the Standard results are not and 16% of MR Probability results are <=-2 when the Standard results are not.

Table 5 Simulations - coincident high and low G_i^* z-scores (>=2;<=-2) between Standard and each of Partition, LR1, LR2, and MR Probability results

>=2						
Dataset	D.1.1	D.1.2	D.1.3	D.2.1	D.2.2	D.2.3
(Standard count)	(7)	(297)	(371)	(5)	(127)	(122)
	Counts					
Partition	3/23	11/41	186/391	0/1	5/17	51/92
LR1	5/20	110/169	40/63	3/4	5/20	27/28
LR2	4/12	249/271	124/125	4/6	101/130	66/66
MR Probability	7/7	275/284	0/0	5/5	113/133	92/92
<=-2						
(Standard count)	(7)	(54)	(367)	(0)	(109)	(112)
Partition	6/13	45/54	163/394	0/0	0/2	40/68
LR1	6/20	105/148	34/63	0/0	11/14	33/35
LR2	6/9	316/339	91/96	0/0	91/98	57/57
MR Probability	38/39	407	308/308	0/0	105/144	68/68

Figure 6 illustrates maps of the Standard, MR Probability (50%), and LR1 results for D.1.2 and D.2.2. It can be seen from the maps how the MR Probability (50%) results are identifying high (>=2) and low (<=-2) *z*-score values in areas where the standard method identifies none, and the MR Probability results are simultaneously retaining much of the same general pattern of the Standard results. For D.1.2, the sharp divisions in regional differences in the mean (high standard deviation and distinct grouping of points with either high or low values relative to the global mean) are partly to explain for the large counts of

high and low *z*-scores for the Standard and MR Probability results. If there were only two regions in the dataset, one encompassing the majority of the study area, and the other encompassing a small portion, the counts of high and low *z*-scores would be much more similar to what is seen with a completely stationary dataset. The effect of a regionalization on the *z*-scores of the standard method are to increase the counts of high and low *z*-scores. Similarly, the number of high and low *z*-scores of the MR Probability (50%) results relate to the size and number of regions in the study area and the level of deviation between the values of the region. The general conclusion is that the MR Probability (50%) obtains similar results to the standard G_i^* method for the simulated datasets, but identify subtle variations in the spatial pattern that are lost with the standard method. The difference between the MR Probability and Standard results is more pronounced for the case study (see next sections).

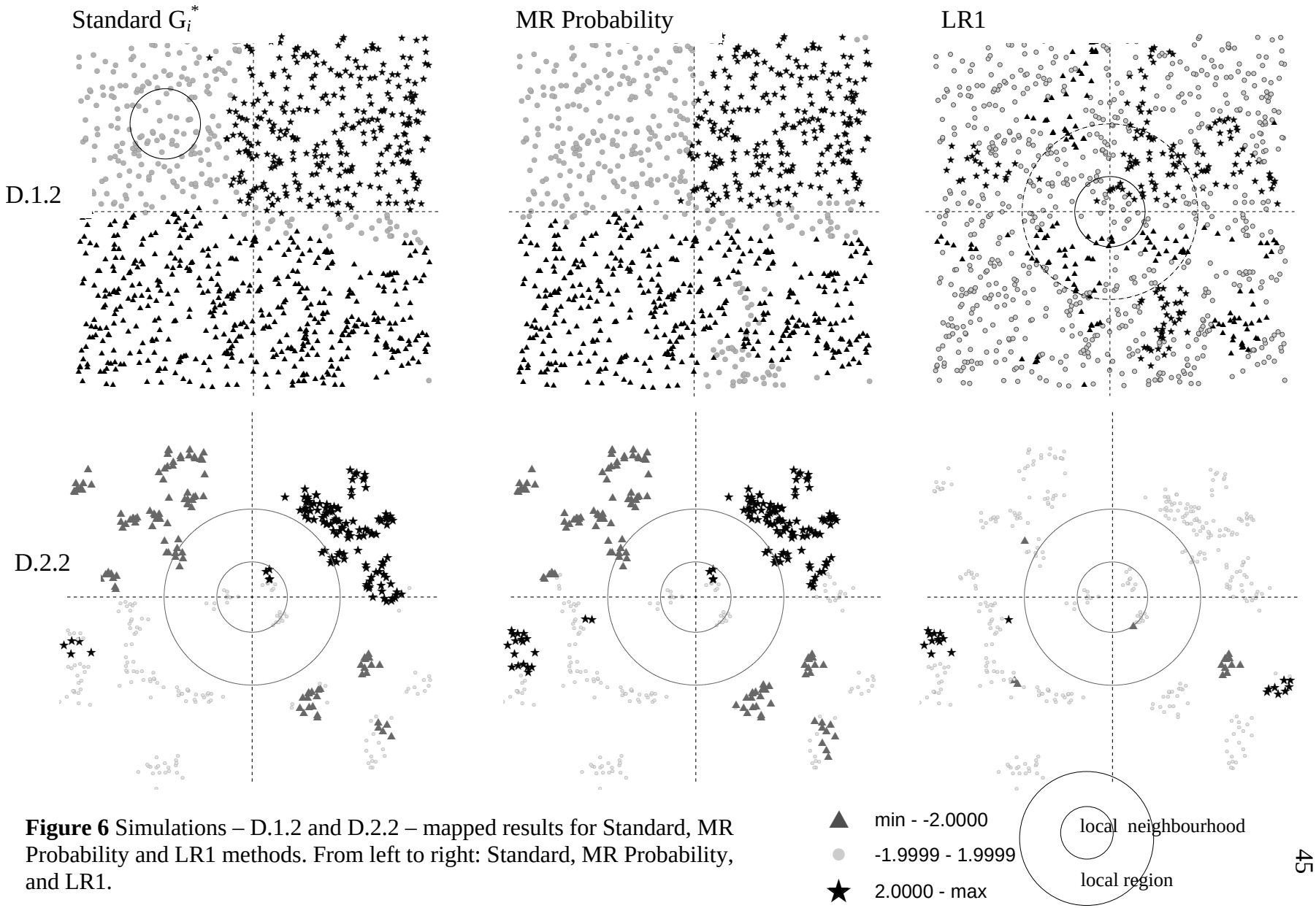


Figure 6 Simulations – D.1.2 and D.2.2 – mapped results for Standard, MR Probability and LR1 methods. From left to right: Standard, MR Probability, and LR1.

DISCUSSION

As expected, where there is little or no regional difference in the mean and thus the spatial process is stationary, there is more agreement between methods as indicated by the distribution for the different methods in the cumulative frequency graph (D.1.1 and D.2.1). The different methods achieve similar counts of high and low G_i^* *z*-scores when the process is stationary, and these counts account for a small percentage of the total counts. Where there are regional differences in the mean, there is disagreement between the results of the different methods also as indicated by the lines of the cumulative distribution graph which show a clearer separation between the results and an order in the relative range of the results for the different methods that is consistent across the different datasets. As expected, very few high or low G_i^* *z*-scores result from any of the methods run on the stationary datasets, D.1.1 and D.2.1, and comparatively there is a considerable increase in the number of high and low *z*-scores for the standard results for the non-stationary datasets (D.1.2, D.1.3, D.2.2, and D.2.3), but for the other methods there is distinct variation in the high and low counts calculated. The LR G_i^* method results LR1 obtain similar values to a true partition. I hypothesize that differences between the LR1 and Partition results are due to the ability of the LR G_i^* to calculate the statistic across transitional boundaries between regions. Even when a true region is not known so that LR can not be adjusted to reflect that, the trend is for lower more conservative *z*-scores to be achieved when a smaller local region is used. The local neighbourhood of the LR G_i^* method retains much of the pattern of the standard on full partition, but allows points at border in the transition area to be compared to global values beyond the region in question.

The MR method results can be taken apart to examine each of the results sets for each local region and this could be used to determine the scale of a true region, should such a region exist.

The MR Probability results evaluate how unusual a location (and its neighbours) is relative to multiple scales of a local region. Additionally, the MR Probability results allow regions with distinctly different means compared to the surrounding regions to contribute to the statistic. For the simulated datasets, the MR Probability results receive similar *z*-scores to Standard results and for similar locations but are able to pull out more subtle local variations. By comparing the results of several different methods, LR, MR, and Standard, one is able to make inference on the type of spatial process that is occurring for a dataset.

Extreme regions (regions with unusually high or low means relative to the remaining regions) will have an effect proportional to their size in the study region. This is the case for the regional datasets D.1.2 and D.2.2, but particularly the former. In D.1.2, the NE quadrant has a very strong influence over all results of the MR methods. If this region were much smaller, for example, a tenth the size, one would observe far fewer high and low *z*-scores for the datasets and a much more heterogeneous spread of *z*-scores across the points (I did in fact run such an experiment to confirm this). Of course, the Standard method run on such a dataset would also calculate fewer high and low *z*-scores, with a few small, albeit important, differences from the MR Probability results in what locations receive high and low *z*-scores.

PART B

In Part B, I demonstrate potential methods for determining an optimal local neighbourhood and local region, and for detecting transitional areas in the study area.

OPTIMAL LOCAL NEIGHBOURHOOD AND LOCAL REGION

Selecting a neighbourhood size is more of a concern for modeling than for exploratory analysis. When conducting exploratory spatial data analysis (ESDA) several neighbourhood sizes can be used and the results compared. Such an approach is helpful to investigate the scale at which spatial processes act. For modeling, however, a local measure is sometimes used to create a weights matrix for incorporating spatial autocorrelation (association) into an environmental model, as for instance is done in two recent studies Lin and Lu (2006) and Aldstadt and Getis (2007). The multi-region approach can be used to determine an appropriate local neighbourhood size. However, the LR G_i^* introduces a new concern for modelers. What is the best local region to use? I will demonstrate how a local region can also be selected using the MR G_i^*.

In the case of using local spatial association methods to create a weights matrix to be used in regression models, it would be necessary to identify a single set of neighbourhoods/regions. I can do this using the multi-region approach by searching for a local neighbourhood / local region combination that identifies ~5 % of locations as significant (Figure 7) (the tails of the distribution of the results). Under complete spatial randomization one would expect high and low scores to account for ~5% of total scores, ~2.5% for each tail. For the stationary random datasets (D.1.1 and D.2.1) the high and low G_i^* *z*-scores remain at ~5% of total scores across all increments and no changing pattern is observed with increasing distance of the local region.

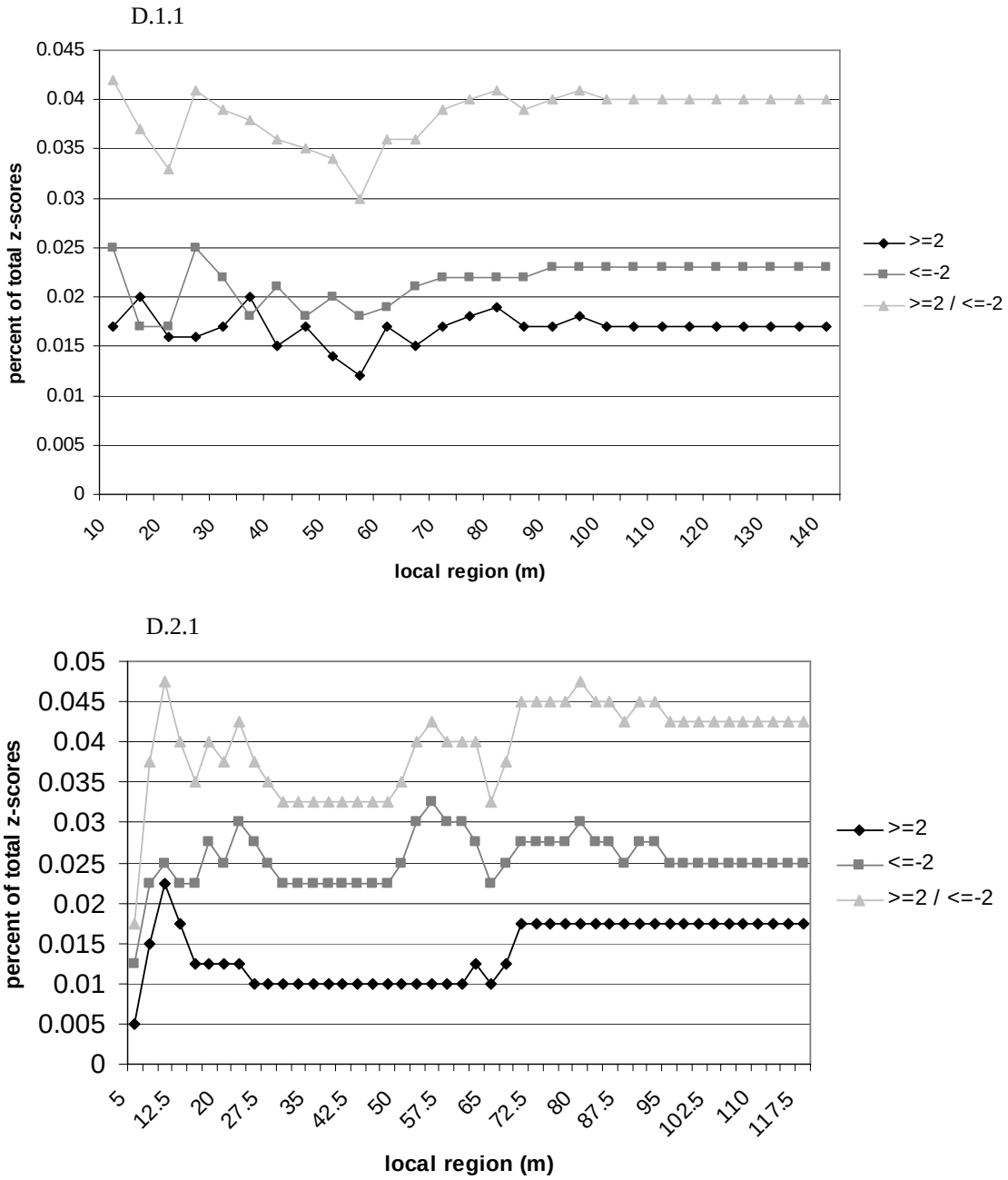


Figure 7 – Simulations – D.1.1 and D.2.1 – Graphs of MR Probability results across each local region using a local neighbourhood of 2.5 metres.

For non-stationary D.1.2 combined high and low *z*-scores of ~5% are achieved from 10 to 42.5 metres of the local region beyond which there is a rapid increase in total high and low scores until they level off at 16% for high scores at 90 m and 8% for low scores at 118 m (Figure 8).

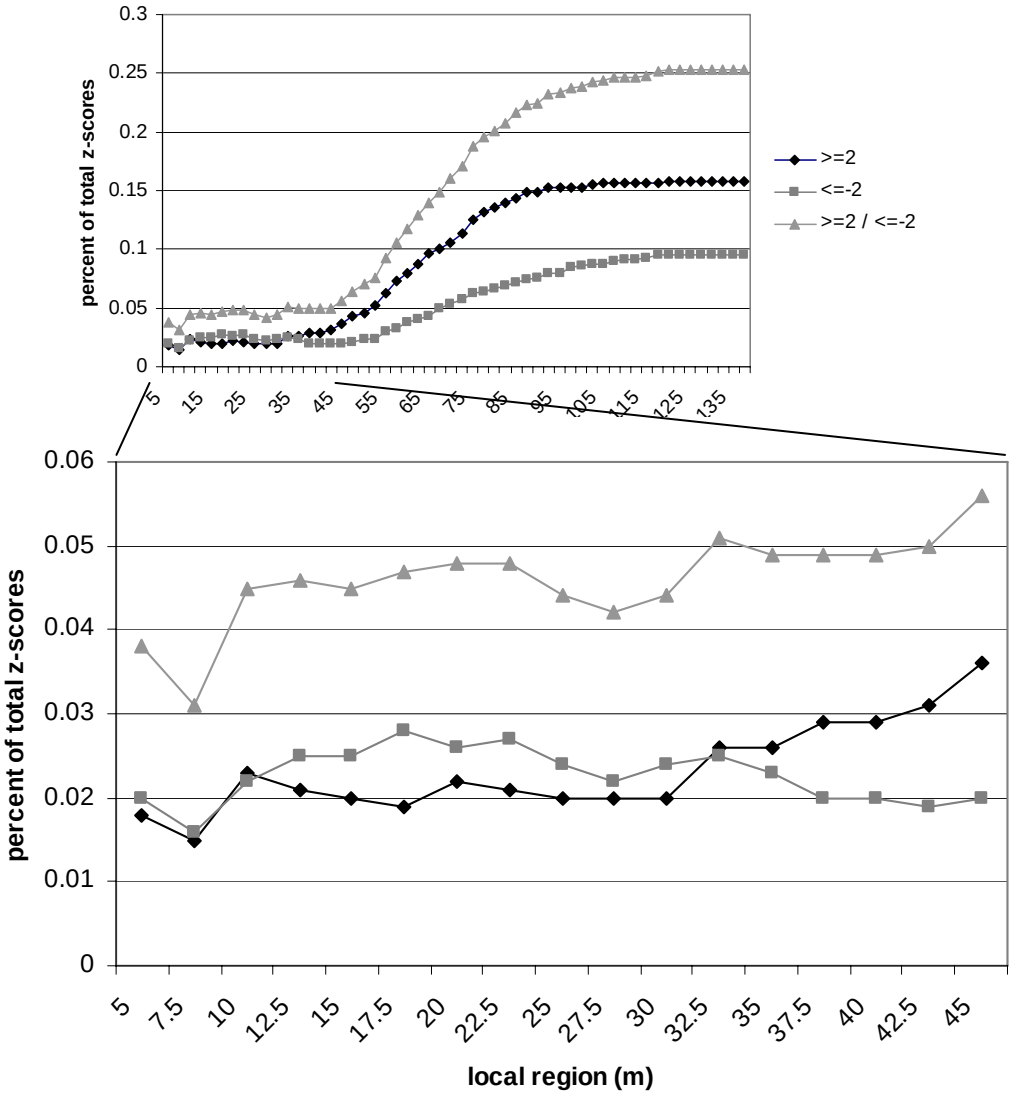


Figure 8 Simulations – D.1.2 – graphs of MR G_i^* probability results for each local region, using a local neighbourhood of 2.5 metres.

A conservative approach would be to choose the smallest local neighbourhood size possible to account for fluctuation in the size of a local region across the study area. Consequently, I determined a local neighbourhood of 10 metres to be appropriate for D.1.2 since it is the

smallest size of the local region to capture ~5% of combined high and low scores. Next, I ran the MR G_i^* method using a local neighbourhood of 10 metres. For this new set of MR G_i^* *z*-scores (Figure 9) the local region size of 35 metres occurs just before a major departure between the percentages of high and low scores (up to this region size the graph indicates a close agreement/symmetry between high and low scores and a total percentage of high and low scores below 5%). I would therefore select this as the local region since beyond this local region the scores are indicating more variation in the data as indicated by the departure of the relative percentages of high and low scores; at this point there are marked differences between points captured by the local neighbourhood versus the local region and an imbalance

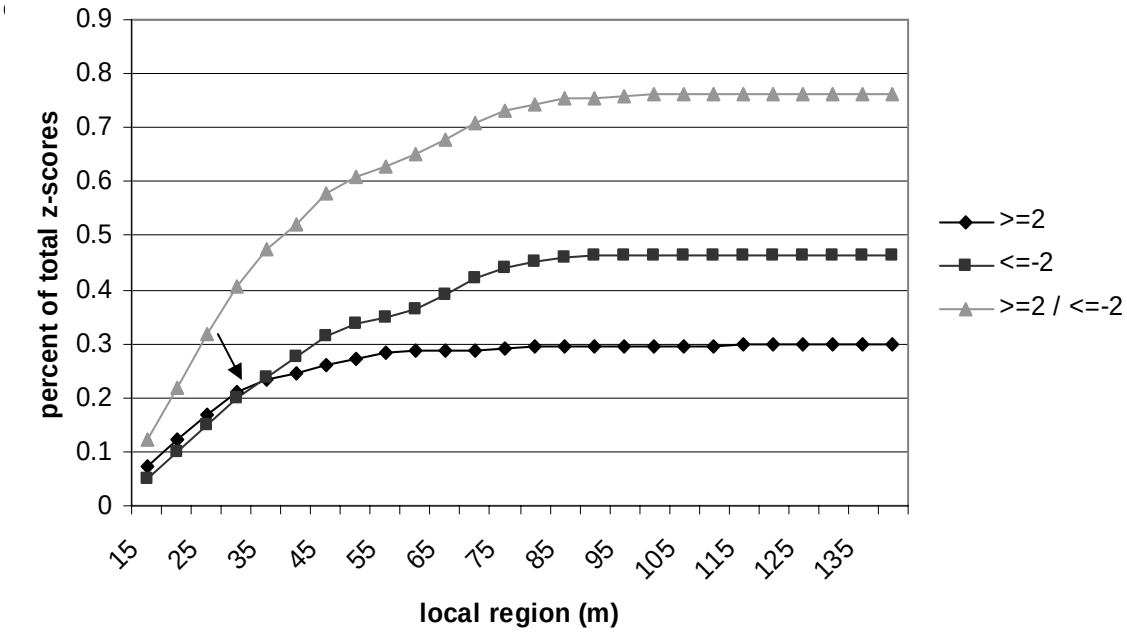


Figure 9 – Simulations – D.1.2 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 10 metres. Here, the local region is incremented by 5 metres.

For D.2.2, identifying the local neighbourhood and local region is an easier task than for D.1.2 because the clustering accentuates the changes across the study area (this is also true for the case study – to follow). I ran the MR G_i^* using a local neighbourhood of 2.5 metres. At a local region size of 5 metres the total percentage of high and low scores is below 5%

although slightly imbalanced between the two (high = 3% low =2%); at a local region size of 7.5 metres the total percentage of high and low scores is above 5% (5.5%), however, the high and low scores are now balanced each accounting for 2.75% of the total scores (Figure 10). I determine that an appropriate local neighbourhood size would lie somewhere between these two numbers (5m and 7.5m). For the local region, I select the first peak at around 15 metres. Because the choice of a single local neighbourhood and local region combination must necessarily generalize the pattern occurring throughout the dataset, the selection of a local neighbourhood and a local region can not in this sense be perfect but the aim should be to choose a combination that will account for the most local variation without being subsumed by regional differences. A way around this imperfection would be to apply this approach on a point by point basis, therefore having a local neighbourhood / local region combination that would change depending on the spatial pattern occurring in the vicinity of any one point.

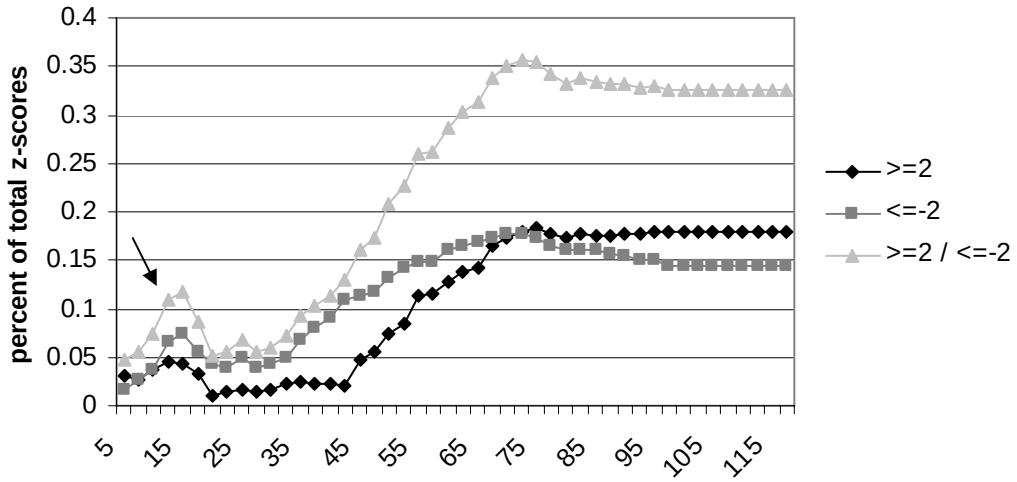


Figure 10 - Simulations – D.2.2 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 2.5 metres.

The datasets with global trends (1.3 and 2.3) can be distinguished from the other datasets by the linear trend that manifests for the MR G_i^* method with a symmetry between high and low scores that spans across the range of local regions (Figure 11).

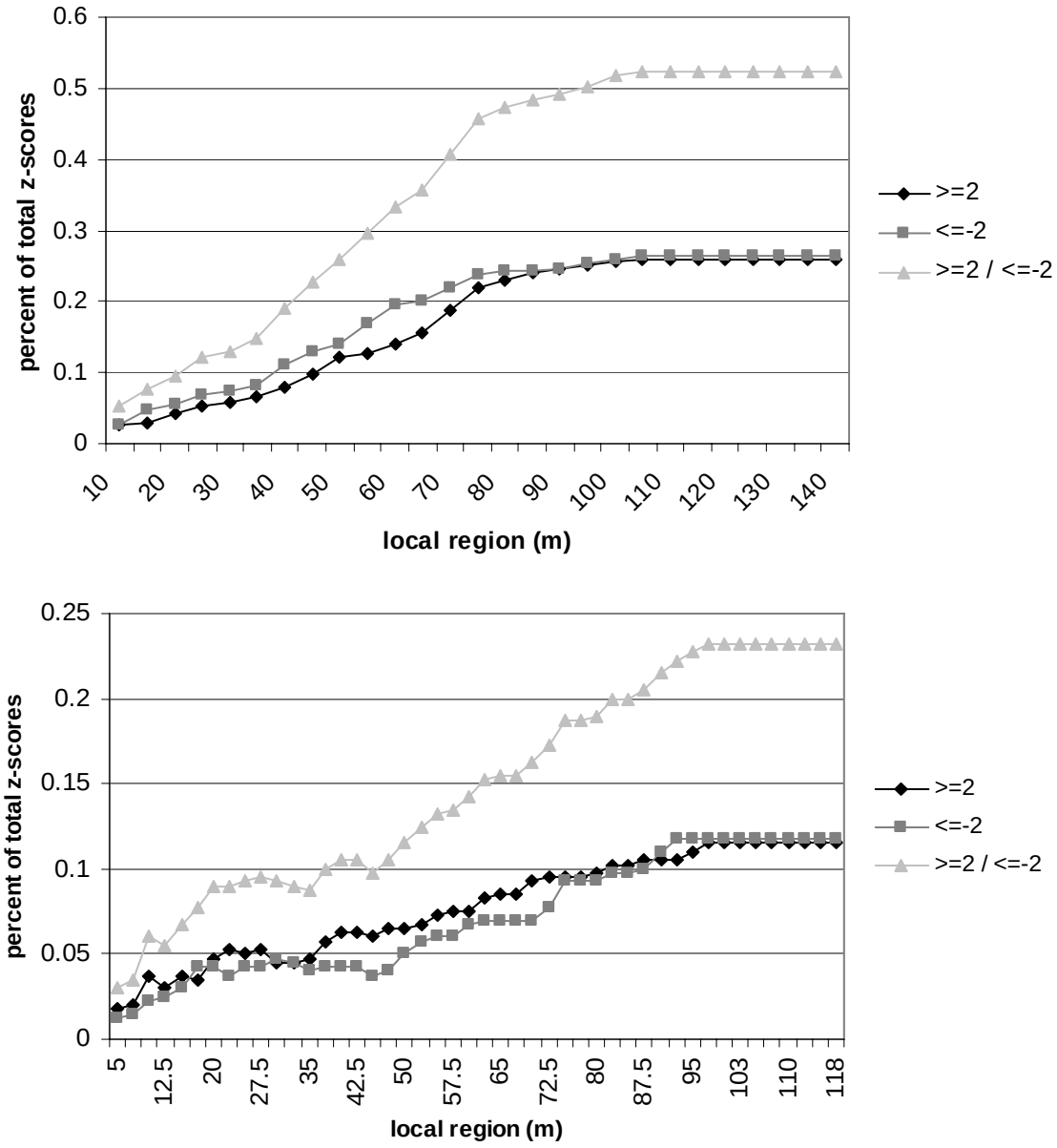


Figure 11 – Simulations – D.1.3 and D.2.3 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 2.5 metres.

MULTI-REGION DETECTING TRANSITIONAL AREAS

A multi-region has the potential to be used to determine the location and dimensions of true regions should they exist. Although it is beyond the scope of this paper to delve far into this subject, I introduce the idea here. The set of *z*-score values of any point across the multiple scales of the MR G_i^* can be graphed and major changes in the *z*-scores identified as the distance at which the local region is capturing a major shift in the value of marks. By graphing multiple points one should be able to triangulate where regional boundaries exist. Figure 12 illustrates how one can begin to determine the location of regional boundaries. The figure includes graphs of the distribution of *z*-scores over multiple scales and accompanying maps showing the size of the local region where major shifts in *z*-scores occur, for two sets of points, one pair for D.1.2 and the other for D.2.2. There is also potential for detecting transitional areas by identifying those locations where the Standard results are high, or low, and the MR Probability methods are not, and vice versa, but I leave this for future research.

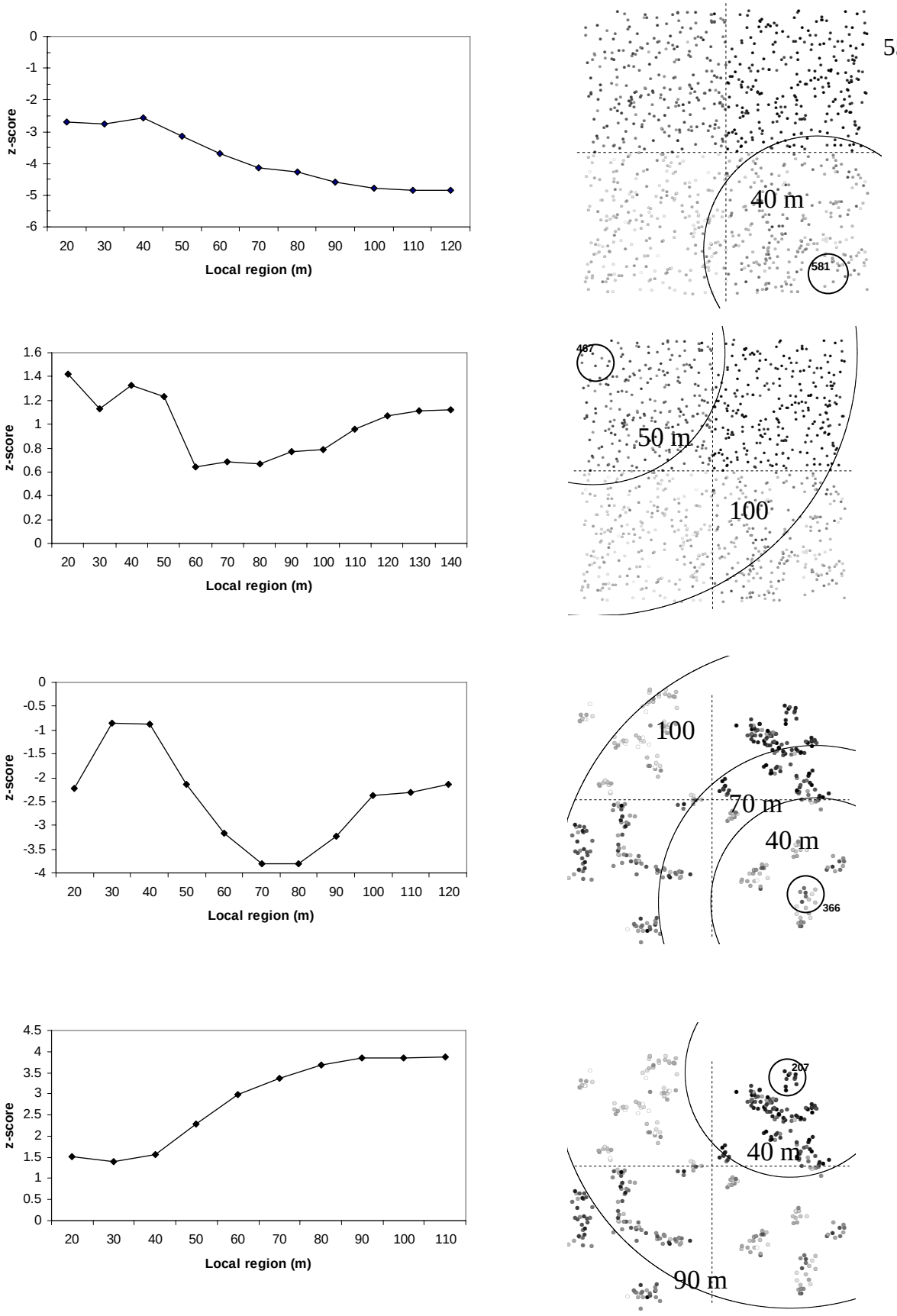


Figure 12 Simulations – D.1.2 and D.2.2 – Detecting transitional areas on a point-by-point basis. Graphs of the distribution of *z*-scores over multiple scales of the local region (left) accompany maps showing the size of the local region where major shifts in *z*-scores occur. The top two examples are from D.1.2 and the bottom two from D.2.2. The smallest circle is the local neighbourhood.

DISCUSSION

There is potential to use a multi-region approach to determine optimal local neighbourhood and local region sizes. This can be done by considering all of the dataset together or on a point-by-point basis. Although here I have selected the local neighbourhood size and local region size by visually assessing graphs of the multi-region probability scores at various scales of the local neighbourhood and region, for a point-by-point approach this would best be done programmatically. Future work would involve designing an algorithm to be able to detect the “pits” and “peaks” in the multi-region probability scores for individual point locations.

CHAPTER 5 – CASE STUDY

INTRODUCTION

The mountain pine beetle (*Dendroctonus ponderosae*) is the most significant agent of mortality in mature lodgepole pine (*Pinus contorta*) forests in western North America (Furniss and Carolin 1977). There are two main factors that have contributed to the successful expansion of the beetle population in British Columbia: 1) the large amount of mature lodgepole pine on the land base, which has tripled in the last century as a result of intensive fire suppression activities (Taylor and Carroll 2004), and 2) several successive years of climatic conditions favourable to beetle survival (Carroll et al. 2004). As of 2005, the mountain pine beetle infestation has affected 7 million ha of British Columbia's forest (Westfall 2006).

Identifying forest stands with the greatest risk for timber losses as a result of mountain pine beetle attack is critical information for mitigation and forest management planning (Wulder et al 2006). Decision support systems that incorporate the movement of pests and disease allow managers to fight outbreaks of these phenomena more efficiently (e.g., Hawksworth et al. 1995). Identifying forest stands prone to attack by mountain pine beetle provides forest managers with information to assist in mitigation and management planning. Areas identified as high risk can be targeted for sanitation harvesting or other mitigation activities (Dymond et al 2006). The dispersal and population density of pests are key elements of evaluating risk; however, landscape-scale movement of pests and diseases can be difficult to measure and predict (Liu et al. 2006). The relative importance of source population density and spread distances to the evaluation of risk associated with mountain pine beetle attack has only begun to be investigated (Wulder et al 2006).

A recent example of why quantifying spatial pattern is so important for managing the mountain pine beetle infestation in British Columbia comes from the work of Wulder et al (2006). Wulder et al (2006) applied spatial methods to calculate beetle pressure parameters for the Shore and Safranyik (1992) index, which has been the primary model used to predict the susceptibility of a forest stand to mountain pine beetle attack in the province of British Columbia. The Shore and Safranyik (1992) risk index is a function of both stand susceptibility and beetle pressure. Susceptibility is the level of damage that may occur when the stand experiences a mountain pine beetle infestation and beetle pressure is the probability that the mountain pine beetle will enter the stand. Wulder et al. (2006) applied a traditional distance based model, as well as an alternative density-based model implemented with a Voronoi tessellation to calculate beetle pressure. They found comparable trends between the two risk rating outputs but that overall, the density-based model of beetle pressure developed using the Voronoi polygons, generated high risk ratings that more closely corresponded with actual attack by the mountain pine beetles.

Although in this thesis I do not explore the option of using these methods for calculating beetle pressure, Wulder et al (2006) have demonstrated that this is an area worthy of further research, and I believe there is potential for a method such as the LR or MR G_i^* to be used to calculate beetle pressure, with the advantages that have been evaluated and discussed throughout this thesis. For instance, the G_i^* methods incorporate the density of points as part of their calculations, where as the simple distance based model (Wulder et al. 2006) does not and this appears to be its main shortcoming over the Voronoi-polygon method.

The case study is meant as an example and evaluation of how the LR and MR G_i^* works on real datasets. The case study is not meant as an analysis of the mountain pine beetle, *per se*, since an adequate link between pattern and processes influencing the density and dispersal of the mountain pine beetle would require a full consideration of other environmental variables including such things as wind, climate, and forest composition. Furthermore, a thorough analysis would need to involve an evaluation for all years that data were collected, not just for the two years I have selected here for demonstration purposes. A full analysis of the mountain pine beetle with respect to the patterns observed using the LR and MR G_i^* is beyond the scope of this thesis, but I do discuss how one might begin to take the observed patterns and try to link them back to other environmental variables and processes.

In Part A of this chapter I evaluate the performance of the LR and MR G_i^* methods on the Morice TSA mountain pine beetle infestation datasets (in the same manner which I explored them in Part A of Chapter 4 on the simulated datasets). In Part B, I extend the evaluation to consider other important aspects of a multi-regional approach including: 1) detecting transitional areas, 2) determining local neighbourhood and local region sizes, and 3) conducting randomization tests to evaluate the appropriateness of high and low cutoffs (>=2; <=-2) . First, however, I describe the case study data and study area.

DATA COLLECTION AND STUDY AREA

From 1995 to 2005 forestry consultants conducted helicopter-aerial surveys of mountain pine beetle infested trees in the Morice Timber Supply Area (TSA), centred on Houston British Columbia, at 6025800mN, 658250mE UTM Zone 9 (Figure 13). The TSA covers approximately 15,000 km^2 (1.5 million hectares). Geographical positioning systems were used to mark the centroids of clusters of red-killed trees, those trees having been infested approximately one year prior and displaying red colouration of their leaves (i.e., needles). For the case study I apply the methods described in this paper to compare data collected in 1996 and 2001 (Nelson et al 2006). Figure 14 shows kernel density maps of the number of infested trees for the two selected years of mountain pine beetle data. By 1996, the beetle infestation in the Morice TSA is widespread, but in part because it originated in the north and then spread south and east, and additionally because where it does occur it is still localized in densely infested areas, the general pattern is a heterogeneous spread of high and low values across the landscape. By 2001, the density of infestation across the landscape has become more homogeneously spread, more akin to what others have described as the continuous nature of the mountain pine beetle infestation at high levels of infestation (Carroll and Safranyik 2004), and this has lead some researchers (Nelson et al. 2006) to evaluate the advantages of representing this data as a continuous surface, more specifically by calculating kernel density estimator surfaces from point datasets and outputting as a raster surface. I maintain the datasets as points in this study, but do consider some of the issues identified in Nelson et al. (2006) with regards to shortcomings of point datasets and point-based analysis. I discuss the characteristics of the 1996 and 2001 datasets in further detail below, but first I describe the data more generally.

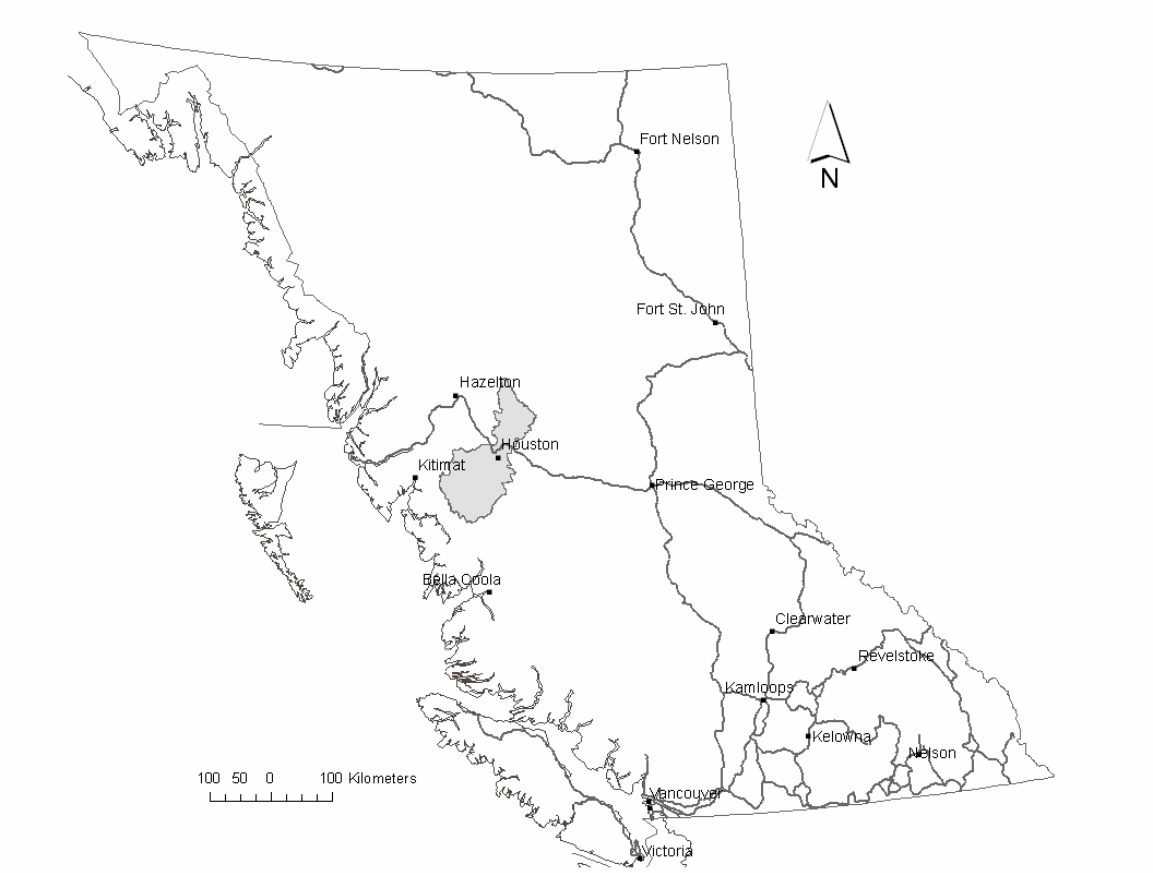


Figure 13 – map of British Columbia showing the location of the Morice Timber Supply Area (TSA) (shaded), where the case study data was collected.

The maximum area represented by a point is typically 3 hectares, equivalent to a circle with a radius of 100 m (Nelson et al., 2006). Research has indicated that errors associated with the heli-GPS points (as determined by direct comparison to ground surveys) were small; when estimating the numbers of attacked trees, 92.6% of heli-GPS points have

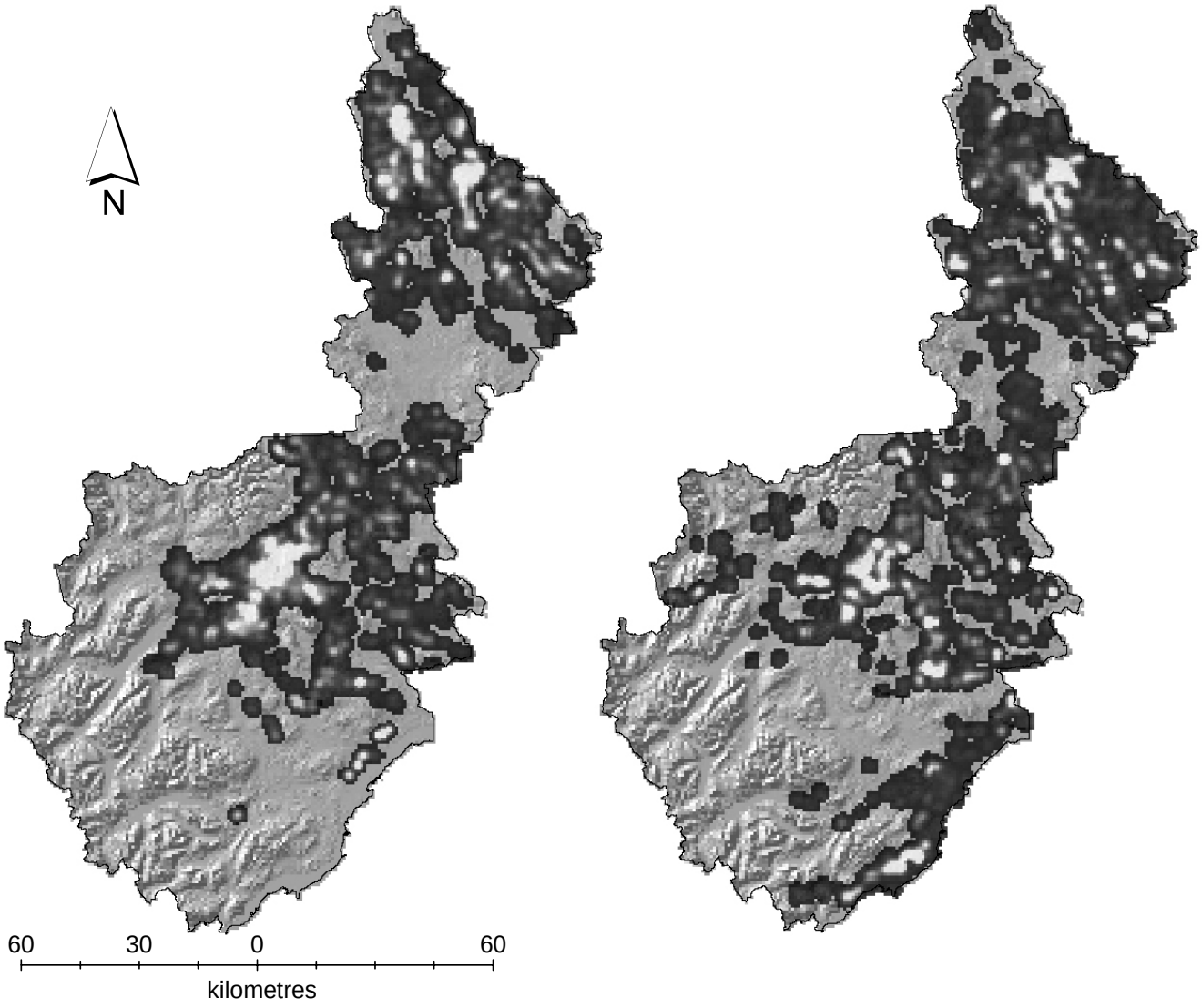


Figure 14 – Case Study – kernel density maps of the marks (number of infested trees) for 1996 and 2001 in the Morice Timber Supply Area. Cellsize = 200 m. Dark = Low values; Light = High values

errors of ± 10 trees (Nelson et al., 2006). To avoid problems of non-stationarity of spatial processes across the study area, previous research manually partitioned the study area into north, central and south regions based on the timing of the mountain pine beetle attack (Nelson et al 2006). It is one of the goals of this thesis to demonstrate how the LR and MR G_i^* can be used to avoid having to partition the study area.

To begin to understand the results of the standard and modified G_i^* for the case study data, it is useful to describe the datasets for 1996 and 2001. From the histograms with rug plot for the two different years (Figure 15), it can be seen that in 2001 there are several very high extreme values (compared to those of 1996). However, one can see that apart from the difference of a few extreme values the distributions of the 1996 and 2001 datasets are quite similar. Then why are the G_i^* results so different between the years? To answer this one needs to look more carefully at the spatial distribution of the values.

In 1996, the mountain pine beetle infestation is still in its early development phase and there is a clear distinction between the number of trees infested in the north versus the central region and the south. Additionally, to this north-south trend, in 1996 across the landscape there are a scattering of locations with very high counts relative to the other

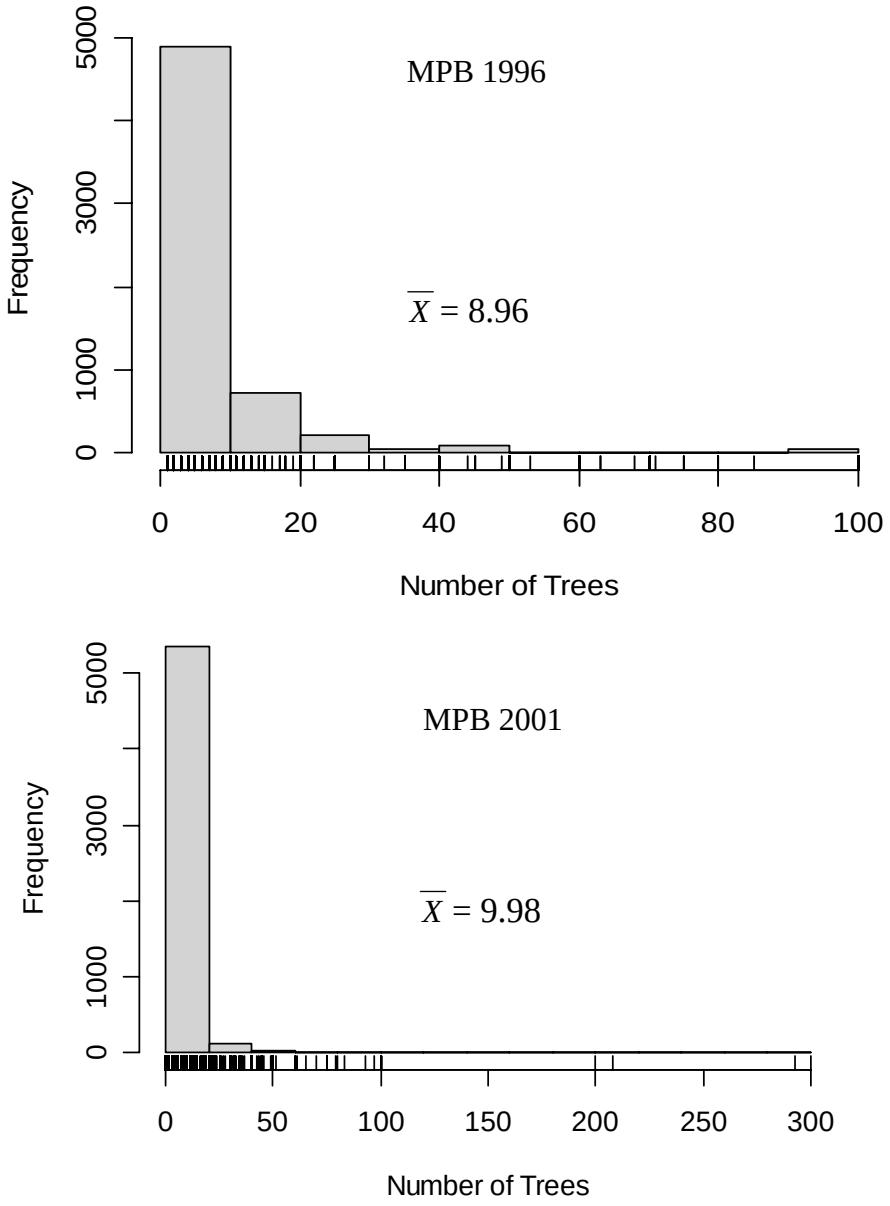
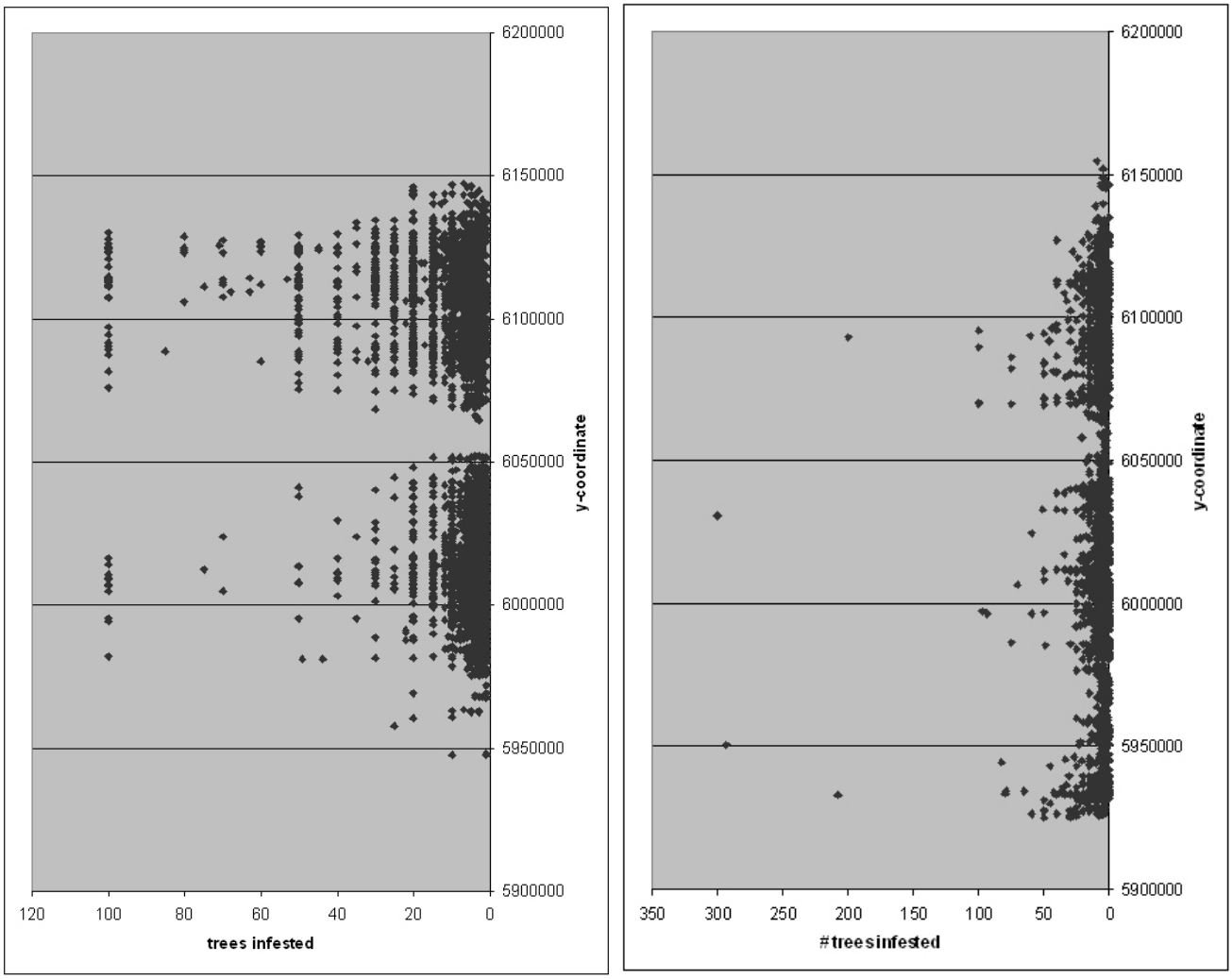


Figure 15 – Case Study - Histograms with rugplots for the case study datasets. The x-axis represents the number of trees associated with each point in the dataset.

points while in 2001 this is no longer the case. From an ecological perspective early spot infestations near the beginning of an epidemic occur as small areas of high values surround by large areas of low values (as in 1996), whereas once the epidemic is more developed the surface of high values becomes more continuous. From Figure 16 one can see that the attribute values (marks) are distributed bimodally in 1996 across the y axis and are more evenly spread in 2001. What is not easily discerned from the graph for 1996, is that there are higher counts of high valued points in the north region than in the south. This can be seen by observing the table included in Figure 16 which compares frequency classes of values between the north and south of the study area (split between the two modes). There are far more points in the classes of values >=20 and subsequent classes for the north region than for the south region. In contrast, there is a much more even split of values between the north and south sections of the 2001 data (split at the same geographic location as for the 1996 data) (see table in Figure 16).

1996 2001



	South	North
<10	2700	1593
10-20	336	676
20-30	75	336
30-40	20	115
40-50	10	37
50-100	16	103
>=100	12	47

	South	North
<10	2363	2194
10-20	420	313
20-30	78	59
30-40	25	15
40-50	9	15
50-100	20	11
>=100	3	6

Figure 16 – Case Study – 1996 and 2001 – scatterplots of the marks (number of trees infested) along the y-axis for 1996 and 2001.

PART A

In Part A, the author evaluates aspects of the LR and MR G_i^* methods with respect to the mountain pine beetle case study data and as an example of how these methods can be applied in an ecological setting.

METHODS

For both years of the case study, I use a distance band for the radius of the local neighbourhood of 2500 metres for the initial evaluation of the local and multi-region approaches. This distance is approximately suitable given knowledge of average maximum beetle dispersal (not including the rarer event of long range dispersal via above-canopy winds) from capture recapture studies which finds 90-95% of captured beetles to be within 400-500 m radius of the release site, and 99% of captured beetles to be within a 3.2 km radius (Turchin and Thoeny 1993). In addition to the 2500 metre distance band, I also employ other sizes of the local neighbourhood to complete the evaluation of the multi-region approach.

Because it is not possible to know *a priori* what the true partitions are for each year of the case study, I estimated the true regions in the study area (presuming there are regional differences) with an *ad hoc* comparison of iterative partitions of the study area using a Komolgorov-Smirnov test, which I call KS-partition. Appendix A provides further details on how this *ad hoc* method was conducted. The shortest side of the smallest region was 15 km and 22.5 km for 1996 and 2001 respectively. Therefore, the LR1 for 1996 has a local region with a radius of 7.5 km and for 2001 a radius of 11.25 km is used.

Given the difficulty of defining the exact local region, the focus for the case study is more on the multi-region approach than the local-region approach. It should be noted that the KS-partition method was initially employed to give a comparison against which to assess the LR G_i^* results. However as I furthered my understanding of the MR G_i^* method, I had much more confidence in the ability of the MR G_i^* to identify regions over the KS-partition. I do however include the results of the KS-partition method in Figure 18 as this gives a rough idea of the concordance between the region identifying capabilities of the KS-partition and MR G_i^*.

Based on the results of the KS-partition, divisions were made at 6005000mN, 6020000mN, 6080000mN, 6110000mN, and 6132500 mN for the 1996 data, and at 5952500mN, 5975000mN, and 6072500mN for the 2001 data (UTM NAD 83 Zone 10).

A Standard G_i^* was run on the entire study area and then using the LR and MR G_i^*. Methods were then compared and hypotheses formulated about why the methods perform differently for the years 1996 versus 2001.

RESULTS

Multi-Region Versus Standard

The combined total counts of low and hig *z*-scores <=-2 and >=2 account for a larger percentage of total *z*-scores for all methods for the 1996 data than the 2001 data (Table 6).

Figure 17 shows maps of the standard results for 1996 and 2001 and it is immediately apparent how in 1996 the high values in the north and low values in the south are causing a sharp dichotomy of high and low G_i^* *z*-scores in the study area.

Table 6. Case Study - percent of significant locations (>=2 and <=-2) Standard and MR Probability results.

	1996	**2001**
Standard	.381	.089
MR Probability (50%)	.268	.099

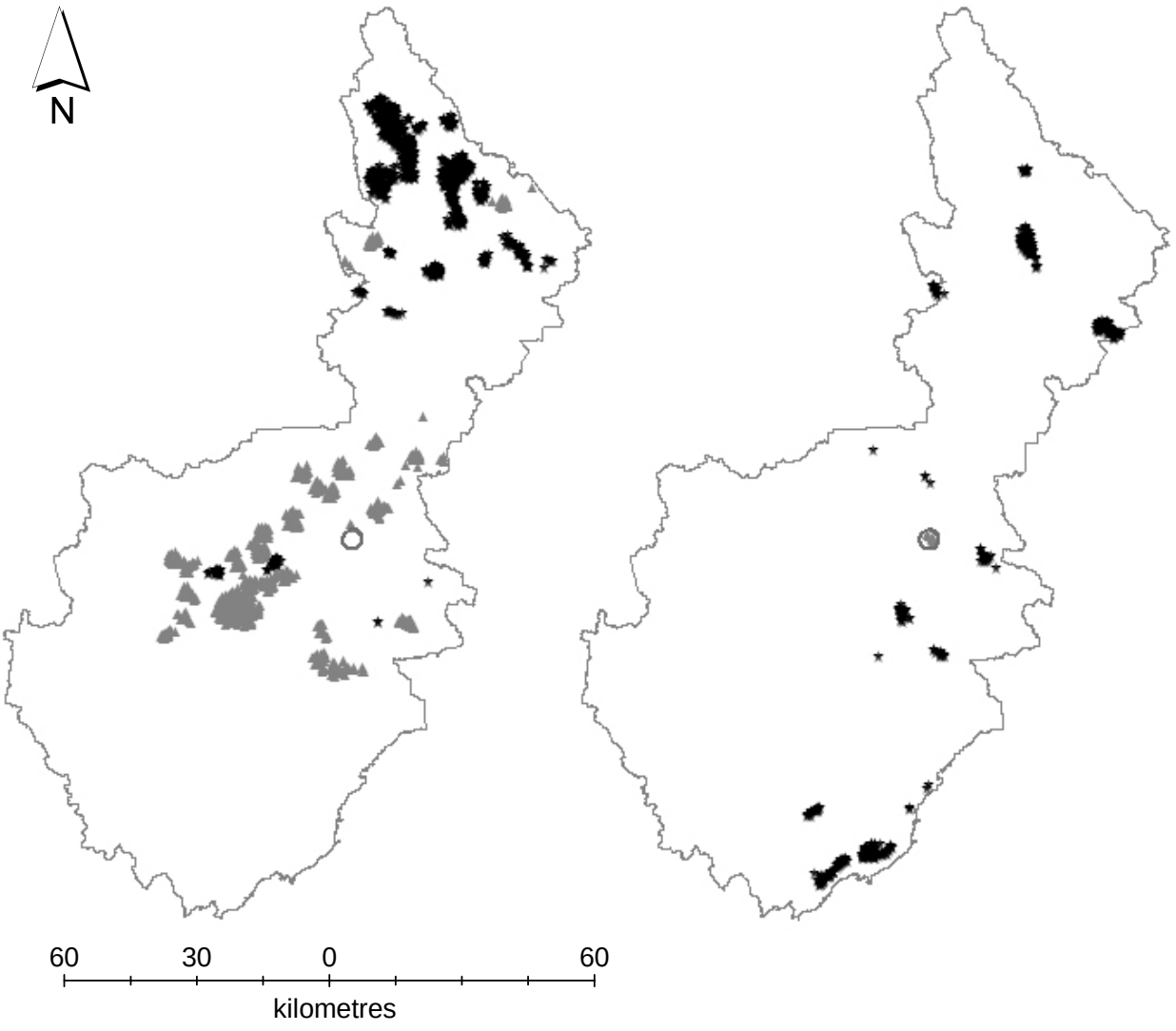


Figure 17 – Case Study – 1996 and 2001 – High and low G_i^* *z*-scores (>=2;<=-2) of the Standard G_i^* method, using a 2500 metre radius for the local neighbourhood. Black star: >=2; Grey Triangle: <=-2.

For 1996, overall the difference in the number of significant high *z*-scores (>=2) between the multi-region probability (50%) and the standard G_i^* is slight – 1145 versus 1105 points >=2. However, the dispersion of these hotspots over the study area is distinctly different for the two methods (Figure 17).

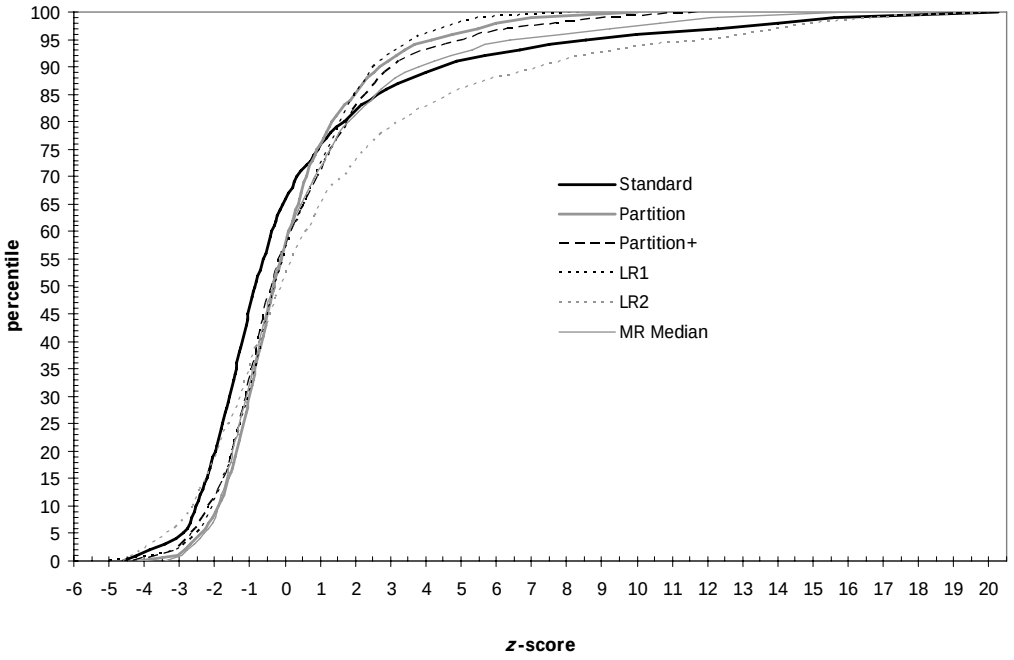
For 2001, the results set that has the greatest number of coincident significant high *z*-scores (>=2) with the Standard results is MR Probability (85% coincidence). The Standard results only had nine significant low *z*-scores (<=-2) (Table 7).

Table 7 – Case Study - coincident *z*-scores >=2 and <=-2. Total counts of *z*-scores >=2 or <=-2 are in round brackets beside the name of the results set; the percentage of *z*-scores >=2 or <=-2 out of the total number of points (1000) is in square brackets next to the actual counts.

1996	**>=2**		**<=-2**	
	Standard (1105)	Partition (902)	Standard (1207)	Partition (520)
Standard	-	641/1105	-	105/1207
MR Probability	840/1145	560/1145	317/485	164/485
2001	**>=2**		**<=-2**	
	(481)	(554)	(9)	(69)
Standard	-	326/481	-	9/9
MR Probability	458/539	400/539	9/9	9/9

From the cumulative distribution graphs one can see that there is a much greater spread between the results for 1996 than there is for 2001 (Figure 18).

1996



2001

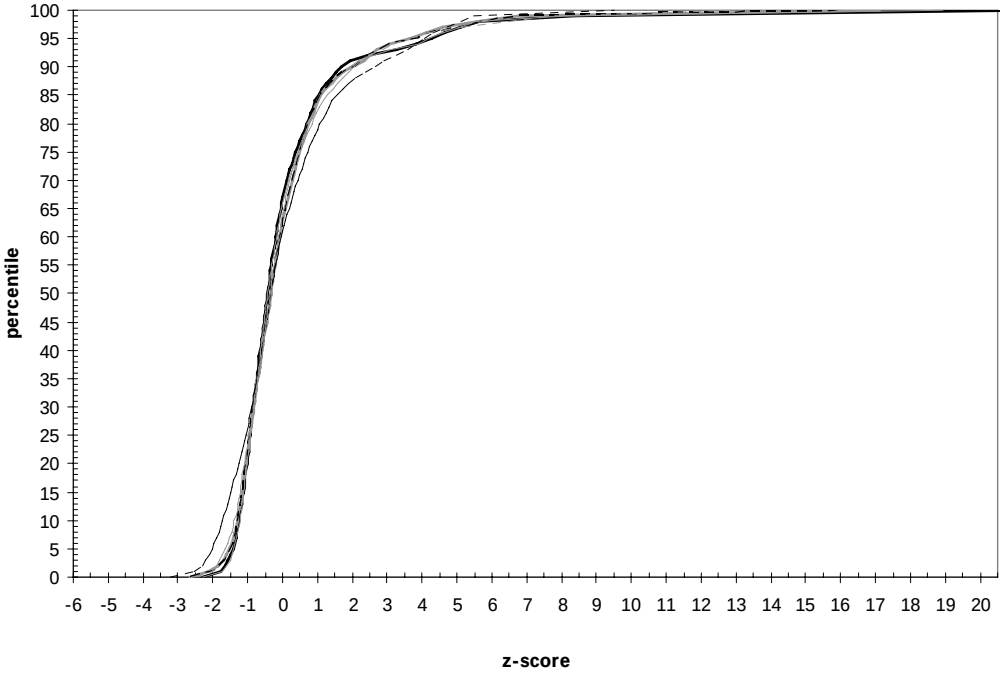


Figure 18 Case Study – 1996 and 2001– Cumulative frequency graphs for the results of the various G_i^* methods.

DISCUSSION

The results for the 1996 data are achieving results similar to what was seen for the non-stationary simulated datasets. On the other hand, the results for the 2001 data are more similar to the results of the stationary simulated datasets than the non-stationary datasets. The results suggest to us that the 2001 dataset has greater stationarity than the 1996 data.

Because of the way spatial association/autocorrelation by definition manifests itself in a way such that near things are more similar to each other than things farther apart (Tobler 1970) one can expect that those points in the centre of area of high or low value clusters will have higher probability scores than those at the edge of these clusters (assuming isotropic processes, that is, the cluster (or cluster of clusters) is equal in size in all directions).

PART B

In Part B, I extend the evaluation of the MR G_i* to consider other important topics including: 1) detecting transitional areas, 2) determining local neighbourhood and local region sizes, and 3) conducting randomization tests to determine significance envelopes for the methods.

DETERMINING A LOCAL NEIGHBOURHOOD AND LOCAL REGION

For modeling, a local measure is sometimes used to create a weights matrix for incorporating spatial autocorrelation (association) into an environmental model. It is in this section that I further the approach discussed earlier in this thesis for determining an

appropriate size for a local neighbourhood and region in the case of modeling spatial autocorrelation.

1996

As discussed for the simulation study, under a CSR process one would expect approximately the upper and lower tails of the distribution (>=2; <=-2) to be ~0.025 for each respectively. Therefore in selecting a local neighbourhood size I aim to select a size that will encompass a homogeneous area within which values are within two standard normal deviations. I determine this local neighbourhood size by observing the percentage of high and low scores (>=2;-<=2) for combinations of a local neighbourhood and a local region. If the local neighbourhood and local region are capturing similar values the percent of high and low scores will be something like what I see for a CSR process, thus, approximately 5% of the total scores will be high or low. If this is the case, then I can rationalize the local neighborhood can be increased to the size of the local region. This approach works best if a small local neighbourhood is used (at least to start). For instance, the initial selection of a 2500 metre local neighbourhood results in a combined percentage of high and low scores that exceed 20% when used with the smallest local region size of 5000 metres. I therefore conclude that at 2500 and 5000 metres there is already a considerable difference across the study area between the values of the local neighbourhood and the local region. To determine an appropriate local neighbourhood size it is necessary to use a smaller initial local neighbourhood for the MR G_i^*. In the case of the 1996 mountain pine beetle data, I reset the local neighbourhood to 500 metres and using a local region of 1000 metres achieve combined high and low

results G_i^* *z*-scores of 5.2%. Local regions beyond 1000 metres have a higher percentage of high and low G_i^* *z*-scores (Figure 19). To go lower than 1000 metres for the local neighbourhood would mean excluding a larger number of points that would be isolated from the rest of the dataset at this distance (can not calculate statistic if there are too few points in the local neighbourhood). The local neighbourhood was set at 1000 metres for this reason and because 5.2% is approximately what would be expected under complete randomization.

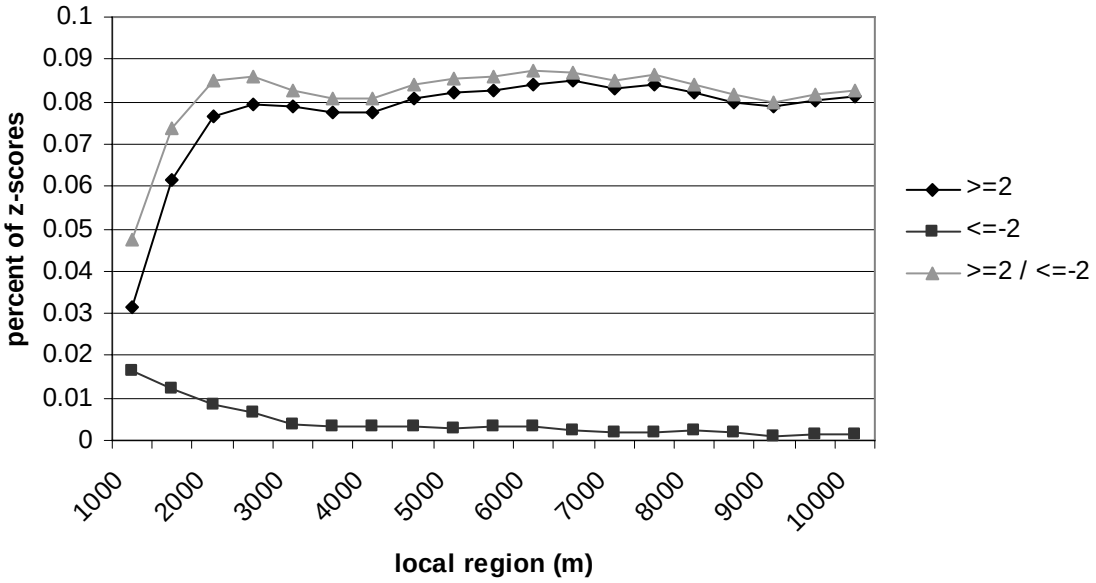
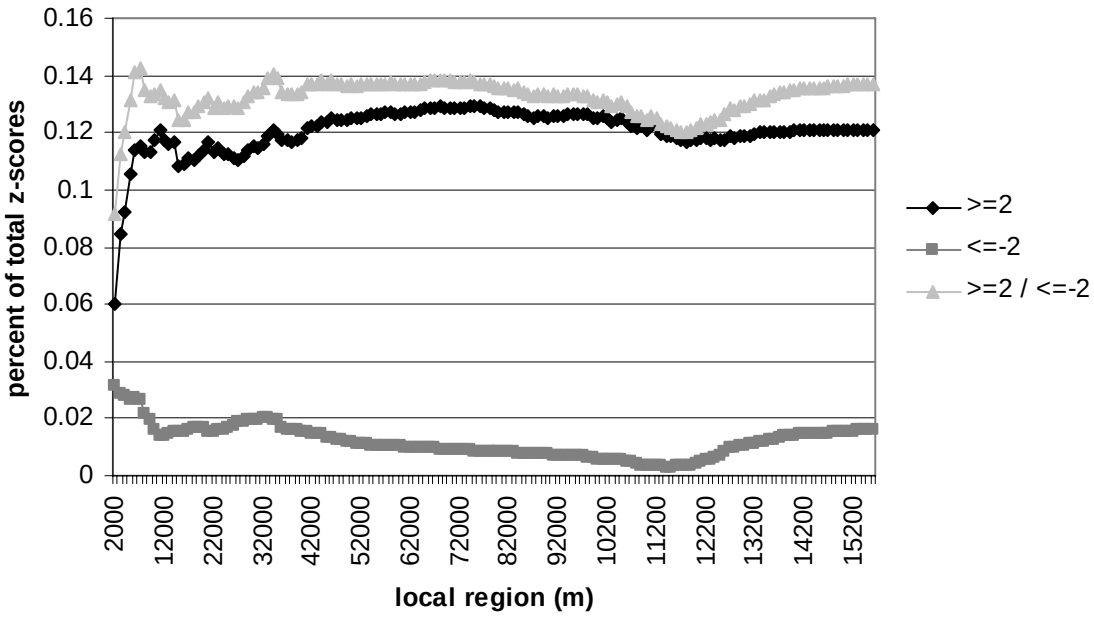


Figure 19 – Case Study – 1996 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 500 metres.

Next, I ran the MR G_i^* with the selected local neighbourhood of 1000 metres with the goal of now determining an appropriate local region. From the graph of the resulting percentage of high and low scores across the multiple regions (Figure 20), I determined an appropriate local region to be at 6000 metres, corresponding roughly with the first peak in high scores and pit of low scores. Local regions beyond 6000 metres show a

much less marked change in high and low scores (Figure 20). I choose the local region to account for the most intense areas of change. To go beyond this size for the local region would be to risk incorporating the effects of non-stationarity into the model.

Figure 20 – Case Study – 1996 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 1000 metres.



2001

Similar to the 1996 dataset, the initial local neighbourhood selection of 2500 metres results in high and low MR G_i^* *z*-scores that exceed 20% even when coupled with the smallest local region of 5000 metres. I therefore recalculate the MR G_i^*, using a local neighbourhood of 500 metres and find the combined percentage of high and low scores for a local region of 1000 metres is 4.3%, slightly lower than expected. With a local neighbourhood of 500 metres and a local region of 1500 metres the total high and low scores increases to 6.4%, slightly higher than expected. I would thus expect the most appropriate local neighbourhood to lie between these two local regions, 1000 metres and 1500 metres (Figure 21). For the sake of comparison with the 1996 dataset, I decided to use the region that achieves more conservative results, 1000 metres, rather than estimating a new local region distance in between the two regions, although keeping in mind the purpose of this approach is to obtain the largest local neighbourhood possible without exceeding a reasonable level of heterogeneity/homogeneity inside the neighbourhoods across all locations (no greater than ~5% of combined high and low scores).

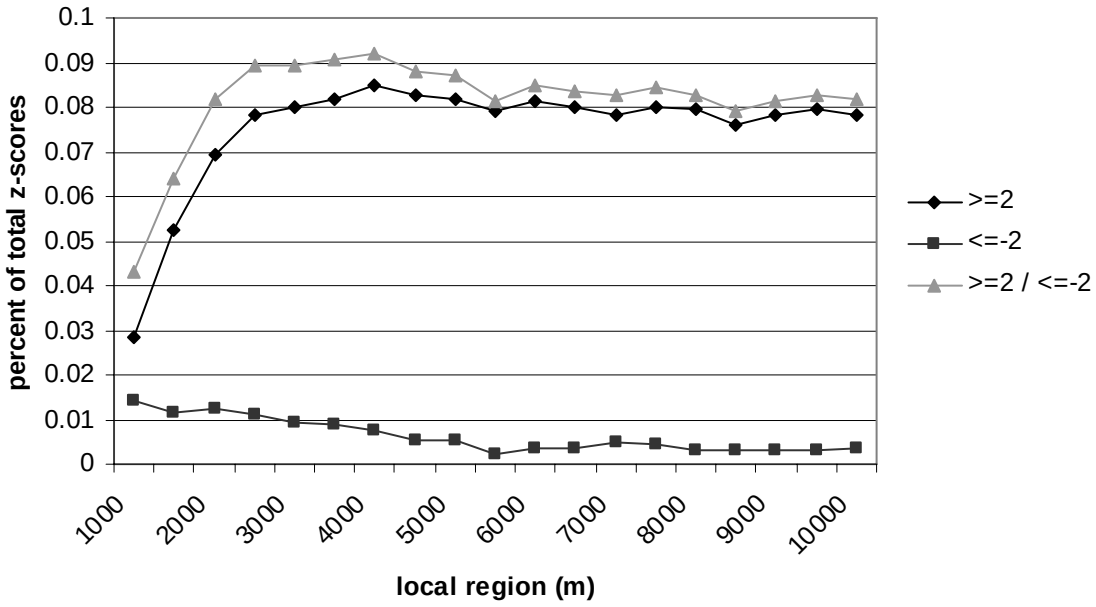


Figure 21 - Case Study – 2001 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 500 metres.

I ran the MR G_i^* using a local neighbourhood of 1000 metres. From interpretation of the graph of the MR G_i^* results (Figure 22), I determined the appropriate local region to be at 6000 metres roughly at the first and highest peak of the high G_i^* *z*-scores and first levelling off of the low G_i^* *z*-scores. After this first peak, the high scores decrease until at ~25000 metres the scores level off and then fluctuate above and below ~7% of total high scores for the remaining increments of the local region.

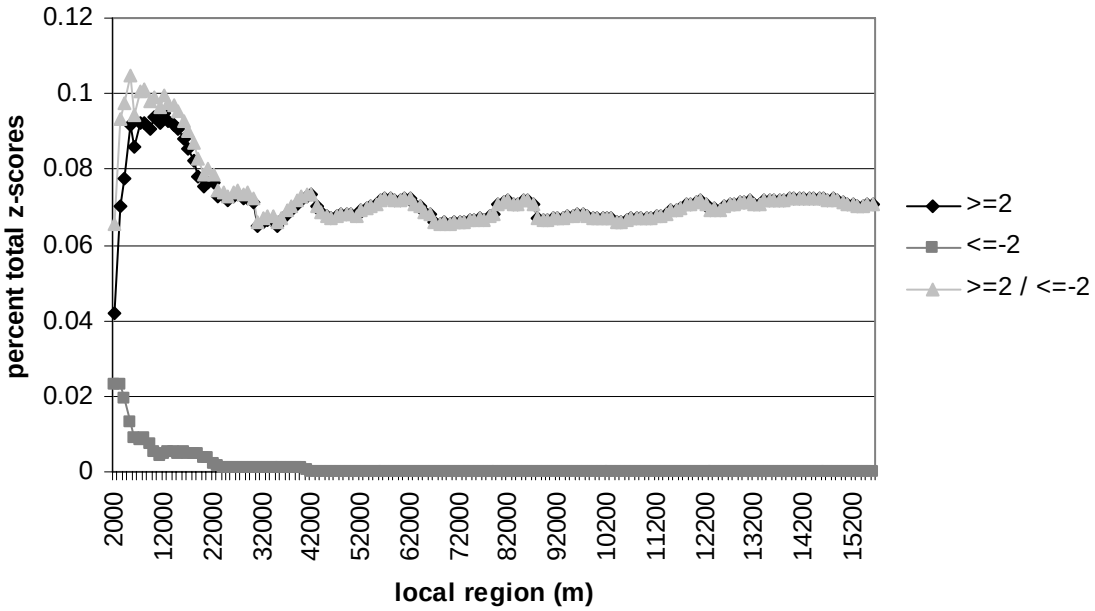
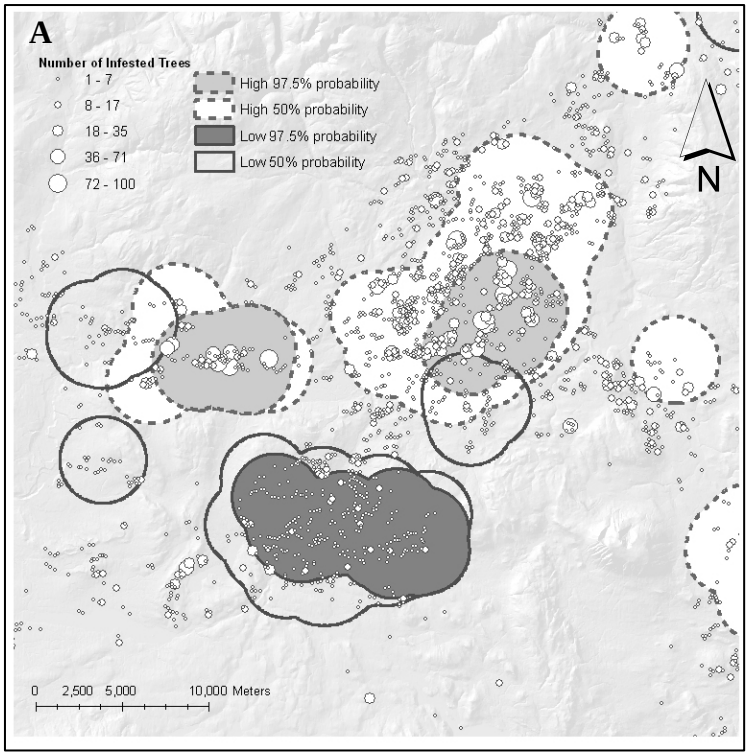


Figure 22 - Case Study – 2001 – graph of MR G_i^* probability results for each local region, using a local neighbourhood of 1000 metres.

In Figures 23 and 24 I map the two permutations of the MR G_i^* discussed above, for each year of the dataset respectively. The two permutations are the MR Probability G_i^* using a local neighbourhood of 2500 metres and the MR Probability G_i^* using the local

neighbourhood of 1000 metres as determined from the graphs of the results across the multiple regions of a local region. I map high and low results by buffering the respective high and low points by the size of the local neighbourhood used and then dissolve so that overlapping high or low points become one polygon. This makes visualization of "hotspots" and "coldspots" considerably easier than when they are left as overlapping points.



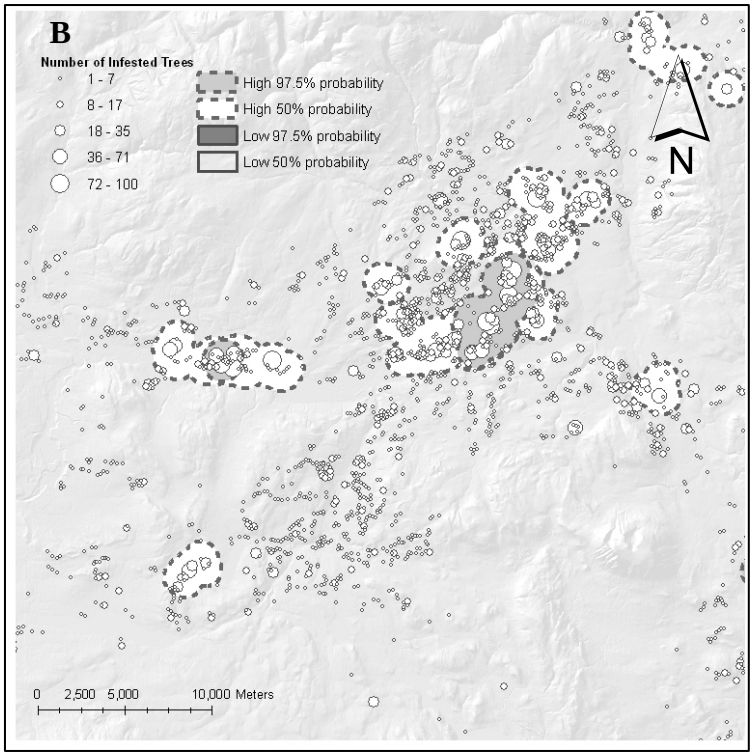
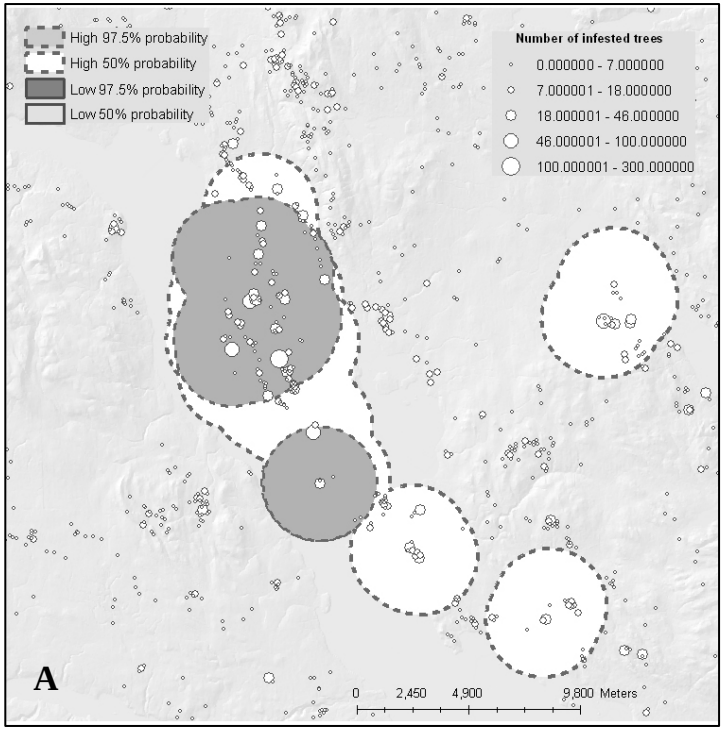


Figure 23 Case Study - 1996 – Comparison of mapped results of methods. A. high and low results of MR G_i^* Probability using a local neighbourhood of 2500 metres (chosen to capture between 8 and 30 points and as representative of average dispersal distance of the mountain pine beetle) B. high and low results of MR G_i^* Probability using a local neighbourhood of 1000 metres (as determined from evaluation of peaks and pits and percentages of high and low scores from the graph of results of the MR G_i^*)



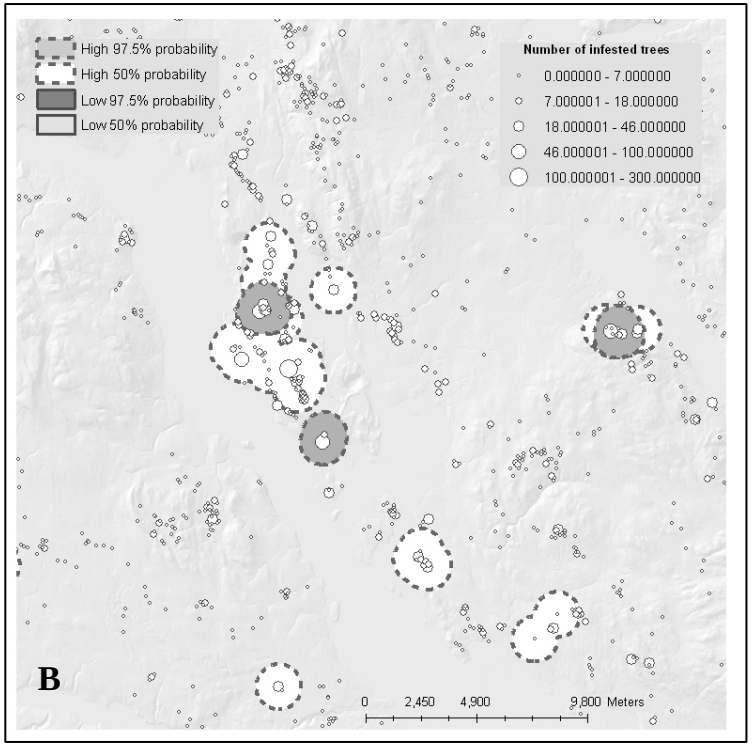


Figure 24 Case Study – 2001 – Comparison of mapped results of methods. A. high and low results of MR G_i^* Probability using a local neighbourhood of 2500 metres (chosen to capture between 8 and 30 points and as representative of average dispersal distance of the mountain pine beetle) B. high and low results of MR G_i^* Probability using a local neighbourhood of 1000 metres (as determined from evaluation of peaks and pits and percentages of high and low scores from the graph of results of the MR G_i^*) C. high and low results of LR G_i^* using a local neighbourhood of 1000 metres and a local region of 6000 metres (as determined from graph of MR G_i^* scores).

TRANSITIONAL AREAS

In this section, I demonstrate how the multi-region approach can be used to identify transitional areas in the dataset. For the 1996 data I plot the percent of high and low G_i^* *z*-scores for each increment of the local region, so that I can see how these change as I increase the increment size. I can identify important transitional areas by identifying the peaks and pits in the graph of MR G_i^* results (Figure 25). I can then observe the mapped high and low *z*-scores on the map and begin to interpret why the major shifts may be occurring (Figure 26). Figure 26 displays the maps for important peaks and pits in the MR G_i^* Probability results.

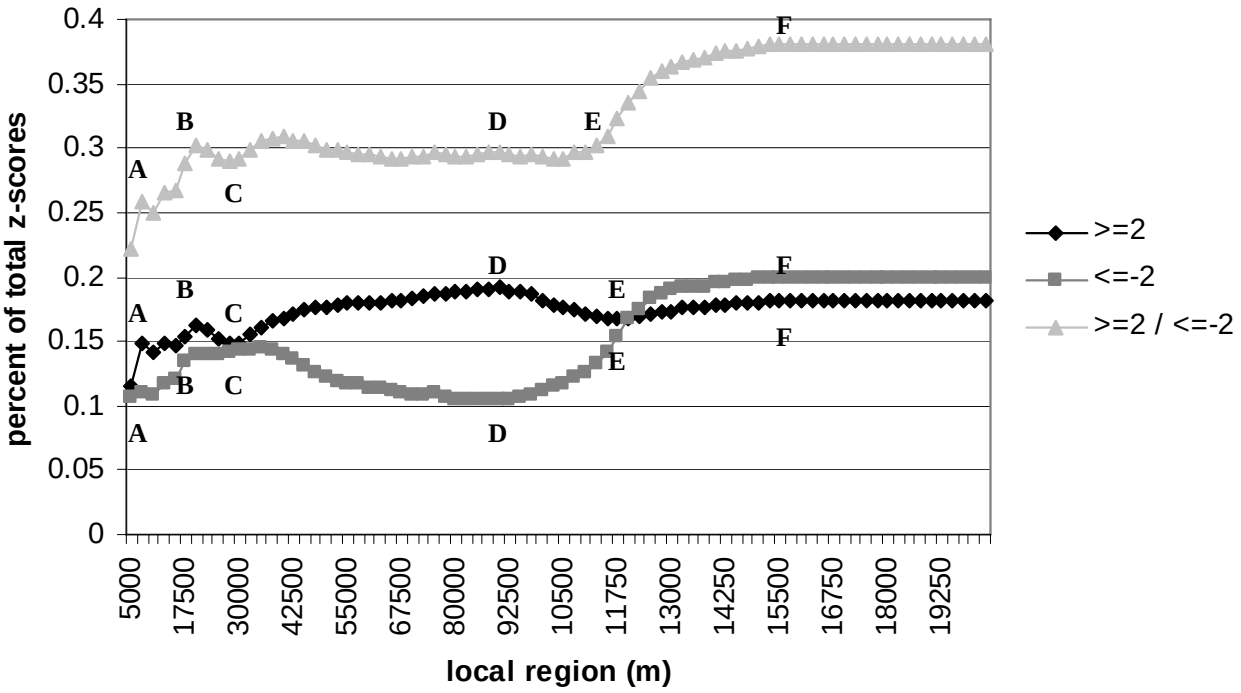


Figure 25 – Case Study – 1996 – graph of MR G_i^* probability results with major pits and peaks labeled and associated with mapped results in Figure 24. Letters identify important shifts in slope the lines of the graph. Letters are as noted in text.

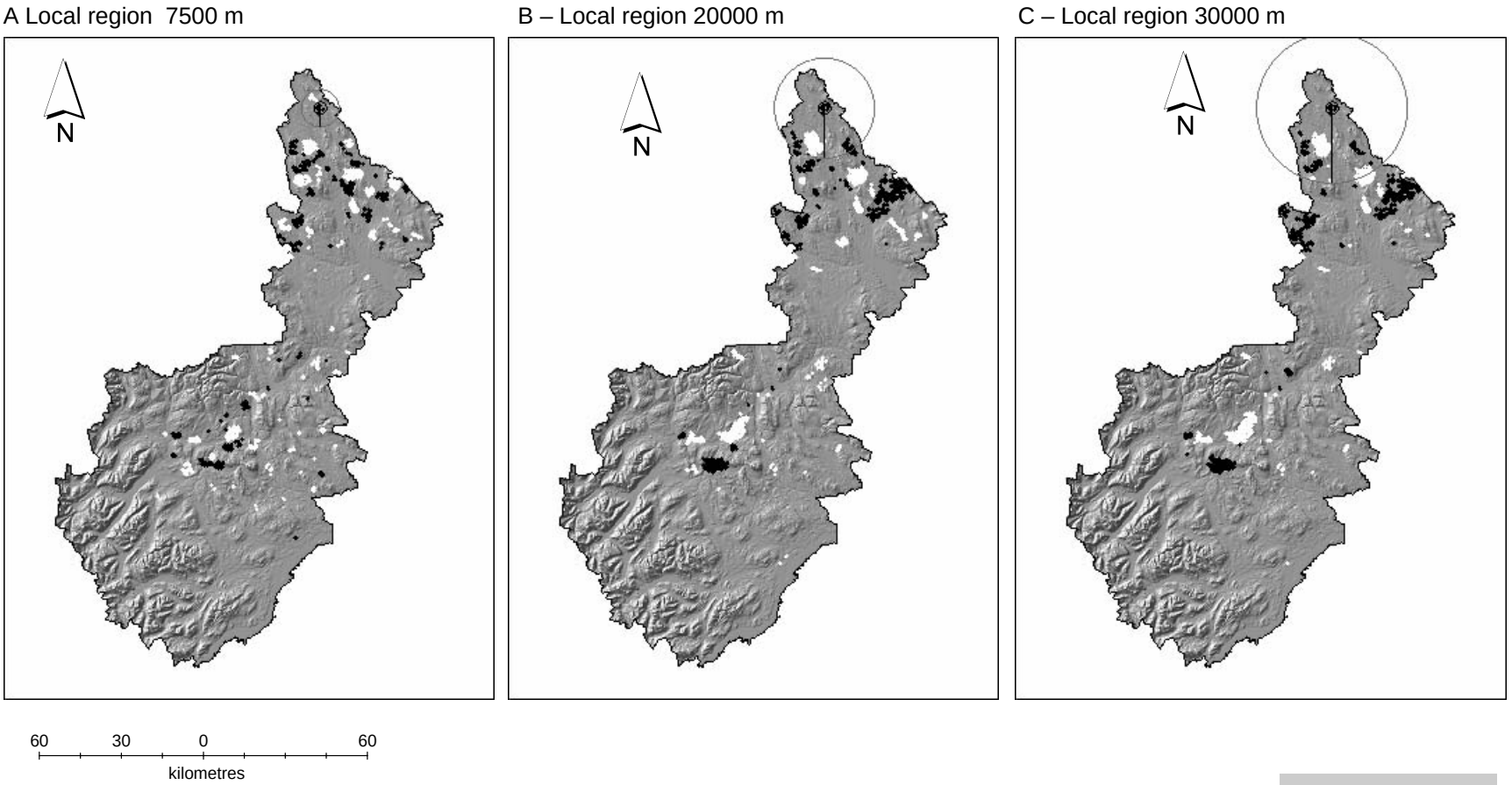


Figure 26 – Case Study – 1996 – maps of MR G_i^* Probability results for various local regions identified as peaks and pits for the graph of the results in Figure 26. The local neighbourhood is 2500 metres. The black symbols are used to represent points with high positive spatial association of low values (G_i^* *z*-score <=-2) aand the white symbols are used to represent high positive spatial association of high values (G_i^* *z*-score >= 2)

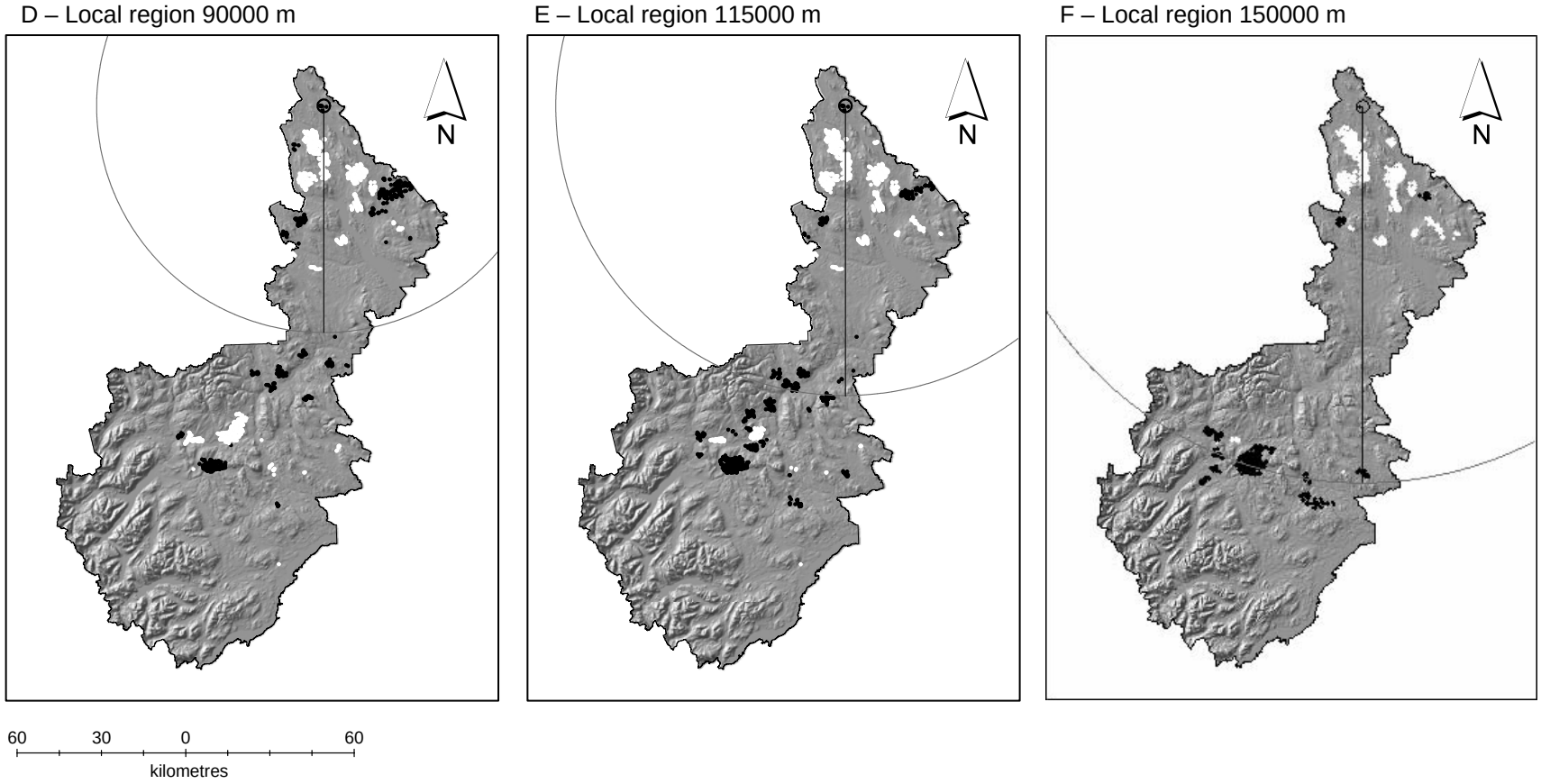


Figure 26 continued. Case Study – 1996 – maps of MR G_i^* Probability results for various local regions identified as peaks and pits for the graph of the results in Figure 26. The local neighbourhood is 2500 metres. The black symbols are used to represent points with high positive spatial association of low values (G_i^* *z*-score <=-2) and the white symbols are used to represent high positive spatial association of high values (G_i^* *z*-score >= 2)

Identifying transitional areas can be done more precisely on a point-by-point basis. As displayed in Figure 27 , I can graph the G_i^* *z*-scores for a single location for each of the increments of the local region, identify where there are unusual shifts in the *z*-scores and plot these increments relative to the point in question on a map. In Figure 27, the dashed line represents unique regions identified by my *ad hoc* KS-partition method.

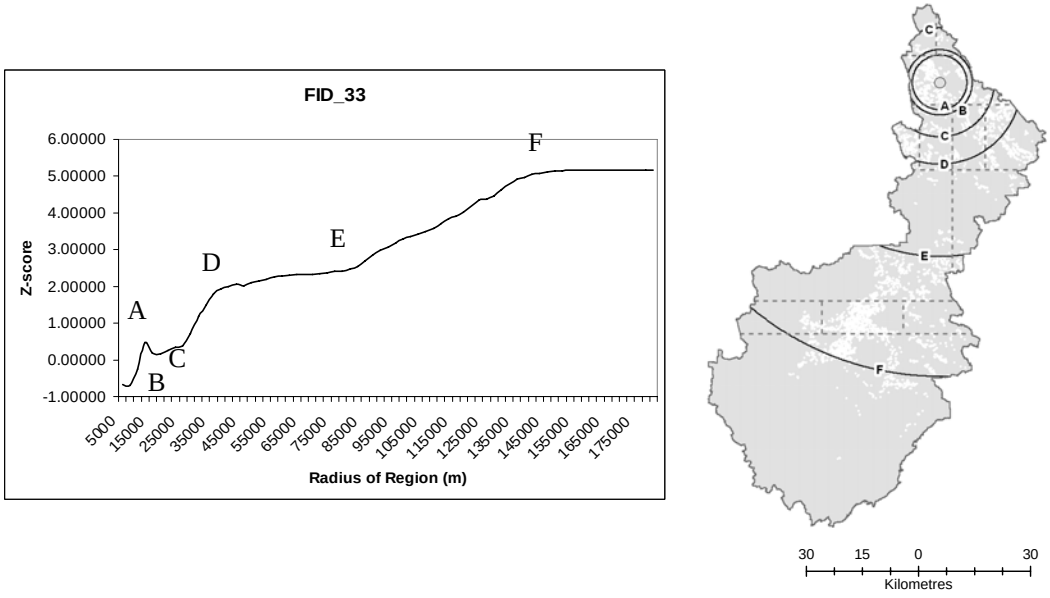


Figure 27 – Case Study – 1996 – Detecting transitional areas on a point-by-point basis. This figure shows two examples of using the MR G_i^* to detect transitional areas. Orthogonal dashed lines represent areas demarcated as separate regions using the KS-partition method (see Appendix A). The local neighbourhood is 2500 metres.

RANDOMIZATION

To investigate statistical significance for the results of the methods I explored two randomization procedures. The first involves keeping the spatial location of points in the study area static while randomizing the associated marks *n* times. The second involves randomizing spatial locations of points and randomizing associated attribute values *n* times.

To evaluate the appropriateness of the high and low cutoffs (>=2; <=-2) of the LR G_i^* I ran a randomization procedure to determine significance envelopes. The marks associated with points were randomly shuffled and the LR G_i^* recalculated for 100 randomizations. Spatial locations remained static through out the procedure. A local neighbourhood of 1000 metres and a local region of 6000 metres were used. The procedure was completed for the 1996 and the 2001 datasets.

The purpose of randomizing marks is to isolate the effect of spatial clustering by location from spatial clustering by mark. By reshuffling the mark values among the point locations 100 times, I am able to identify areas where there is a high probability under the current position of points and conditions determining the presence or absence of other points that certain location will have high or low scores regardless of their associated marks. Significance envelopes were determined for each point location by calculating the 2.5 and 97.5 percentiles across the 100 randomization results for that location. I then could determine statistical significance of the results by observing at which locations the

real LR G_i^* *z*-score falls outside of these significance envelopes. The results show that the high and low cutoffs (>=2;<=-2) are producing lower percentages of high and low scores than would be the case if I used the randomization significance envelopes. For the 1996 dataset, the majority of high and low real LR G_i^* *z*-scores (>=2;<=-2) fall beyond the significance envelopes of the randomization procedure. Of the high scores identified by the high cutoff (>=2) for the real LR G_i^* method, 82.11% fall outside of the significance envelopes and 17.89% fall inside. For the low scores (<=-2) only 3.66% fall within the significance envelopes, the remaining 96.34% fall outside. An advantage to using randomization-based significance envelopes is that the influence of boundaries will be reflected in the significance envelopes and therefore edge correction solutions are unnecessary

For the spatial randomization, I conditioned the randomizations to occur spatially only where there were no lakes, marshes, roads, or non-forested areas, and at elevations lower than 1200 metres. Figure 28 is a map of the results for 1 of the 100 spatial randomizations. For *n* I used 100 randomizations. More randomizations would be ideal, but computationally expensive and for this initial evaluation I felt that going beyond 100 was unnecessary. The attribute values used were the same for the original 1996 dataset. For this evaluation I used just one increment of the multi-region. Ideally this would be done for all or several sizes of the local-region. The local neighbourhood used was 2500 metres and the local region 5000 metres. Additionally, to cut down on processing time, the randomization procedures were conducted for the northern region only.

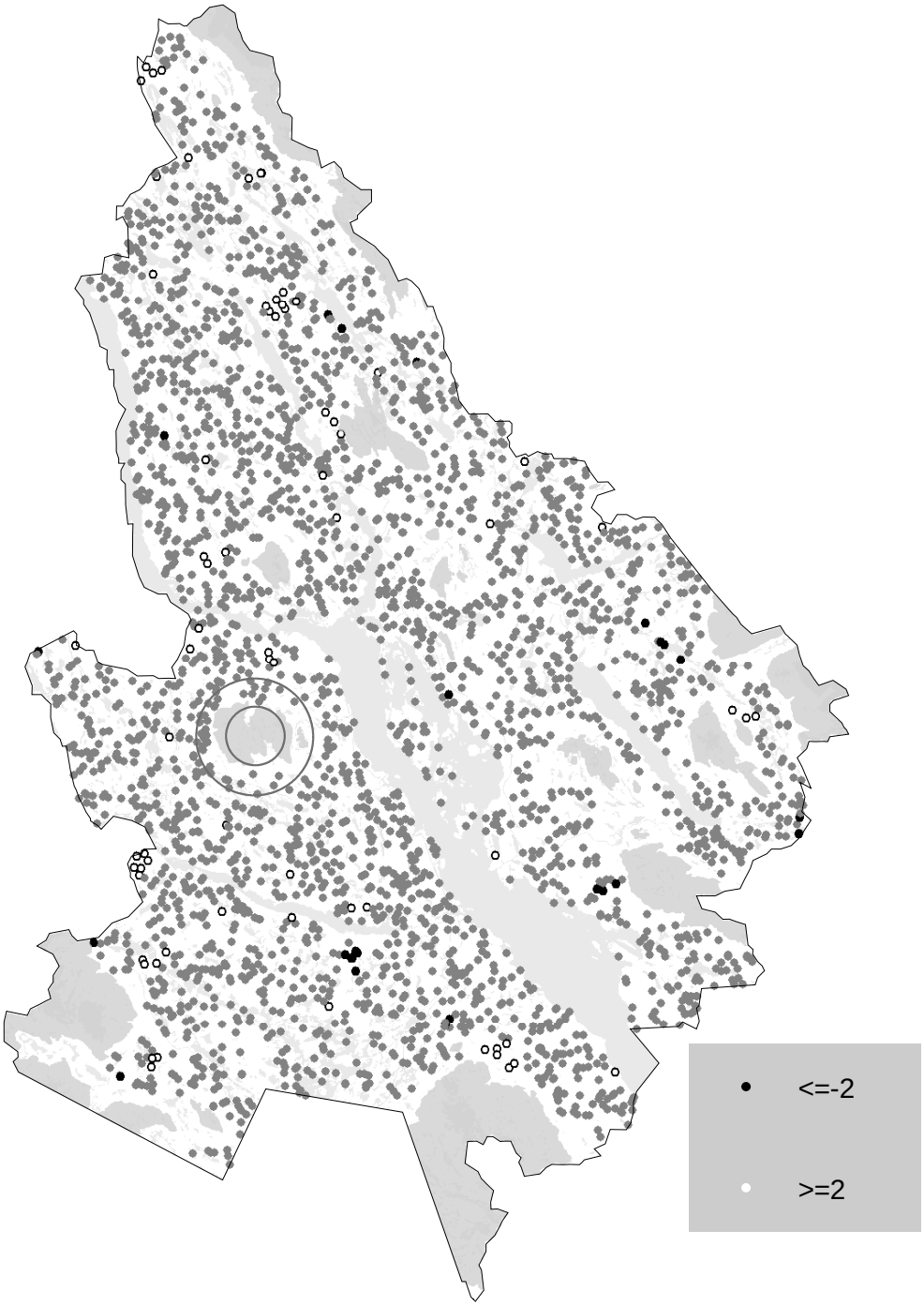


Figure 28 – Case Study – 1996 North – spatial randomization example. This figure shows an example of one randomization of random points generated, including calculated G_i^* *z*-scores (white: >=2, black: <=-2). Grey shaded areas are the lakes, non-forested areas, high elevations, and other areas conditionally selected out of the randomization space.

Figure 29 – Case Study – 1996 North – spatial randomization; results of 100 randomizations. Dark cells represent locations where over 100 simulations where results were on average for that cell >=2 15-30% of the time (grey), or over 30% of the time (dark grey). The size of one side of a cell is 1000 metres, less than half the local neighbourhood size of 2500 metres.

This randomization procedure involves randomizing the location of points (within the set conditions, see above) and subsequently also the location of associated marks. The purpose of this procedure is to determine if the conditions are somehow influencing the perceived clustering of points and also given the non-normal distribution of attribute values in the dataset, does this skew the high and low thresholds that I have set at >=2 and <=-2. I follow this procedure to create a reference set (a null hypothesis) of complete spatial and mark randomness. Because the spatial location is changing for each of *n* randomizations, I can not map the results as points. One solution is to create a mesh and calculate the average G_i^* *z*-scores for each randomization and across all randomizations. This is demonstrated in Figure 29.

Note that the mean values for high and low scores are approximately what is expected under CSR and therefore I can conclude that neither the conditions (lakes, mountains, etc.), the non-normal distribution of attribute values, nor the method itself are causing non-true results (at least at the summary level) (Table 8). However, there does appear to be a greater likelihood of receiving a high score than a low score.

TABLE 8 – Case Study – 1996 North - Summary statistics of high and low score probabilities (>=2;<=-2) for 100 randomization of the spatial location of data points.

	Mean	Min	Max	Standard Deviation
>=2	0.031	0.022	0.042	0.005
<=-2	0.009	0.004	0.017	0.003

Spatial querying the results of the randomization (1000 metre mesh) and the local region method using 2500 metre neighbourhood and 5000 metre local region

indicates no tendency for high G_i^* *z*-scores to land where the randomization procedure shows high probabilities for the occurrence of a high score, or for low G_i^* *z*-scores to land where the randomization procedure shows high probabilities for the occurrence of a low score. In fact there is only one location with a low score (<=-2) that is coincident with a mesh cell that has a probability of >=10 percent of receiving a G_i^* *z*-score <=-2 (where 101 cells of the mesh have probabilities greater or equal to 10 percent for being low - <=-2), and only 37 locations with a high G_i^* *z*-score (>=2) that are coincident with mesh cells that have a probability >=10 percent (where 361 cells of the mesh have probabilities great or equal to 10 percent for being high). Of high or low G_i^* *z*-scores, 136 are coincident with mesh cells that have either probabilities >= 10 percent of being high or low. However, if I look at the probability of being high or low together and observe where these extreme values are landing there are indeed some areas that are consistently landing in areas identified by the null hypothesis randomization as having a higher probability (than the surrounding areas) of receiving a high or low G_i^* *z*-scores. Twelve point locations of the LR G_i^* results are coincident with mesh cells with probabilities greater or equal to .3 of being either high or low, of which there are 39.

DISCUSSION

The multi-region approach provides a way in which to define appropriate neighbourhoods and local regions. With the local neighbourhood I seek to capture the largest neighbourhood size while still ensuring that for the majority of locations (~95%) there is no significant difference between the local neighbourhood and a sub-local neighbourhood of a radius half the size. I select a local region to capture the most rapid changes in significant differences between the local neighbourhood and local region while guarding against including values from adjacent regions where the spatial process is markedly different. Although the demonstration here uses the summary values of total high and low scores for a set local region, it would also be possible to apply this approach on a point by point basis, which would likely give more precise delineations of local neighbourhoods and regions.

A LR or MR G_i^* advantage of calculating level of infestation based on immediate surrounding which may include one more unidentified (or unquantifiable) variables that make susceptibility different in that particular region of the landscape. In this regard, taking the total tree count and comparing to average intensity across the study can give erroneous results because in some areas high susceptibility does not result into the same number of infested trees as high susceptibility would in another area. A simple example is that the percentage of pine in an area such as defined by biogeoclimatic zones will change across the landscape; in the valley bottoms pine may dominate a forest and account for upwards of 100% of the trees species in a stand whereas at higher elevations such as the Englemann-Spruce-SubAlpine Fir (ESSF) biogeoclimatic zone pine is interspersed with other species namely Subalpine Fir and Englemann spruce.

Susceptibility for pines in these stands may be of equivalent probability to homogeneous pine stands however will not reflect as intense an infestation because of the lower intensity of pine in these areas. One way to account for this would be to conduct analysis one biogeoclimatic zone at a time, however there are likely cases where I do not know what variable to partition a landscape on. Based on Tobler's law, and my general understanding of spatial autocorrelation in the natural environment and more specifically for tree plant species, it is to be expected that nearby areas will have higher probability of being in similar environments.

Transitional areas can be detected by summarizing the results across all points for each local region of the MR G_i^* into percent of high and low scores for that region and then by comparing the changing percentages of high and low scores as the local regions changes. Alternatively, transitions can be more precisely defined by evaluating changes in G_i^* z-scores for different sizes of the local region for individual points. Either way, both methods demonstrate how a multi-region approach can be used to detect regions and transition areas between regions. By buffering and merging the high MR G_i^* Probability scores, I overcome some of the shortcomings of visualizing the intensity of point datasets, such as is the case with proportional symbols where the visualization is obscured by symbol overlap and thus prevents meaningful interpretation (Nelson et al 2006). Representing high MR G_i^* probability scores (high for positive spatial autocorrelation of high or low values) as polygons also lends the results to further analysis potential regarding spread and dispersal by using spatial-temporal methods such as STAMP (Robertson et al 2007). By creating polygons from the points with high and low MR Probability scores, I prepare the dataset for further research that will tie into

underlying ecological variables. For instance, I can use these polygons to evaluate if local scale differences in underlying variables such as homogeneity/heterogeneity of forest stand composition due to forest management have any relationship to the location (spatially and temporally) of unusual clusters (*z*-scores >=2 or <=-2) of mountain pine beetle infested trees.

From the randomization tests, I can conclude that at the chosen scale the majority of high and low G_i^* *z*-scores in the study area are not the results of forced clustering by conditions such as lakes, non-forested areas, etc. (although may be influenced by unaccounted variables such as species composition in forest stands, or more directly, the presence of pine in a stand). The spatial location of points, more specifically their clustering, does not appear to be the primary factor influencing the location of high and low G_i^* *z*-scores (as observed in the randomization of marks only).

CHAPTER 6 – CONCLUSION AND FUTURE DIRECTIONS

The LR G_i^* method set for a local region of a size similar to that of a true region, calculates results similar to those of a standard G_i^* on partitions of the study area. When the spatial process is stationary across the entire study area, a LR G_i^* with a local region of any size will receive results similar to that of a standard G_i^* for the entire study area. The LR G_i^* is advantageous over a partition approach because it can be applied with no knowledge of the borders of the true partitions. However, it does not answer the question, what is the true partition? I believe this question can be answered in part by systematically running the LR G_i^* at different sizes of the local region (as is done for the MR G_i^*) and observe at what scale the distribution of *z*-scores and counts of high and low *z*-scores change (this can be done by observing on a graph at what scales individual points have a marked change in the calculated *z*-score), however this is the focus of further research.

The MR G_i^* allows for a detailed dissection of the nature of spatial association in the study area through the observation of values for individual record over changing sizes of a local region. The MR G_i^* does not require exact knowledge of the size of a true region, but potentially could be used to identify the size of regions where they exist and the scale of the spatial process governing the pattern of a spatial dataset. Again, this is the focus of ongoing research. The MR G_i^* captures unusual spatial associations for locations at both large and small scales. It is able to identify unusual local spatial associations in a study area that the standard method misses and does so by dampening the effects of regions with extreme values for the global parameters. The MR G_i^* also has practical appeal. A

user can visually assess the potential effect of regional differences of mean in a study area by mapping summary statistics of the MR G_i^* (median or probability), or query the results for any particular distance of region around all points to determine if there is anything unusual about certain locations in the study area when compared to a set surrounding area.

In summary, the important difference between the LR and MR method is that the LR considers an area, local region as set by the user, immediately around the local neighbourhood while the summary value of the MR method considers the whole study area. The MR method allows high and low regions to have a strong influence on the results of other regions, but allows for small variation that is absent in standard method. The LR method on the other hand excludes areas beyond a certain distance that a user might consider unimportant to influencing that local grouping of locations. Whether one method is better than the other is perhaps a poor question, as each method fills a different requirement. My first recommendation for using these methods is to determine whether regional differences in the mean are a problem for the dataset being analyzed. A sure sign that there are regional differences in the mean in the data for a study area is a high count of high *z*-scores (>=2) and / or low *z*-scores (<=-2) when using the standard G_i^*. If this is the case, the next step would be to determine the location of regional boundaries, first by mapping and observing the attribute values (in case those regions should be very obvious; of course, this could also be done before running any test), and second by graphing the *z*-scores for individual locations across multiple scales of the LR G_i^* to identify where major shifts in global values are occurring. If the need of the researcher is to identify

unusual values of points at a local scale, then using a local region that is of similar size or smaller than the largest true region for the LR G_i^* will be more appropriate than the standard G_i^* or a summary of the multi-region G_i^* (e.g., median) and will result in fewer overall high and low *z*-scores than either of these methods. It would be important when reporting the results to include the size of the local region (and local neighbourhood) used.

The results of these methods work best when examined relative to one another rather than alone, and it is through the comparison of the results of the different methods that the identity of the underlying spatial process will begin to emerge. A final note is to emphasize that the methods introduced in this paper are intended primarily for exploratory analysis and hypothesis formulation.

References

Aldstadt J, Getis A (2006) Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. Geographical Analysis, 38, 327-343.

Anselin L (1995) Local indicators of spatial association-LISA. Geographical Analysis 27(2),93-115.

Boots B (2002) Local measures of spatial association. Ecoscience 9(2), 168-176.

Burra T (2002) Conceptual and practical issues in the detection of local disease clusters: a study of mortality in Hamilton, Ontario. The Canadian Geographer, June.

Cliff AD, Ord JK (1973) Spatial autocorrelation. Pion, London.

Burrough PA (1987) Spatial Aspects of Ecological Data. In: Jongman RHG, ter Braak CJF and van Tongeren OFR (eds) Community and Landscape Ecology. Netherlands: Pudoc Wageningen, 213-251.

Carroll AL, Taylor SW, Régnière J, Safranyik L (2004). Effects of climate change on range expansion by the mountain pine beetle in British Columbia, in T.L. Shore, J.E. Brooks, J.E. Stone (Eds.), Mountain Pine Beetle Symposium: Challenges and Solutions, Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, BC, Information Report BCX-399, 298.

D'Agostino RB (1971) An omnibus test of normality for moderate and large size samples. Biometrika, 58(2), 341-348.

Davis JH, Howe RW, Davis GJ (2000) A multi-scale spatial analysis method for point data. Landscape Ecology, 15, 99-114.

Diggle, P (1983) Statistical Analysis of Spatial Point Patterns. London, Academic Press Inc.

Dymond CC, Wulder MA, Shore TL, Nelson T, Boots B, Riel BG (2006). Evaluation of Risk Assessment of Mountain Pine Beetle Infestations. Western Journal of Applied Forestry, 21(1), 5-13.

Eck JE, Spencer C, Cameron JG, Leitner M, Wilson RE (2005). Mapping Crime: Understanding Hot Spots. U.S. Department of Justice Office of Justice Programs National Institute of Justice.

Fortin M–J, Dale M (2005) Spatial Analysis A Guide for Ecologists. Cambridge University Press, New York.

Furniss RL, Carolin VM (1977). Western Forest Insects, United States Department of Agriculture Forest Service Miscellaneous Publication Number 1339, 654.

Fotheringham AS, Zhan FB (1996) A comparison of three exploratory methods for cluster detection in spatial point patterns. Geographical Analysis, 28, 200-218.

Getis A, Griffith D (2000) Comparative spatial filtering in regression analysis. Geographical Analysis 34, 130–140.

Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. Geographical Analysis 24(3), 189-206.

Griffith D (1992) What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. L'Espace geographique 3, 265-280.

Haining R (1990). Spatial Data Analysis in the Social and Environmental Sciences. Cambridge: Cambridge University Press.

Haining R, Wise S, Ma J (1998) Exploratory spatial data analysis in a geographic information system environment. The Statistician 47(3), 457-469.

Hawksworth FG, Williams-Cipriani JC, Eav BB, Geils BG, Johnson RR, Marsden MA, Beatty JS, Shubert GS, Robinson DCE (1995). Dwarf Mistletoe Impact Modeling System User's Guide and Reference Manual, United States Department of Agriculture Forest Service Report MAG-95-2, 120.

Hernandez D, Clementini E, Felice PD (1995) Qualitative distance. Spatial Information Theory, Spring-Verlag, Lecture Notes in Computer Science, 988, 45-57.

Huang Z, Svensson P (1993) Spatial query language and analysis In: Advances in Spatial data Bases, Springer-Verlag, Lecture Notes in Computer Science, No.692: 413-436.

Laffan SW (2002) Using process models to improve spatial analysis. International Journal of Geographical Information Science 16, 245-257.

Legendre P, Dale MRT, Fortin M-J, Gurevitch J, Hohn M, Myers D (2002) The consequences of spatial structure for the design and analysis of ecological field surveys Ecography 25, 601-15.

Legendre P, Fortin M-J (1989) Spatial pattern and ecological analysis. Vegetatio 80, 107-138.

Lin and Lu (2006) Evaluating Local Non-Stationarity when Considering the Spatial Variation of Large-scale Autocorrelation. Transactions in GIS 10(2), 301-318.

Liu D, Kelly M, Gong P (2006). A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery. Remote Sensing Environment, 101, 167-180.

Nelson T, Boots B, Wulder MA (2006). Large-area mountain pine beetle infestations: Spatial data representation and accuracy. The Forestry Chronicle, 82, 243-252.

Nelson TB, Boots B, Wulder M (2004). Beetle infestations to characterize pattern, risk, and spread at the landscape level. In T.L. Shore, J.E. Brooks and J.E. Stone (eds.). Mountain Pine Beetle Symposium: Challenges and Solutions, October 30–31, 2003, Kelowna, British Columbia, Canada. pp. 164–173.Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, British Columbia, Information Report BC-X-399.

Okabe A, Boots B, Sugihara K (1992) Spatial tessellations: concepts and applications of Voronoi Diagrams, published by John Willey & Sons Ltd 1992.

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis 27(4), 286-306.

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. Journal of Regional Science 41(3), 411-432.

O'Sullivan D, Unwin D (2004) Geographic Information Analysis. Hoboken NJ, John Wiley and Sons.

Pélissier R, Goreaud F (2001) A practical approach to the study of spatial structure in simple cases of heterogeneous vegetation. Journal of Vegetation Science 12, 99-108.

Robertson C, Nelson TA, Boots B, Wulder MA (in press) STAMP: Spatial-temporal analysis of moving polygons. Journal of Geographical Systems.

Safranyik L, Linton, DA, Silversides R, McMullen LH (1992). Dispersal of released mountain pine beetles under the canopy of a mature lodgepole pine stand. Journal of Applied Entomology, 113, 441-450.

Shapiro SS, Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52, 591-611.

Sokal RR, Oden NL, Thomson BA, Kim J (1993) Testing for regional differences in means: Distinguishing inherent from spurious spatial autocorrelation by restricted randomization. Geographical Analysis 25(3), 199-210.

Taylor SW, Carroll AL (2004). Disturbance, forest age, and mountain pine beetle outbreak dynamics in BC: A historical perspective, in TL Shore, JE Brooks and JE Stone

(Eds.), Mountain Pine Beetle Symposium: Challenges and Solutions, Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, BC, Information Report BC-X-399, 298.

Tobler, WR (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46, 234–40.

Tukey, J.W. (1977) Exploratory data analysis, Addison-Wesley, Reading, Mass.

Turchin P, Thoeny W (1993) Quantifying dispersal of southern pine beetles with mark-recapture experiments and a diffusion model. Ecological Applications 3(1), 187-198.

Unwin A, Unwin D (1998) Exploratory spatial data analysis with local statistics. The Statistician 3, 415-421.

Wagner HH, Fortin MJ (2005) Spatial Analysis of Landscapes: Concepts And Statistics. Ecology 86(8), 1975–1987.

Westfall J (2006). 2005 Summary of Forest Health Conditions in British Columbia, British Columbia Ministry of Forests and Range, Forest Practices Branch, Victoria, BC, Pest Management, Report Number 15, 50.

Wulder MA, White JC, Dymond CC, Nelson T, Boots B, Shore TL (2006). Calculating the Risk of Mountain Pine Beetle Attack: a Comparison of Distance- and Density-Based Estimates of Beetle Pressure. Journal of Environmental Informatics 8(2), 58-69.

Wulder M, Boots B (1998) Local spatial autocorrelation characteristics of remotely sensed imagery assessed with the Getis statistic. International Journal of Remote Sensing 19(11), 2223-2231.

APPENDIX A – KS PARTITION METHOD

Partitioning method

A systematic approach was developed for partitioning the study region into sub-regions with approximately similar means throughout (sub-regions approach or achieve the assumption of stationarity). The method is *ad hoc* in that it is not documented anywhere, however it is in keeping with Arbia (1989) suggestion that a partitioning scheme should attempt to maintain the largest amount of data, such that the fewer divisions made, the less information lost. The method was used for its simplicity rather than striving for a more optimal partitioning that might be achieved by more rigorous, yet more time consuming and more complex methods. The approach involves subdividing the study area into equal parts and compares the subsamples with a KS test for two independent samples. If neighbouring samples are considered to be from the same population they are merged into one area, otherwise a further set of divisions is made and the process repeated. The partitioning method is a simple and practical approach to subdividing the region, without using any prior knowledge of the landscape of the study region or otherwise subjective decisions on where boundaries should be placed. Divisions can be made in the horizontal (as shown in Figure 1), vertical, or horizontal and vertical directions. Other methods for partitioning a study area in to homogeneous sub-regions that show great promise but are not investigated further here include studies by Dale and Powell (2001), Pelisser and Goreaud (2001) and Perry et al (1999). Figure 1 illustrates the method.

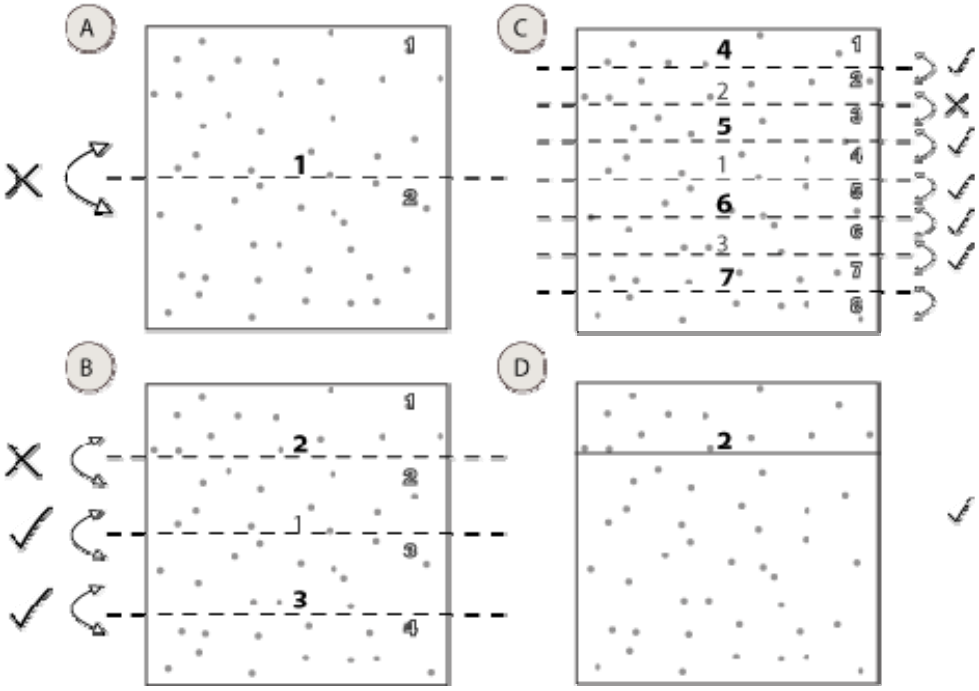


Figure 1 – KS – Partition method

Rules for partition method

The following rules are followed:

1. For a division to be made into a permanent partition boundary it must be statistically significant boundary at two levels of scale. For instance, in the diagram above the first division (A) creates significant divison between points from above the line and points from below the line, but this significance disappears when the study area is further divided into quarters. Therefore this division does not become a permanent boundary. However, the second set of divisions into quarters (B) identifies a significant difference between the first and second quarter, and the location of this division remains the same when the

uarters are further divided into eighths (C). Therefore this becomes a permanent partition boundary.

2. An exception to rule number 1 is if the subdivision reaches a size beyond which any further division would create final partitions that would be too small to be useful for meaningful analysis. Such a minimum scale for the smallest division would be set by the user based on prior knowledge of the process occurring or perhaps by some method similar to that used to determine an optimal distance band. In this situation, partition boundaries would be determined from the previous significant division, regardless of whether it occurs at two levels of scale or not.