2006

Detecting hot spots of mountain pine beetle infestations in the forests of British Columbia: An approach using local spatial autocorrelation

Rodrigo Tapia-McClung
Wilfrid Laurier University

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-21498-5
Our file Notre référence
ISBN: 978-0-494-21498-5

DETECTING HOT SPOTS OF MOUNTAIN PINE BEETLE INFESTATIONS IN THE FORESTS OF BRITISH COLUMBIA:

AN APPROACH USING LOCAL SPATIAL AUTOCORRELATION

by

Rodrigo Tapia-McClung
(Physics, B.Sc., Universidad Nacional Autónoma de México, 2004)

THESIS
Submitted to the Department of Geography and Environmental Studies
in partial fulfillment of the requirements
for the Master of Environmental Studies degree
Wilfrid Laurier University
2006

Abstract

Mountain pine beetle (*Dendroctonus ponderosae Hopkins*) is an endemic species in the forests of British Columbia that has become epidemic and reached infestation levels like never before. Different approaches have been taken in order to try and manage the forest and understand the processes affecting the behavior of mountain pine beetle.

No single model has been entirely successful in unearthing the complexity of mountain pine beetle behavior. In this thesis, large spatial data sets of mountain pine beetle attacks, obtained from helicopter and ground surveys, and further adjusted for the incorporation of uncertainty, are studied using a spatial autocorrelation approach in a pattern-based analysis.

The study of spatial patterns is carried out by simulating possible scenarios of the observed data set. Moran's I is used to obtain an overall measure of spatial autocorrelation of the global pattern and Local Indicators of Spatial Autocorrelation, specifically Local Moran's I, are used to identify local pockets of high levels of infestation (hot spots). Using a significance criterion, regions that have intense infestations are screened to retain those that are more pervasive, thus having a more robust set of results that can be more reliable. Different levels of significance can be used to allow for a more 'liberal' or 'strict' screening of results.

Study of the sensitivity of the data model and detection approach is carried out by comparing the locations of hot spots obtained with different detection methods. A comparison between results derived from data sets containing only aerial data and those containing aerial and field data is useful to determine the impact and effectiveness of sending crews to groundtruth aerial surveys.

Acknowledgements

Financial support for this research has been provided by the Mountain Pine Beetle Initiative, Wilfrid Laurier University, the GEOIDE (Geomatics for Informed Decisions) Network and the Dirección General de Relaciones Internacionales (Secretaría de Educación Pública - México).

I would like to thank all who have supported me throughout this process. Special thanks go to Dr. Barry Boots, a great person, a football fan and a very patient advisor. I owe him the opportunity of coming to Canada and pursue graduate studies. Thank you.

I would also like to thank Dr. Rob Feick for his help and guidance throughout the several stages of this project, for reading and commenting on earlier drafts of this manuscript. Thanks to Dr. Bob Sharpe for giving me the opportunity to explore beyond the realms of spatial statistics and letting me gain teaching experience in his labs. I would also like to thank Dr. Sean Doherty for his assistance and insights in different topics, especially near the end of the process.

I would like to thank my family for their support. It has been very difficult to go through all of this while being alone but they have always been there for me. I thank them for that and for helping me clearing my mind in moments of despair and distress.

I would also like to acknowledge the many good friends scattered all over the place that have shared their opinion countless times and helped me, in one way or another, to get through and finish! I sincerely thank you all! All of you have contributed somehow to this work and it is also yours.

Table of Contents

List of Tables

List of Figures

§ Chapter 1

Introduction

According to the British Columbia Ministry of Forests, the province is experiencing the largest infestation of the mountain pine beetle (*Dendroctonus ponderosae Hopkins*) ever recorded and there is no foreseeable end. From 1994 to 2003, 4.2 million hectares have been infested by mountain pine beetle in the entire province (British Columbia Ministry of Forests, 2003). Local communities whose economies are based on the forest are and will continue to be affected by the infestation as long as there is no scientifically driven management policy.

Forest values affected by mountain pine beetle infestations include: landscape aesthetics, water quality, wildlife habitat and timber supply that have economic implications on about 30 communities around the province and affects 25,000 families in British Columbia. Previous infestations have been tracked through time and outbreaks have been delineated using helicopter surveys since the 1960's. It is interesting to note the work of Wood and Unger (1996), which is a document containing a lot of information with respect to infestations and outbreaks all over the province from 1910–1995. It includes sections on the history of mountain pine beetle beetle in British Columbia for different forest regions and it is particularly suitable, for this study, to briefly summarize the information available for the Prince Rupert Forest Region, in which the study area for this research is located.

For this region, Wood and Unger (1996) present with great detail reports on the infestation levels for each year from 1970 to 1995. They also present the reader with information on the proportion of thousands of hectares and trees affected each year to track the history of infestations throughout time. It is interesting to note that during the 1980's infestation levels where reported to be unusually high compared to the previous decades affecting over 13,000 hectares and up to 1.4 million trees, reaching its peak in 1987. During the early 1990's infestation levels decreased reaching its lowest point around 1992-1993 and during 1994-1995 infestation levels started increasing again.

At the landscape level, infestation is currently taking place at epidemic proportions. For this study, landscape level or scale is understood as an area that is spatially heterogeneous in at least one factor of interest (Turner *et. al.*, 2001).

The general purpose of this study is to explore spatial and spatial-temporal behavior of the mountain pine beetle population distribution over large areas. A specific goal is to detect the locations that are exceptionally intensely affected (hot spots) using spatial analysis. My research will explore how sensitive the identification of hot spots is to the method used to represent the data (the spatial data model) and to the techniques used to identify them (spatial analysis technique).

Little is known about the mountain pine beetle's behavior at the landscape level. Previous studies have been carried out on a more local scale (e.g., individual tree stands) and behavior identified at that scale includes, but is not limited to: aggregation of individual beetles in response to chemical signals, the effect of spatial distribution of individual trees on the spatial pattern of the infestation, and differences in spatial patterns of infestations due to population distribution (Geiszler *et. al.*, 1980; Mitchell and Preisler, 1991; Logan *et. al.*, 1998). Different analyses of mountain pine beetle populations have been carried out and it has been recognized by researchers that it is imperative to consider spatial analysis in its studies (Bentz *et. al.*, 1993; Logan *et. al.*, 1998).

As mentioned before, most of the studies that include spatial analysis for mountain pine beetle populations have been carried out at a more local scale, thus aiding in the understanding of the processes that govern the behavior of the beetle at that particular scale. It should not, then, be assumed that the same process will operate at a landscape level, which is the scale considered in this research. Since infestations are occupying large extents of the forests, it is necessary to gain insights into the behavior at a larger level.

One of the reasons why landscape level studies have not been carried out much in the past is the complexity of mountain pine beetle behavior and the limitations posed by computational hardware (*e.g.*, processor speed and storage). Typically, studies were carried out for small study areas, since they offered the advantage of being a reasonably manageable analysis and computational requirements were not overwhelming. Also, the unavailability of large data sets was a major limitation for this approach. These limitations have been partially surmounted as large data sets have been collected and are readily available, computers are faster and storage media are able to support these larger data sets, thus making it possible to analyze mountain pine beetle spatial behavior within larger areas of study.

Basically, two approaches can be applied to the study of mountain pine beetle infestations: pattern-based or process-based. In general terms, a spatial pattern is the expression of one or more spatial processes (Getis and Boots, 1978), while a spatial process is one that reflects changes in its state due to the spatial properties of the attribute (Haining, 1993). In the case of the mountain pine beetle, pattern-based studies typically use pine mortality to explore the nature of mountain pine beetle spatial behavior. Process-based studies focus on direct observation or modeling of mountain pine beetle emergence, dispersal, and host selection. In this project, the pattern-based approach has been chosen since it provides an advantage at the landscape scale being used. This will be further explained in a later chapter.

The research goal is to locate the areas of major infestation. Once these have been identified, the primary issue will be to compare the hot spots previously identified using another approach (Kernel Density Estimated – KDE – surfaces) with those obtained with the present analysis (Local Indictors of Spatial Autocorrelation – LISA). As it will be explained in more detail in Chapter 2, for some years data were collected by aerial and ground surveys and the analysis will be carried out using these two different sets of data. It will be important to assess the differences of utilizing field data obtained from the surveys and aerial data.

It is worth noting that this thesis does not relate profoundly to the mountain pine beetle entomology or pine biology. There are several books, articles and other sources of information to which the reader is referred if further knowledge on this topic should be required. As examples of such sources, the following can be listed: Safranyik *et. al.* (1974); Geiszler *et. al.* (1980); Preisler and Mitchell (1993); Amman *et. al.* (1988), among many others. Also Nelson (2005) has a brief, yet comprehensive, summary of mountain pine beetle entomology and pine biology and further references can be found therein.

The present work is organized as follows. In Chapter 2, the area of study is presented along with the data sets. Special emphasis is given to the distinction between the two different data sets that will be used for this analysis and the way they were obtained. In Chapter 3, a brief summary of previous research is presented and the research methods used here are explained. Chapter 4 presents the results of this study and finally, Chapter 5 summarizes and presents the conclusions of this work, as well as some ideas for future research.

§ Chapter 2

Area of Study and Data

2.1 Morice Timber Supply Area

The area of study of this research is the Morice Timber Supply Area, (54°24′ N, 127°38′ W). Morice is part of the British Columbia Ministry of Forests, Nadina Forest District and is centered on the small town of Houston (Figure 2.1). It covers approximately 1.5 million hectares and it has been one of the areas impacted by mountain pine beetle infestation.

It is delimited by the Cascade Mountains to the west and Tweedsmuir Park to the south. The topography of the region shows a trend to be more mountainous towards the southwest. Within the region are two large lakes: Babine Lake in the north and Ootsa Lake in the south, as well as three major rivers: the Bulkey, the Morice and the Nadina. The area is dominated by lodgepole pine (*sp. Pinus*) and spruces (*sp. Picea*).

2.2 Data

The mountain pine beetle infestations have been monitored in the study area since 1995 using point-based, global positioning system aerial surveys. Aerial surveys

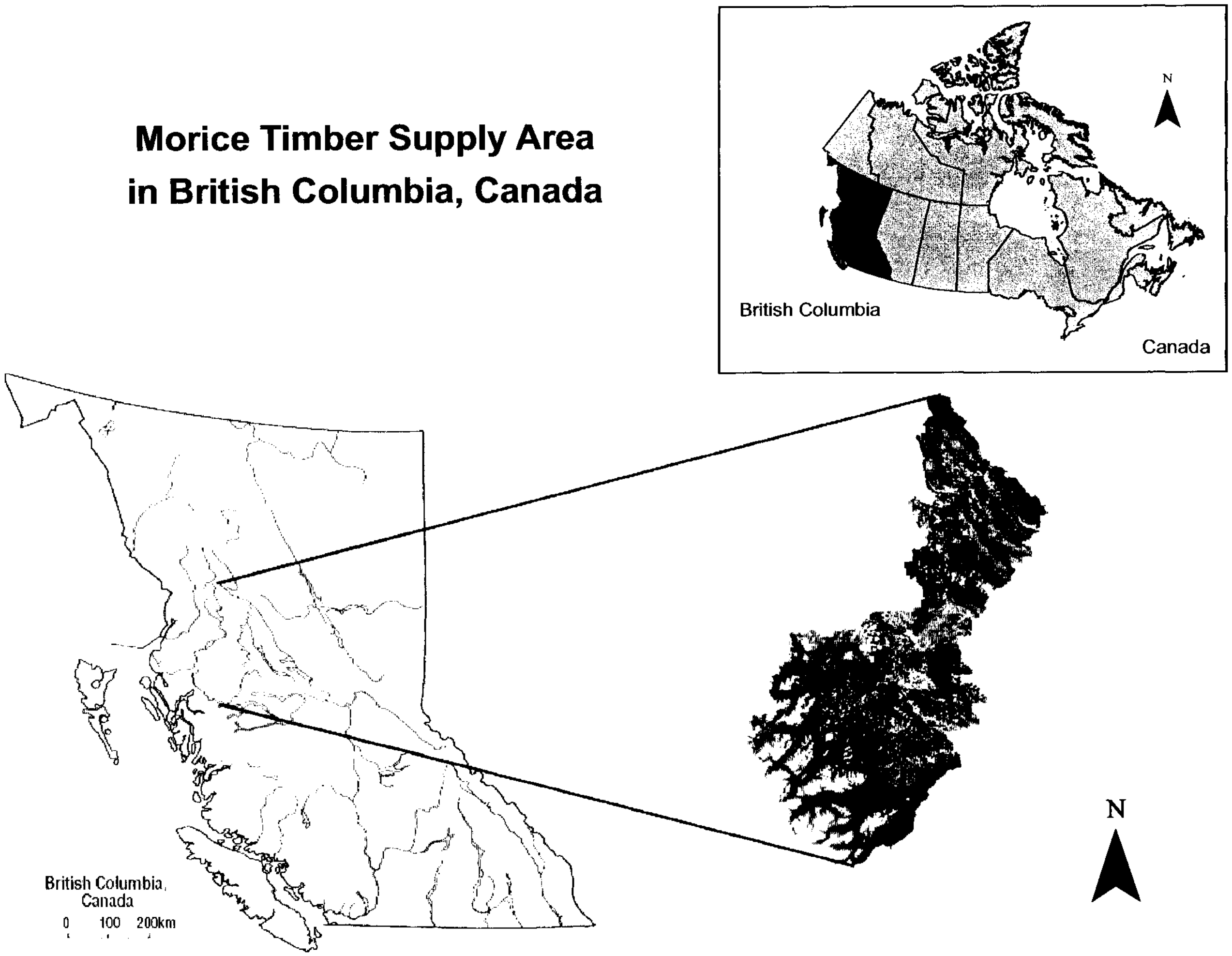


Figure 2.1: Location of the Morice Timber Supply Area. Source: Natural Resources Canada.

use indicators of pine mortality to monitor mountain pine beetle activity. When pine trees are attacked by mountain pine beetles, crown foliage changes successively from green to yellow, brown, red and finally to grey leaving the tree just with the stem and branches (Safranyik *et. al.*, 1974).

Pine trees are usually attacked during summer and the first visible change in foliage color usually occurs the next spring after the attack. Typically one year after the attack the foliage is yellow green or yellow brown. Two years after the attack the foliage is usually red, and by the third year, needles fall off the tree.

During aerial surveys, if a cluster of infested trees is detected it will be identified and a cluster center is mapped with a point. For each cluster, the number of infested trees is estimated and the infesting insect species recorded. The maximum area represented by a point is approximately 0.031 km^2, equivalent to a circle with a radius of 100 meters; each data point represents from 1 to a maximum of 300 trees. From 1995 to 2002 a total of 42,632 points were identified during aerial surveys, of which field data were collected for 6,151 points, between 1999 and 2002. Table 2.1 shows the data available for each year, for both aerial and ground surveys.

Survey Sites	1995	1996	1997	1998	1999	2000	2001	2002	Total
Aerial	2,181	6,076	8,461	2,418	4,657	5,310	5,226	8,308	42,637
Field	0	0	0	0	223	104	3,004	2,820	6,151

Table 2.1: Annual number of aerial and field survey sites.

Field data are obtained when ground crews locate the infestation clusters that were identified during aerial surveys and determine the cause of tree mortality. If trees were killed by mountain pine beetle, the number of green trees under attack, the number of trees attacked the previous year and two years previously as well as the number of trees attacked, and now gray, are recorded. The crews also note the presence of any non-mountain pine beetle infestations on the field. It is important to note, however, that even though green-tree attack is recorded it is not used in this study in order to have comparable results with what has been obtained from the helicopter surveys.

When data have been collected in the field, records are updated and changed and it is expected that those databases containing only information from aerial surveys will be less accurate than the ones containing field data. Field data sets will often contain more records than their aerial counterpart since typically, when crews are inspecting areas in the field, more clusters of infected trees are found and recorded.

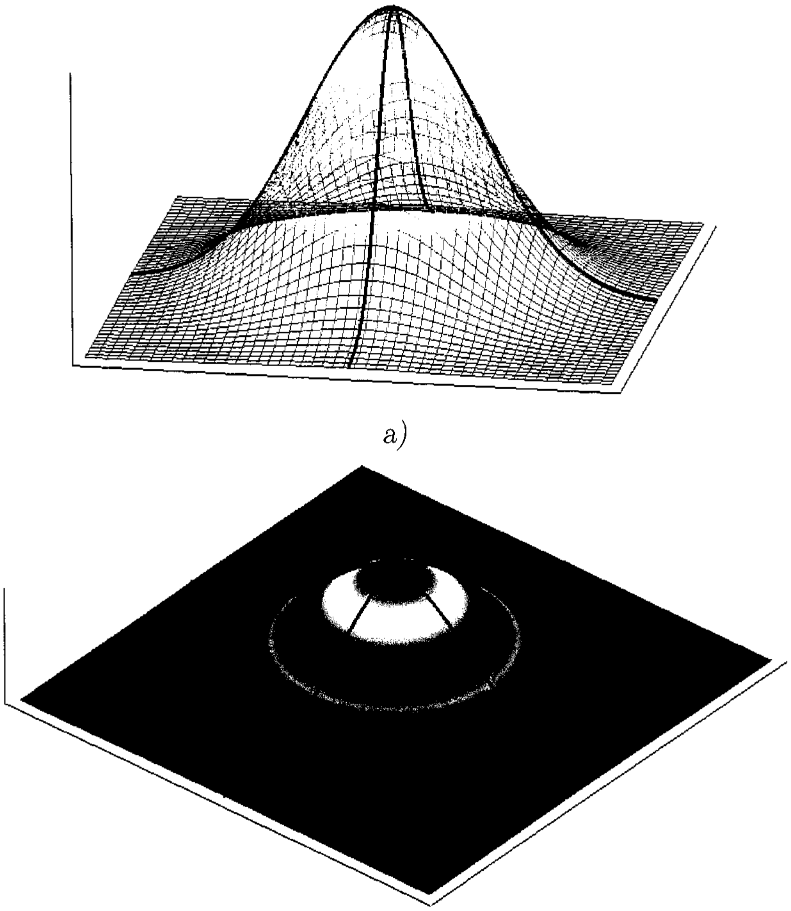
It is important to mention that field data are used in this analysis in order to identify errors in aerial data and adjust these accordingly. It will provide insights to see whether considering error, as identified in the field data, provides an advantage over the data collected from the air. This work will involve two aspects to address the problem: one with what has been called 'Adjusted Aerial Data' and the other which was called 'Adjusted Aerial-Field Data', which are presented below.

2.2.1 The Adjusted Aerial Data

These data sets are available from 1995 to 2002 and will be useful to check for consistency of hot spots throughout time. Data have been collected in the field for some of these years, thus making it possible for a comparison of field and aerial data at those sites where both are available. This in turn allows for identification and modeling of error in the aerial data. Once this information is available, it is used to adjust the aerial data of those years for which no field data was collected.

In order to incorporate adjustments on spatial error and uncertainty, simulated data sets of what the real data might have been, given knowledge of the error obtained from the field data, are created. For detailed information on the error and data accuracy from mountain pine beetle aerial and field surveys, Nelson (2005) contains a comprehensive discussion on this topic. The simulations used in this research were obtained in the following way: spatial error for each point is estimated by field crews to be within $\pm 25m$ (Nelson *et. al.*, 2006). Values are randomly drawn from a standardized normal distribution and then scaled between ± 25 and added to both the x and y coordinates of each point. This is a valid approach since for either the $x-$ or $y-$axis, the distribution of GPS error is approximately normally distributed (Leva *et. al.*, 1996). In other words, one could consider that for each simulation a point is randomly displaced from its position based on a probabilistic approach using a two-dimensional normal distribution. Figure 2.2 shows an illustration of this idea. In this Figure, the $x-$ and $y-$ axes represent the spatial coordinates of any given point while the $z-$ axis represents, and is shaded according to, the probability distribution of spatial error that will be assigned to each point in every simulation. Also, in this Figure, each black line represents the normal distribution for each axis, $x-$ and $y-$. It can be seen that smaller displacements for spatial locations have the higher probabilities (near the center of the distribution), while greater displacements are less likely to occur since the probability greatly decreases as we move away from the center of the distribution.

As it was recognized in Nelson (2005), the number of infested trees has uncertainty associated to it and is important to try and take this into account. In order to do this, attribute error was simulated by drawing values from a distribution. A cursory exploration suggested it was necessary to use more than one distribution and, based on natural breaks in the frequency, three categories for attribute values were used: one to five, six to ten, and greater than ten, to properly reflect changes in the distribution of attribute data depending on the number of infested trees. It was found



a)

b)

Figure 2.2: Spatial error probability distribution for both axes. Illustrations are shaded according to probability values ranging from purple for lower to red for higher levels: *a)* shows a wireframe so the normal distribution along each axis can be seen; *b)* is shown as a continuous surface and probabilities can be more easily identified.

that the distribution for each category appeared to follow a Gamma distribution. For each class, a two-parameter Gamma distribution was fitted and used to simulate the attribute values for the realizations.

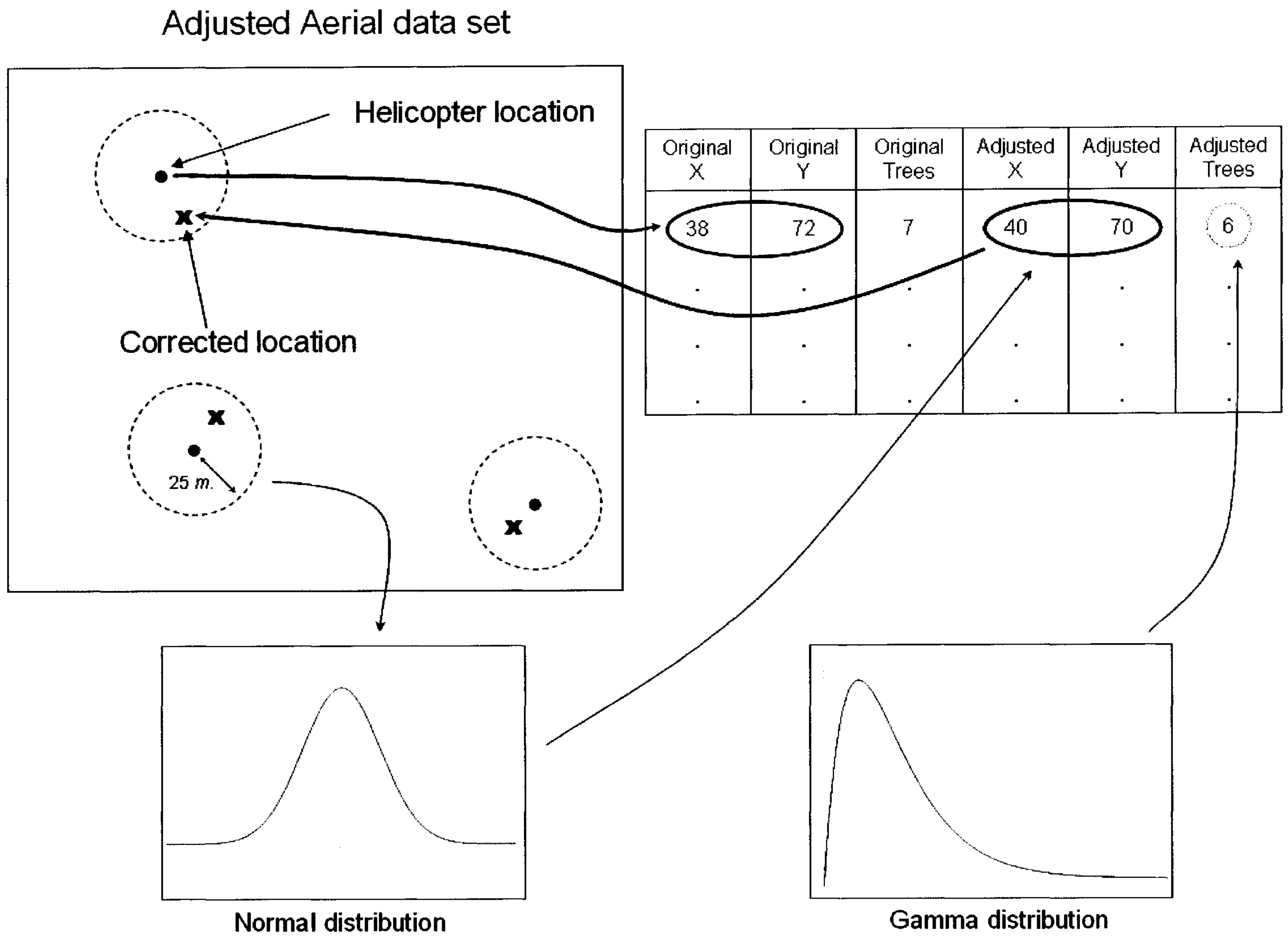
Ground crews found locations where there were no infested trees and assigned these locations a value of zero. Since the Gamma distribution does not contain zero values, there was a limitation when simulating attribute values and keeping the zero values obtained from field surveys. In order to overcome this problem, the percentage of zeros was kept and the number of zero values assigned to each year were treated like random draws from the distribution. With parameters estimated and the percentage of zero values fixed for each year, attribute values for each point were simulated by randomly drawing values from the Gamma distributions. An illustration of the conception of the Adjusted Aerial data sets is shown in Figure 2.3. This illustration can be explained in general terms as follows: the coordinates obtained from the aerial survey are subject to be simulated by means of displacing the original locations by adding random values drawn from a normal distribution (scaled between ±25), while the number of trees is simulated using a Gamma distribution.

2.2.2 The Adjusted Aerial-Field Data

The Adjusted Aerial-Field data sets are available from 1999 to 2002. These data sets were derived from the previous Adjusted Aerial data sets by including the field data values at sites where these values were available, hence the label Adjusted Aerial-Field. These data sets will give the possibility to explore two aspects: test for the impact of the corrected data set (Adjusted Aerial *vs.* Adjusted Aerial-Field) on hot spot detection, and compare the detected hot spots with those recognized previously using another method. Previous studies dealing with detecting mountain pine beetle hot spots in the same study area have been carried out using the Adjusted Aerial-Field data sets. It is worth noting that to date, no previous work on comparing the impact of including field data or not to correct the data sets has been carried out. Also, field data counts are not adjusted, and are considered to be closer to true than those obtained via helicopter surveys. For the Adjusted Aerial-Field data sets, spatial error for each point was also simulated by displacing the original locations $\pm 25m$ by randomly drawing values from a normal distribution for each axis. An illustration of the conception of the Adjusted Aerial-Field data sets is shown in Figure 2.4.

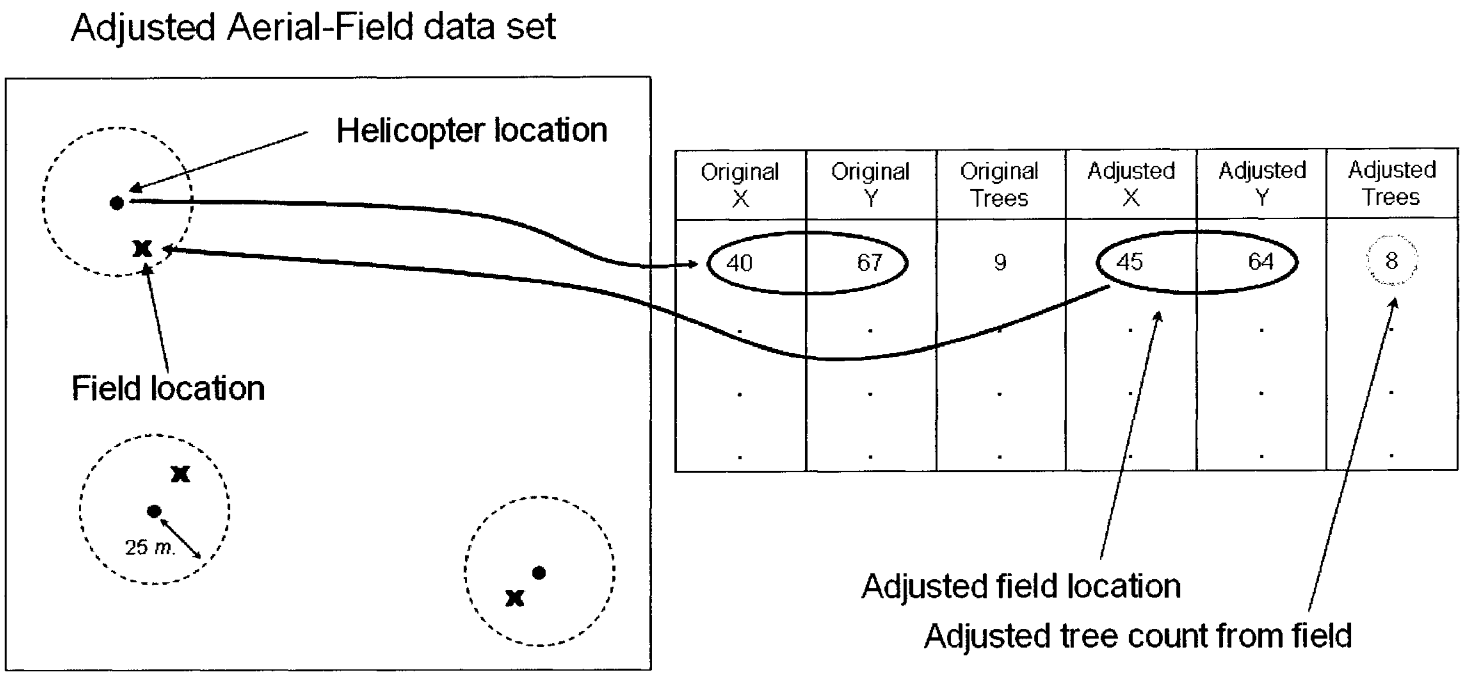
In this case, this illustration can be explained in general terms as follows: original coordinates from the aerial survey are displaced by adding random values drawn from a normal distribution (scaled between ±25), while the number of trees is either kept if data field was collected for any given location or is corrected using a Gamma distribution.

For each case a total of 100 simulations was generated in order to create 100 different spatial representations of the point data, for hot spot detection using local spatial autocorrelation techniques. These steps are explained in the next chapter. Illustrations showing the idea behind the generation of these realizations are shown in Figures 2.5 and 2.6. An illustration showing a summary of the process for obtaining the simulations for both data sets is shown in Figure 2.7.



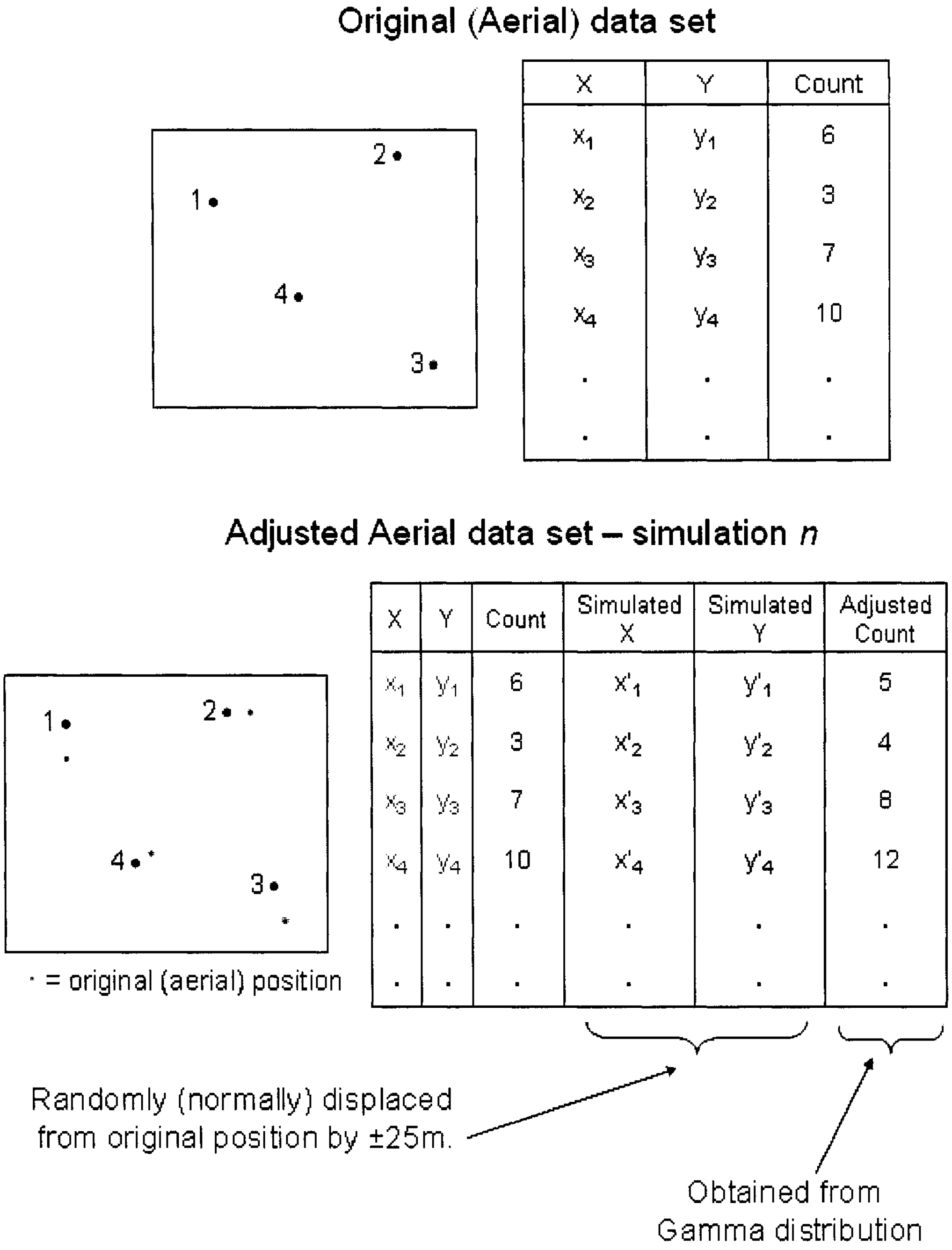
Original X	Original Y	Original Trees	Adjusted X	Adjusted Y	Adjusted Trees
38	72	7	40	70	6
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Figure 2.3: Illustration of the construction of the Adjusted Aerial data sets. Original locations are displaced using a normal distribution; number of trees are adjusted using a Gamma distribution.



Original X	Original Y	Original Trees	Adjusted X	Adjusted Y	Adjusted Trees
40	67	9	45	64	8
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

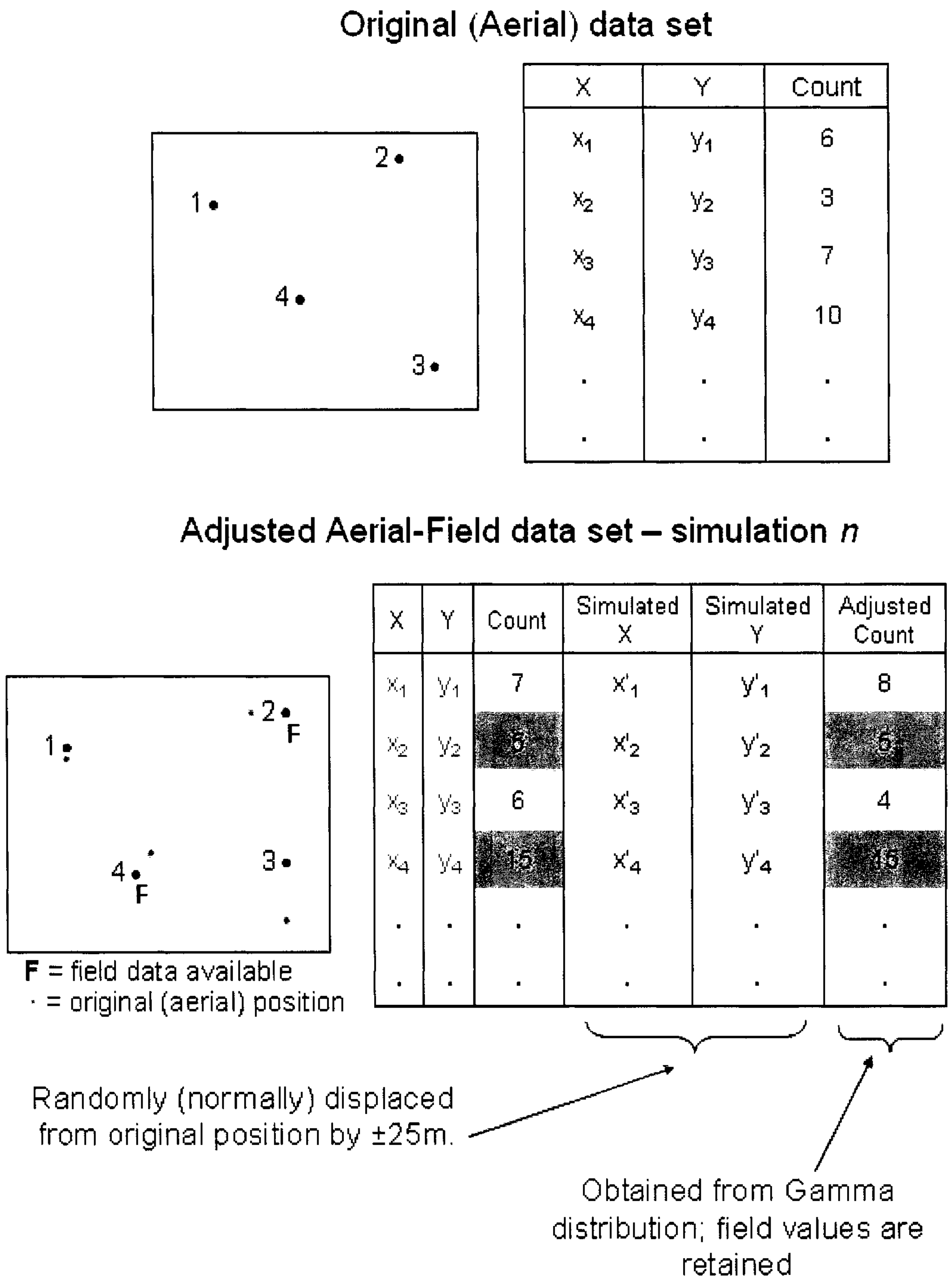
Figure 2.4: Illustration of the construction of the Adjusted Aerial-Field data sets. Original locations are corrected when visiting the site and then displaced using a normal distribution; number of trees are corrected when visiting the sites.


Original (Aerial) data set
1
2
4
3
Adjusted Aerial data set – simulation n
1
2
4
3
· = original (aerial) position
Randomly (normally) displaced from original position by ±25m.
Obtained from Gamma distribution

X	Y	Count
x_1	y_1	6
x_2	y_2	3
x_3	y_3	7
x_4	y_4	10
.	.	.
.	.	.

X	Y	Count	Simulated X	Simulated Y	Adjusted Count
x_1	y_1	6	x'_1	y'_1	5
x_2	y_2	3	x'_2	y'_2	4
x_3	y_3	7	x'_3	y'_3	8
x_4	y_4	10	x'_4	y'_4	12
.	.	.	.	.	.
.	.	.	.	.	.

Figure 2.5: Illustration of an example of the process to obtain the Adjusted Aerial data set simulations. Original locations are displaced using a normal distribution; number of trees are adjusted using a Gamma distribution.



Original (Aerial) data set

X	Y	Count
x_1	y_1	6
x_2	y_2	3
x_3	y_3	7
x_4	y_4	10
.	.	.
.	.	.

Adjusted Aerial-Field data set – simulation *n*

X	Y	Count	Simulated X	Simulated Y	Adjusted Count
x_1	y_1	7	x'_1	y'_1	8
x_2	y_2	5	x'_2	y'_2	5
x_3	y_3	6	x'_3	y'_3	4
x_4	y_4	15	x'_4	y'_4	15
.	.	.	.	.	.
.	.	.	.	.	.

F = field data available
· = original (aerial) position

Randomly (normally) displaced from original position by ±25m.

Obtained from Gamma distribution; field values are retained

Figure 2.6: Illustration of an example of the process to obtain the Adjusted Aerial-Field data set simulations. Visited field locations are displaced using a normal distribution; number of trees are corrected when visiting the sites or adjusted using a Gamma distribution if no field data are available.

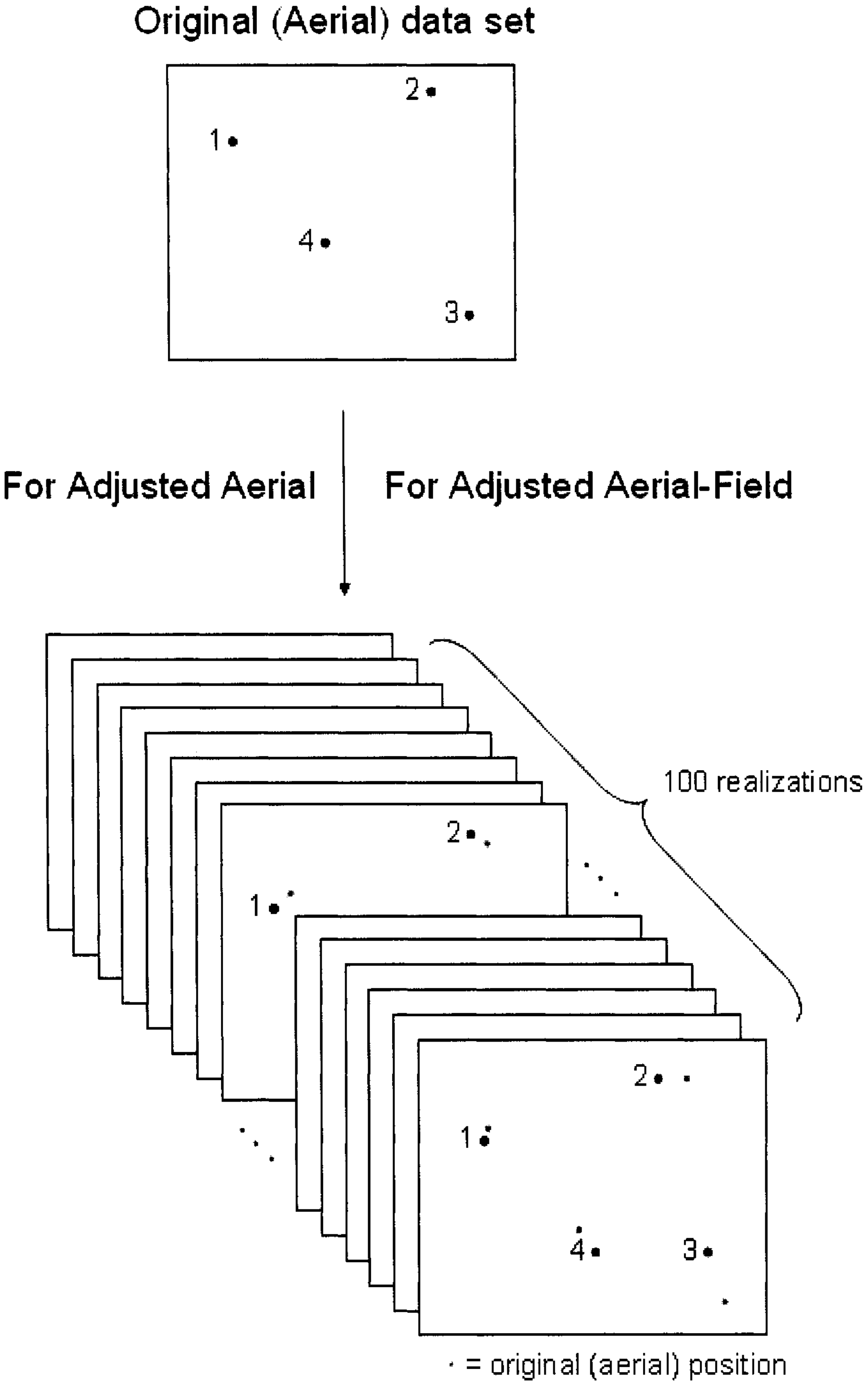

Original (Aerial) data set
2•
1•
4•
3•
For Adjusted Aerial
For Adjusted Aerial-Field
100 realizations
2•
1•
2•
1•
4•
3•
· = original (aerial) position

Figure 2.7: Illustration of the summary of the process to obtain the 100 simulations for both the Adjusted Aerial and the Adjusted Aerial-Field data sets.

§ Chapter 3

Research Methods

The spatial behavior of the mountain pine beetle is not well understood, but it seems to be increasingly recognized that spatial dynamics play an important role (Bentz *et. al.*, 1993; Powell and Rose, 1997; Logan *et. al.*, 1999; Powell *et. al.*, 2000). As was mentioned in the Introduction, there are two approaches for studying mountain pine beetle infestations at the landscape level: pattern-based or process-based. Pattern-based studies typically use pine mortality to explore the nature of mountain pine beetle spatial behavior. On the other hand, process-based studies focus on direct observation or modeling of mountain pine beetle emergence, dispersal, and host selection.

In this study, the pattern-based approach is preferred since it provides an advantage at the landscape scale insofar as it is more useful for exploratory analysis. This is because the processes of mountain pine beetle dispersal and host selection are not well understood to derive a good pattern that can be compared to the ones observed on the ground. In other words, process-based analysis poses further technical limitations to the study since there is insufficient knowledge of the processes to derive good models.

In this case direct data on the spatial process, the mountain pine beetle behavior, are difficult to obtain. With the spatial pattern approach it is possible to make inferences regarding processes but unfortunately, the relationship between these two is complex because spatial patterns are usually the result of several interacting processes and it can be either very difficult to identify or assess the impact of all of them.

3.1 What is a Hot Spot?

The term "hot spot" has been coined and used in many different disciplines and can have several meanings. It could very well mean that something out of the ordinary has occurred (Ord and Getis, 2001); that there is an excess of an event in space or time compared to an expected value (Wartenberg and Greenberg, 1992) or that an unusual absence of an event has occurred (Sokal *et. al.*, 1998); and more recently, it has been understood to refer to accessibility to a wireless network in a certain place (Manjunath *et. al.*, 2004). It is clear from the above that one definition is not enough and the suitability of a particular definition depends on what is being studied. The term is so widely employed and is so flexible that typically, in order to have a consistent definition, it is necessary to take into account the spatial pattern, hypothesis, techniques and available data, but it is also important to consider that the definition of a hot spot will influence the method that is used for its detection.

In general terms, it is safe to say that hot spots are locations where something "unusual" is occurring. That is to say that, whatever it is that is being measured, is showing different values from the expected ones. In other words, these values are exceeding a threshold. Thresholds can be defined absolutely or relatively and either choice will be useful in some instances, but will suffer from certain disadvantages as well.

It is useful to choose relative thresholds when analyzing a specific data set, but it is important to be aware that the results will very likely be unique to that data set and, in general, will not be comparable to a different data set. Relative thresholds may be useful for detecting 'local extreme' values, where absolute thresholds are helpful for identifying extreme values across data sets. It is important to note that when global characteristics do not change much over different data sets that are being analyzed, the use of relative or absolute thresholds will not make a big difference. Also, when using relative thresholds, hot spots will always be identified. This does not necessarily hold when using absolute thresholds.

Perhaps it is useful to clarify the difference between each threshold type by presenting an example. Consider a data set that contains the heights of a sample of the population of a certain region. When interested in finding the tallest people in this area it is possible to take two approaches. For the first one, think of the data set as heights sorted in an ascending order and then define, in a totally arbitrary way, that a person will be considered to be part of the 'tallest people' if their height falls within the top 15% of the heights. In the second case, consider a person to be part of the 'tallest people' if their height is above a certain pre-defined level, say for example $1.80m$. It is clear from these two definitions that the outcomes of these two classification schemes are likely to be different simply because the first will always contain elements, while the second not necessarily. It is possible that the data set will not contain heights above $1.80m$, therefore the second method would not contain any elements. On the other hand, it is easy to see that considering the top 15% of the individuals in the data set will inevitably give us people that are classified amongst the 'tallest'.

In this example two different classification thresholds are used: a relative one in the first case and an absolute one in the second. In other words, the relative threshold helps to identify those people who, within or relative to the data set, are the tallest. The absolute threshold helps identifying those whose heights are above a specified level, an absolute measure with respect to the whole data set.

3.2 Previous Studies in Mountain Pine Beetle Spatial Analysis

There are several studies of Mountain Pine Beetle involving both pattern and process-based approaches. Previous process-based studies, have focused on understanding reactions of beetles to different stimuli as well as obtaining data on beetle movement and 'migration'. One disadvantage of this approach is the limited coverage one can have, as these studies are only suitable at the tree or stand level, therefore they are not useful for a study at the landscape level. Examples of this type of work are Safranyik *et. al.* (1989, 1992); Turchin and Thoney (1993); Powell and Rose (1997); Powell *et. al.* (2000); Logan *et. al.* (1998).

Also, starting from a process-based study point of view, in the cited papers, a model for the mountain pine beetle pheromone ecology and single tree processes is considered, and Logan *et. al.* (1998) simulates the spatial process and generates an expected spatial pattern of pine mortality. Direct pattern studies include those of Mitchell and Preisler (1991) and Preisler and Mitchell (1993), in which the spatial pattern of individual trees is studied as well as the behavior of attacking beetles with respect to host trees. The approach of these process-based studies has been useful to shed some light into mountain pine beetle dispersal and host selection at the stand level. Unfortunately, beetle 'migration' usually spreads across the landscape covering several kilometers, therefore making it difficult to use the results from these studies at the landscape level.

Pattern-based studies at the landscape level seem to be scarce, but Nelson (2005) is worth noting. In this work, hot spots of infestation were obtained by converting the original data from marked point patterns – that is, cluster centroids with the number of infested trees – to continuous surfaces by means of using Kernel Density Estimator (KDE) methods (Diggle, 1985; Gatrell, 1994; Bailey and Gatrell, 1995). Hot spots were defined using the relative upper 10% threshold of intensity values of the obtained surfaces. One disadvantage of defining hot spots in this fashion is that it is an arbitrary way of doing it and does not give a helpful indicator that can be traced through time, as the upper threshold is likely to change from year to year, making it hard to compare results across years. Spatial-temporal analysis involved finding significant differences between pairs of KDE surfaces.

In this research, the definition of a hot spot consists of the following parts:

- Spatial Autocorrelation
- Statistical Significance
- Concentration Measurements

each of which will be elaborated throughout the rest of this chapter.

3.3 Spatial Autocorrelation and Hot Spots

For this research it is suitable to define a hot spot as the location of an abnormality in the spatial pattern that is being observed, the spatial pattern being a realization of a spatial process (Haining, 1993; Tiefelsdorf, 2000). The goal is to identify those locations that have infestation levels that are significantly above an expected value, that is, showing dissimilar and abnormal values to the rest of the areas. Because of mountain pine beetle processes it is expected that data values will be clustered, or spatially autocorrelated, since the underlying idea behind the process is that beetles tend to locate themselves in certain regions of space, thus creating a spatial pattern.

Large and mature trees are preferred by mountain pine beetles for several reasons. Thicker phloem provides optimal food for the beetle, a thicker bark protects the beetles from cold and predators, and older trees are more easily attacked, colonized and killed (Safranyik *et. al.*, 1974; Geiszler *et. al.*, 1980; Preisler and Mitchell, 1993), suggesting that patches of older trees are more likely to be attacked if there is an infestation in the vicinity. It also has been suggested that pine mortality depends on stand density (Amman *et. al.*, 1988) and could be a reason for host selection by mountain pine beetle. Mountain pine beetle dispersal is not clearly understood but it is influenced by chemical signals, temperature, light and wind direction (Safranyik *et. al.*, 1989; Powell *et. al.*, 2000). The combination of tree age, stand density, chemical signals and seasonal changes are factors that affect the location of beetles and may trigger the accumulation of beetles in certain regions, but in general it is very difficult to propose a proper model that takes all the known variables into account and accurately reproduces the observed spatial patterns at the landscape level.

From air and field surveys, information is available on infested trees, and a hot spot will be understood as a location of major levels of infestation. However, since data will show some form of spatial pattern, it is necessary to search for these intensely infested areas on a more local scale. It is important to bear in mind that local patterns and fluctuations may be reflected in the global pattern, but not necessarily, since the possibility exists that the local pattern is something very uncommon that is not picked up by the global pattern. It is also possible that localized areas may behave in an opposite fashion to the general trend of the pattern.

Spatial autocorrelation is a measure of the relation between spatial data from near locations. According to Tobler's first law of geography: everything is related to everything else, but nearby things are more related than distant things (Tobler, 1970). Furthermore, as stated by O'Sullivan and Unwin: "...any set of spatial data is likely to have characteristic distances or lengths at which it is correlated with itself, a property known as self-correlation or *autocorrelation*" (O'Sullivan and Unwin, 2003).

One way to assess the degree of spatial autocorrelation is by means of global statistics such as Moran's I or Geary's c (Bailey and Gatrell, 1995; Ord and Getis, 1995, 2001; O'Sullivan and Unwin, 2003). Such statistics assume there is a certain degree of stationarity or structural stability throughout space and work very well under these conditions. However, in cases where it is recognized that spatial randomness is not the process underlying the observed spatial pattern, using these statistics might not be entirely appropriate. Still, they are able to provide insights into what is going on globally with the pattern under study. Unfortunately it is difficult to derive better models that take into consideration these complex spatial variations. One alternative for taking into account local instabilities and focus on local patterns is the use of a localized version of these statistics, which allows consideration of the contribution of individual observations and that altogether reflect the overall behavior of the pattern. As defined in Anselin (1995), Local Indicators of Spatial Autocorrelation (LISA) are statistics that satisfy the following criteria:

1. The LISA for each observation gives an indication of the spatial clustering of similar values around that observation.

2. The sum of LISAs for all observations is proportional to a global indicator.

In this sense, the value of the local indicator depends on the value of a variable in a certain specific location, as well as on other observations located in a neighborhood centered in the location under study.

The use of Global Moran's I will help in giving a general idea of the overall spatial pattern and autocorrelation at the landscape level. When interested in detecting, at a much finer scale, the locations that are presenting unusual numbers of infested trees, it is much more appropriate to consider the localized version of Moran's I, that is, the I_i statistic. The expression for Global Moran's I is:

$$I = \left(\frac{n}{\sum_i \sum_j w_{ij}} \right) \left(\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \right), \tag{3.1}$$

or in a more simplified form:

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i \sum_j w_{ij}}, \tag{3.2}$$

where:

- $z_i = (x_i - \bar{x})/s$, are deviations from the mean,
- x_i, are individual observations,
- $\bar{x}$, is the average value of these observations,
- s, is the standard deviation of the sample, and
- n, is the total number of elements.

The summations are taken over all of the observations i and over the number of observations, j, that are contained within a certain neighborhood around each element i. The terms w_{ij} are the elements of the weights matrix, which are used to properly reflect the influence of the neighbors around each observation to the overall result, that is, they are used to indicate how each possible covariance term will affect the calculations. The simplest case is an adjacency matrix in which $w_{ij} = 1$ if the elements i and j are adjacent and 0 otherwise. Adjacency matrices are used throughout this study and the way to obtain them is explained further on in this chapter.

Local Indicators of Spatial Autocorrelation come into play when it is acknowledged that global spatial autocorrelation is expected to be present given the nature of the data, as was mentioned before, and the use of a global statistic is therefore not able to give much more information about the spatial pattern under study. As stated by Anselin (1995): "...the assumption of stationarity or structural stability over space may be highly unrealistic". The use of global Moran's I would completely ignore this potential instability, thus the search for local instabilities is considered to be much more appropriate. Since LISAs are defined to be such that the summation of all the contributions is proportional to the global indicator, they allow for each individual observation to have a value attached to it that can help to shed some light on the behavior of the pattern at local scales. In the realm of spatial autocorrelation, a hot spot is typically understood to be a collection of spatial locations that have a significant local statistic, or in other words, a local spatial cluster. The use of global Moran's I will give some indication of the overall spatial autocorrelation of the pattern and Local Moran's I will be used to search for hot spots by comparing the number of infested trees in the vicinity of each location to the number of trees in each location. The expression for Local Moran's I may be written as (Gatrell, 1994; Bailey and Gatrell, 1995; Fotheringham *et. al.*, 2000):

$$I_i = \frac{(x_i - \bar{x})\sum\limits_{j \neq i} w_{ij}(x_j - \bar{x})}{\sum\limits_{i}(x_i - \bar{x})^2/n} \,, \tag{3.3}$$

or in a more simple way as:

$$I_i = z_i \sum_{j \neq i} w_{ij} \, z_j \ , \tag{3.4}$$

where the elements z_i and w_{ij} are as defined above.

The definition of a hot spot that is used throughout this research and was previously introduced will be further refined by including a statistical significance criterion, which will allow for a more effective use of the Local Moran's I statistic. A potential hot spot will be considered as a location that has a significant I_i value. The general course of action for hot spot detection is the following: calculate I_i for each point, identify locations that have a significant statistic with either positive or negative spatial autocorrelation and use a significance criterion to identify unusually large infestation levels. Thus, a hot spot will be considered as such if it has a significant I_i value and is consistent throughout a fair number of simulations. This will be explained in more detail in the next section. Numerical computations to obtain both Moran's I and I_i will be carried out using the GeoDa software package (Anselin, 2004). An alternate approach would include the utilization of different local statistics such as local Getis G_i or local Geary's c, that are defined similar to I_i (Anselin, 1995).

Since there are 100 realizations for each year, for which each point has been randomly displaced from its original position and the spatial attribute has also been randomly assigned, it would be difficult, if not impossible, to make a comparison between hot spots detected from one simulation to the next. If this were feasible, then there would be the cumbersome process of summarizing these comparisons and to then compare the hot spots throughout the years. Given the difficulty of using the I_i statistic values to directly carry out hot spot detection, a significance criterion is used. This is explained in detail in the following section.

3.4 Statistical Significance and Hot Spots

Analyses can be carried out by including a large number of significance tests or by repeated processes. However, there is a problem with the interpretation of results because, if something is tested several times, it is possible to find something "significant" (Bland and Altman, 1995; Voss and George, 1995; Caldas and Singer, 2006). Indeed, as it has been recognized by Caldas and Singer (2006): "...assessing the significance of multiple and dependent comparisons is an important, and often ignored, issue that becomes more critical as the size of data sets increases. If not accounted for, false-positive differences are very likely to be identified". This problem has been recognized and there is no straightforward way to eliminate it, although several methods have been developed to deal with this. However, its consequences and attempts for solution are beyond the scope of this research. If a more in-depth discussion on this topic is desired, the reader is encouraged to review Caldas and Singer (2006).

It is also important to recognize the fact that tests for both global and local spatial autocorrelation are compared to a null hypothesis of spatial randomness (or no spatial autocorrelation). Another problem is that typically neighborhoods will contain common elements, thus their corresponding statistics will be correlated and we are in the presence of non-independent tests. This ties in with the problem of repeated processes (or multiple comparisons) and the net impact is that it is not possible to accurately interpret the value of the significance (Anselin, 1995; Caldas and Singer, 2006). One way to assess how reliable the obtained results are is to change the significance levels and compare the results, thus having an idea of how dependent and sensitive results are on the chosen significance level. This provides a quick way around the problem of repeated processes without much formality. Typically strict levels of significance (*e.g.*, 0.05, 0.01 and so on) are used in order to try and rule out false positives, that is, values that are incorrectly reported to be true when, in fact, they are not.

Statistical significance is a way of arriving at the conclusion of how 'real' a potential hot spot may be. A significance value can be computed for each point and help assess if, within a certain specified confidence interval, a potential hot spot is in fact a hot spot. Not only is statistical significance useful for discerning the local behavior of data points, but it is also useful in determining if the values of global Moran's I are significant, thus permitting us to recognize if we are indeed in the presence of global spatial autocorrelation or not.

3.4.1 Testing significance for Global Moran's I

Statistical significance is implemented in the GeoDa software package (Anselin, 2004) based on a permutation approach. In this procedure, a reference distribution is calculated for spatially random layouts with the same data values as observed. 9,999 permutations are generated and are used to generate the reference distribution (see Figure 3.1), having a total of 10,000 realizations. The significance level is computed as the ratio of the number of statistics for the randomly generated data sets that are equal to or exceed the observed statistic + 1, over the number of permutations used + 1. In simpler terms, it is defined as the ratio obtained by the division of the number of times a statistic is found to have a value greater or equal to the observed one and the total number of realizations. In the example shown in Figure 3.1, the number of permutations used is 9,999 and six values were found to be greater than or equal to the value of Moran's I for the sample under study. The significance for this example would then be computed as:

$$\alpha = \frac{6+1}{9,999+1} = \frac{7}{10,000} = 0.0007$$

In Figure 3.1 the blue bar represents the observed value of Moran's I and the values to the right those that exceed this observed value. The existence of a large number of values to the right of the blue bar is an indication of a low significance in the observed value, meaning that the data set under study is not showing global spatial autocorrelation.

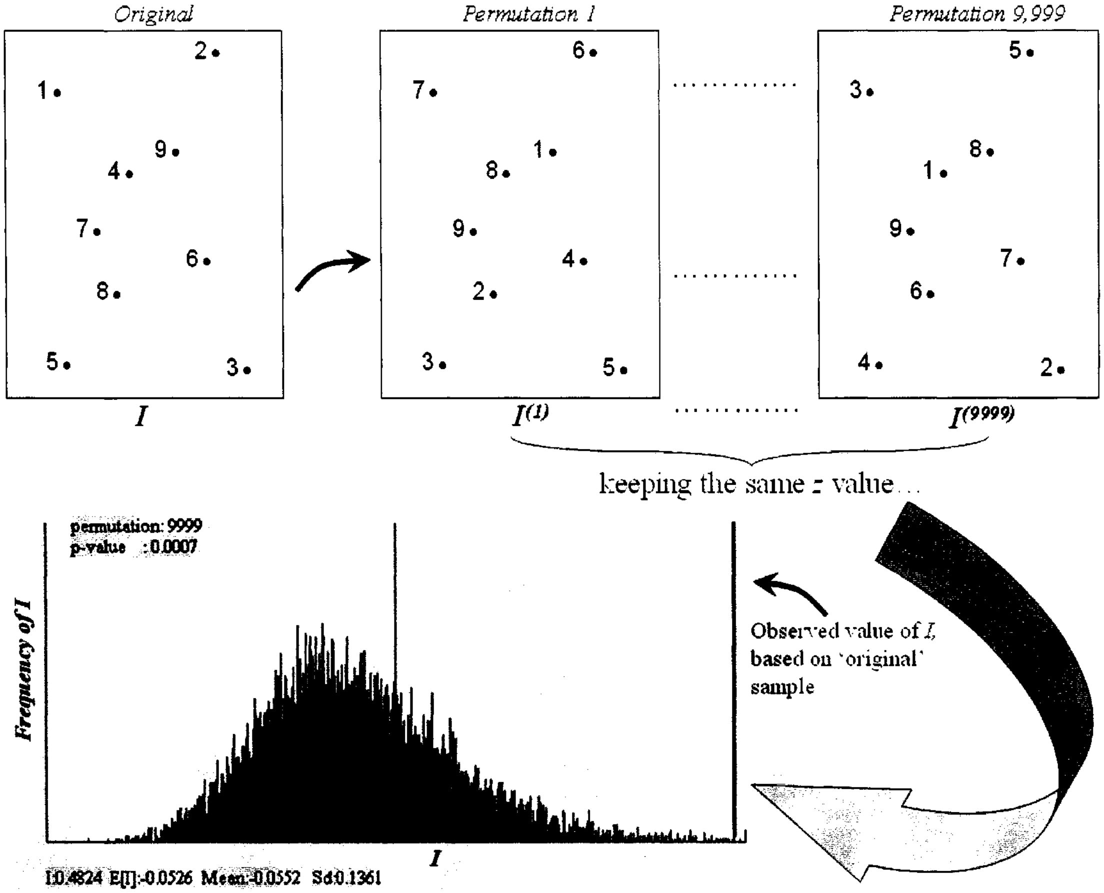


Figure 3.1: Permutations approach used in GeoDa to compute the significance of Global Moran's I. The terms $I^{(1)}, \ldots, I^{(9,999)}$ refer to the value of Moran's I for permutations 1 through 9,999.

3.4.2 Testing significance for Local Moran's I

Statistical significance for I_i is also evaluated in GeoDa and follows the same principles as for Global I, that is, a permutation approach. In this case, however, conditional permutations are carried out within the neighborhood of each point. This means that the value at each point remains fixed, acting as a pivot, and the values at the rest of its neighbors undergo the permutations. GeoDa directly computes the significance level the same way it does for the case of Global I and attaches this value to the point. Figure 3.2 shows an illustration of this process. An observed value of I_i is calculated based on the 'original' configuration of point i and its neighbors. The value at location i remains fixed while intensity values of its neighbors are permutated to calculate the significance of I_i. GeoDa does not display the reference distribution generated for the case of Local Moran's I and it is only shown for the illustrative purpose of understanding the way GeoDa calculates statistical significance for I_i. Significance for Local Moran's I is computed in the same way as for the case of Global Moran's I. In Figure 3.2 the blue bar represents the observed value of I_i and the values to the right are those that exceed this observed value. The existence of a large number of values to the right of the blue bar is an indication of a low significance in the observed value, meaning that there is no significant spatial autocorrelation in the neighborhood around location i.

When 9,999 permutations are used, statistical significance can be classified in four levels: 0.05, 0.01, 0.001 and 0.0001. Each class will hold all those points that have a significance level that is less than or equal to the upper bound of the class (*i.e.*, $\leq 0.05, \leq 0.01$, and so on). This serves the purpose of separating points that are significant from the ones that are not, to further process the valuable information. Once a significance value has been attached to each point in a simulation, this process is repeated for the remaining simulations. After this has been completed every point in every year has a significance value that will allow it to be classified in one or more of the classes mentioned above.

Since there are 100 simulations for any given year, it is possible to determine

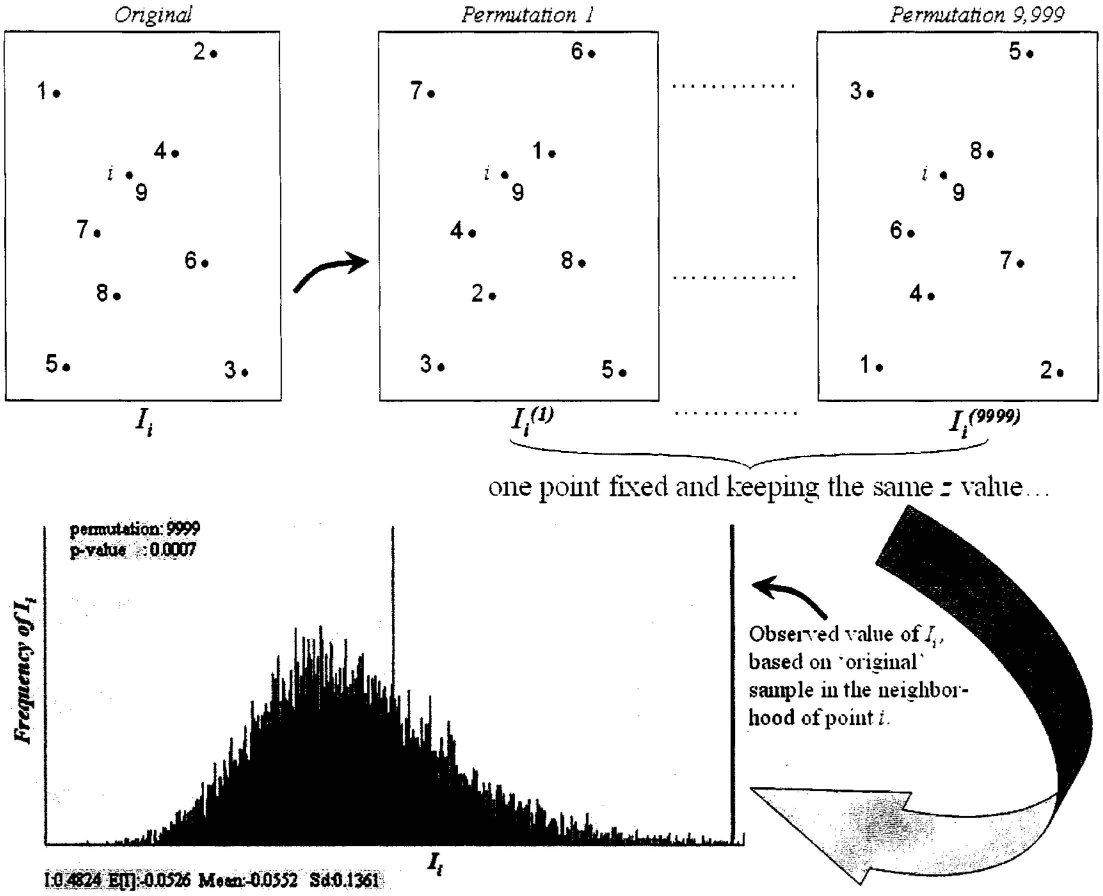

Original
Permutation 1
Permutation 9,999
1 2 3 4 5 6 7 8 9 i
I_i
$I_i^{(1)}$
$I_i^{(9999)}$
one point fixed and keeping the same z value...
permutation: 9999
p-value : 0.0007
Frequency of I_i
I_i
I:0.4824 E[I]:-0.0526 Mean:-0.0552 Sd:0.1361
Observed value of I_i, based on 'original' sample in the neighborhood of point i.

Figure 3.2: Permutations approach used in GeoDa to compute the significance of Local Moran's I. The terms $I_i^{(1)}, \ldots, I_i^{(9,999)}$ refer to the value of Local Moran's I that is evaluated for point i, for permutations 1 through 9,999.

how many times a point is significant at a certain significance level (*e.g.*, 0.05, 0.01, 0.001), out of a hundred; in other words, the significance of a point can be expressed in terms of a percentage.

3.5 Concentration Measurements and Hot Spots

Hot spot detection is carried out using marked points as the data representation and Local Indicators of Spatial Autocorrelation (LISA) as the detection method. As was mentioned before, previous work has been carried out in the same study area using Kernel Density Estimated (KDE) surfaces for these data sets. The driving idea behind this research is to explore if a change in the technique for detection will considerably affect detected hot spots. It will also be interesting to examine the impact of the existence or absence of field data in the results.

For both I and I_i, spatial relationships between points (adjacency matrix) will be defined as distance-based neighborhoods (the term w_{ij} in equations 3.1 - 3.4). More specifically, initially two points will be considered to be neighbors, or adjacent, if the distance between them is less or equal to the minimum distance required for every point in the data set to have at least one neighbor. It is important to note that, typically, adjacency matrices are based on either rook's, queen's or bishop's case (from the way each element moves in a chessboard). However, in this study adjacency is defined in terms of a distance and none of the previously mentioned cases are considered here. As an illustration, if working with a data set consisting of 5 points, an adjacency matrix might look like:

$$\mathbf{W} = (w_{ij}) = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.5}$$

In this example, the adjacency matrix tells the reader that location 1 is adjacent to locations 2 and 5; 2 is adjacent to 1; 3 is adjacent to 4; 4 is adjacent to 3; and 5 is adjacent to 1.

As mentioned before, my analysis will involve two groups of data sets: the 'Adjusted Aerial Data' and the 'Adjusted Aerial-Field Data'. In both cases, Global

Moran's I (equation 3.1) will be used to obtain an indicator of the overall distribution of infested trees. Also, hot spot detection will be carried out using Local Moran's I (equation 3.3) along with a statistical significance criterion.

Once Local Moran's I has been calculated for each point, locations showing a significant value are then classified into one of the four following types of spatial correlation or cluster type: high-high, low-low, for positive spatial autocorrelation and high-low, low-high, for negative spatial autocorrelation. The remaining ones are associated with a cluster type of 'not significant'. Figure 3.3 shows the different cluster types that can be detected using GeoDa, depending on whether the location under study (*Central element* in Figure 3.3) has high or low values and whether its neighbors (*Neighboring elements* in Figure 3.3) have high or low values.

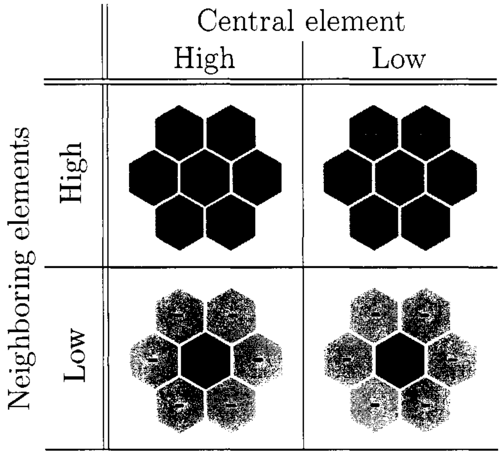


Figure 3.3: Possible cluster types that can be identified for positive and negative spatial autocorrelation.

Attention will be given only to those points with high-high and high-low cluster types, that is, high values surrounded by either high or low values and the rest will be discarded. High-high values suggest clustering of similar high infestation levels while high-low clustering suggests the existence of high values of infested trees surrounded by not-so-high numbers of infested trees, or more localized areas of infestations. In terms of mountain pine beetle and trees high-high points can represent areas of high infestation levels, while high-low ones can indicate locations of an early or late stage of an outbreak.

Once every point in each simulation for each year has a significance value and a cluster type assigned to it, working with this classification scheme allows us to account for the number of times a point is either high-high, high-low or none, at a particular significance level. Therefore it will be possible to say how many times, out of 100, an adjusted point is significantly identified as a hot spot. For analytical purposes, whenever a point is significant at a certain significance level 50 or more times out of the 100 simulations, it will be considered a hot spot.

For data visualization and mapping purposes, data were aggregated by keeping only an average x and y location of the point throughout the whole year; that is to say, for each point all of the x and y coordinates across the 100 simulations were averaged out and only two records were kept ($\bar{x}$ and $\bar{y}$). Information on whether a point is high-high, high-low or none at each significance level is immediately available, but this information was aggregated in order to know what proportion of the times a point was classified as hot (high-xxx).

3.6 Assessing Results

Once the hot spots have been located, the primary issue will be to compare them with those identified by KDE. To do this it is necessary to superimpose the data points on top of the KDE surfaces. This will provide a means of comparing results obtained with each method and data model, and can be done in the following ways:

1. Display detected hot spots as points on top of a KDE surface image and use

them for display as an informal method of visual comparison;

2. Count how many points are contained inside the KDE surface and make the most accurate comparison by means of intersecting the points with the surface.

3. Convert data points to a grid and superimpose this on top of the KDE surface and use map algebra to identify those locations that match.

As was mentioned before, no previous work on comparing the impact of the inclusion of the field data to correct the data sets has been carried out, so it will be important to compare results from both data sets ('Adjusted Aerial' and 'Adjusted Aerial-Field') to determine the importance and usefulness of the field data. It is clear that this task can only be accomplished for those years that have field data available and it will be carried out by studying those locations that, according to LISA, are hot spots. This will provide a way to identify regions in space or individual locations that without the existence of field data would be considered hot spots, when they might be the result of an artifact of the way attribute data has been simulated. It will also provide the possibility of identifying locations that become significant when field data are available, but would otherwise not be classified as such. Should results be similar it will be an indication that field data are not providing a great advantage over pure aerial data, while the opposite will indicate that it is indeed useful to verify the validity of the aerial data by sending crews to the ground to collect data.

A diagram summarizing the hot spot detection process is shown in Figure 3.4. The same procedure is applied to both data sets. When hot spots have been detected for both data sets, a comparison over the years to assess how the mountain pine beetle population has evolved over space and time will be carried out and this will be useful for checking for consistency with hot spots detected in previous years. This idea can be summarized in Figure 3.5.

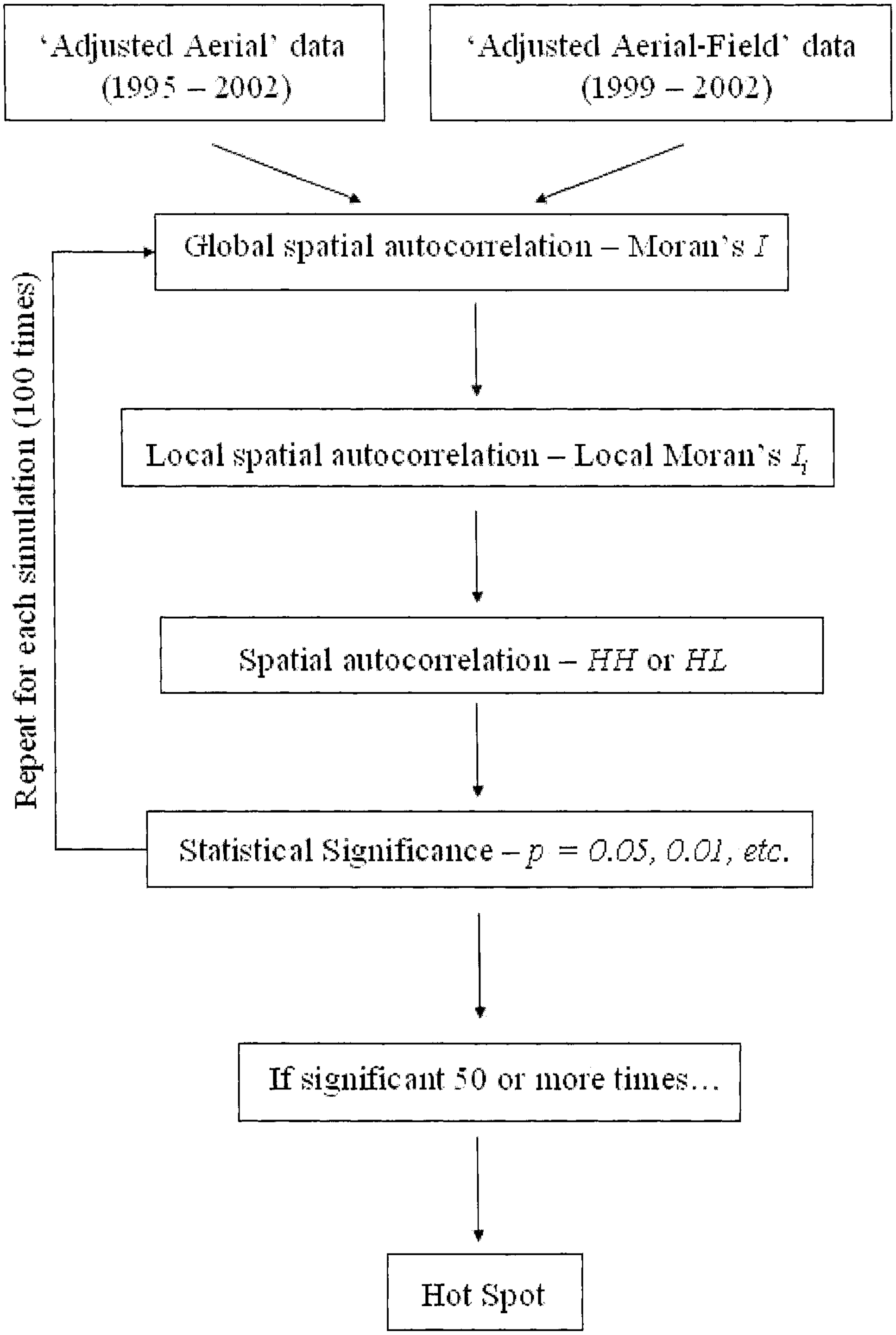

'Adjusted Aerial' data (1995 – 2002)
'Adjusted Aerial-Field' data (1999 – 2002)
Global spatial autocorrelation – Moran's I
Local spatial autocorrelation – Local Moran's I_i
Spatial autocorrelation – HH or HL
Statistical Significance – p = 0.05, 0.01, etc.
Repeat for each simulation (100 times)
If significant 50 or more times...
Hot Spot

Figure 3.4: Hot spots detection approach.

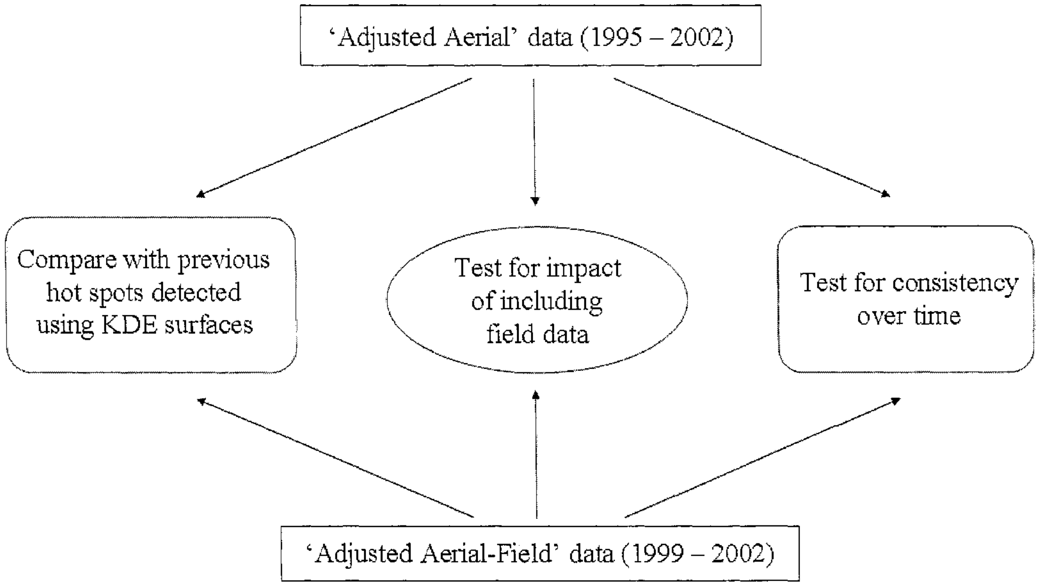

'Adjusted Aerial' data (1995 – 2002)
Compare with previous hot spots detected using KDE surfaces
Test for impact of including field data
Test for consistency over time
'Adjusted Aerial-Field' data (1999 – 2002)

Figure 3.5: Hot spots comparison approach.

3.7 Time Commitments and Computational Effort

As has been mentioned before, in order to compute both Global and Local Moran's I the spatial relationship between points, the adjacency matrix, is defined as distance-based neighborhoods. For this, it is necessary to know the distances between the points in each data set. This calculation is carried out using the GeoDa software package (Anselin, 2004); neighborhoods are constructed based on this information and the weight matrix is built with a custom program which writes this information in the appropriate format to be read by GeoDa.

Within GeoDa both Global and Local Moran's I values are computed, and information on cluster type and significance is also stored in a database file (DBF format), for each simulation for each year. This database file is processed using a custom macro written in Visual Basic for Microsoft Excel. The result is one database (DBF) file, for each year, containing the average location for each point, along with the number of times it is significant for each significance level. DBF files are imported to ArcGIS and hot spot maps are created. An illustration of this process is shown in Figure 3.6.

The most computationally intensive parts of this procedure are the calculations of both Global and Local Moran's I for the different significance levels. Since a permutation approach is used to test for significance, the more strict the significance level is, the more intense it is.

Calculations and data processing were carried out using a Dell Precision 620 workstation, running on an Intel Pentium III XEON double-processor running at 800 MHz, with 512 MB of RAM. In terms of computing time per simulation, distance calculation in GeoDa is quite fast and very well implemented, taking from 5 to 45 seconds, depending on the number of points in each data set. The creation of adjacency matrices based on distance information is fast and efficient taking only a few seconds for each data file. Calculations to obtain the values of Global and Local Moran's I can take anywhere from 10 minutes to close to 1 hour, depending on the number of points of the data file and the significance level. Data processing in Excel is well implemented and takes only a few minutes to complete the summarizing of the information for each year.

Under these circumstances, an estimate of 650-700 hours is required to process the data. For future work involving a similar procedure, it would expected to reduce the amount of time required for data processing by means of using a computer with a faster processor ($\sim$ 3.2 GHz) and more RAM (1 GB).

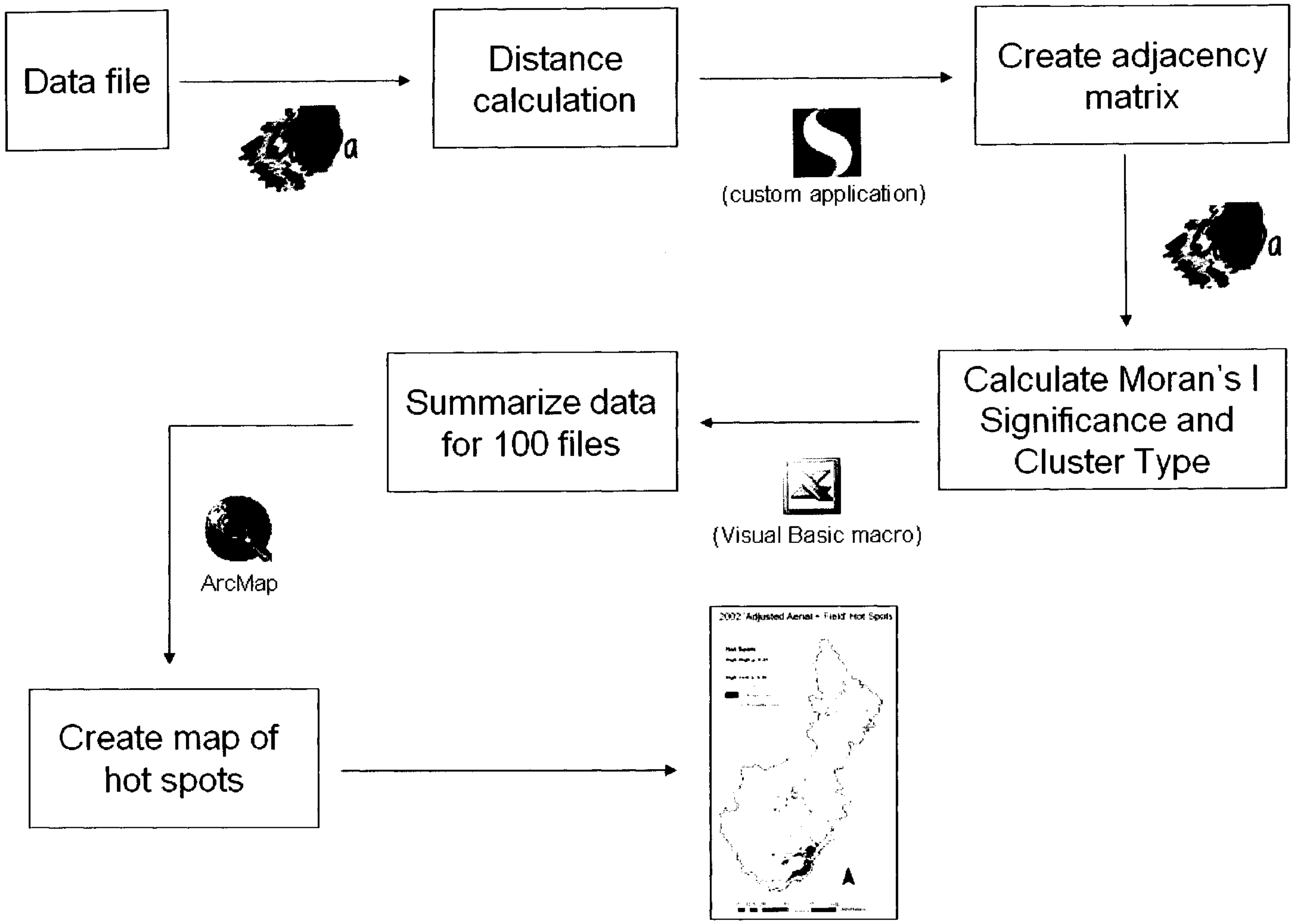


Figure 3.6: Illustration of the computation procedure used in this project for each data set.

§ Chapter 4

Results and Discussion

In order to compute Global and Local Moran's I, adjacency matrices were required and these were obtained in terms of neighboring distances, as explained in the previous chapter. The average minimum, the smallest minimum and the maximum minimum distances, along with the standard deviation, over the 100 simulations, for each year are shown in Table 4.1. From now on, for simplicity, data for years referring to the 'Adjusted Aerial' data sets will be identified as $\langle year\_number\rangle nof$ and those referring to the 'Adjusted Aerial + Field' data set will be identified as $\langle year\_number\rangle f$; for example, $1999nof$ will correspond to the 1999 'Adjusted Aerial' data set, while $2002f$ to the 2002 'Adjusted Aerial + Field' data set.

To get an idea of how neighborhoods have been defined and how many elements are typically taking part in the calculations of Global and Local Moran's I, box plots were created to show the distribution of the number of points in neighborhoods. For this, only a representative simulation was used, since each point from one simulation to the next is displaced within a radius of 25 meters. It can be seen from Table 4.1 that the distances used to define neighborhoods are a few orders of magnitude greater than this displacement. It is therefore not expected for neighborhoods to change a lot across simulations. Table 4.2 shows the median and maximum number of neighbors for a representative of each year.

Year	Average Distance (m)	Minimum Distance (m)	Maximum Distance (m)	Standard Deviation
1995	4,101.4803	4,062.8649	4,146.2723	15.9899
1996	6,265.1188	6,215.0356	6,301.3098	17.5075
1997	7,504.3631	7,469.4766	7,538.1071	15.0069
1998	10,591.4601	10,549.0901	10,625.4311	15.6897
1999*nof*	9,860.5717	9,827.4465	9,899.5218	16.9770
1999*f*	9,863.2975	9,812.0763	9,907.8270	19.2941
2000*nof*	9,467.4395	9,429.2555	9,523.8424	16.7467
2000*f*	9,466.1554	9,420.6375	9,515.7056	18.3779
2001*nof*	6,769.9002	6,725.4676	6,804.8840	16.4005
2001*f*	8,069.7623	8,017.0162	8,108.7347	17.3147
2002*nof*	6,945.6253	6,906.7593	6,990.8036	17.4485
2002*f*	6,945.8614	6,907.8385	6,984.8805	16.4474

Table 4.1: Average, minimum and maximum distances, over 100 simulations, used for defining adjacency matrices for each year. Standard Deviation is also shown. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

Year	Median number of neighbors	Maximum number of neighbors
1995	35	157
1996	149	551
1997	238	820
1998	119	386
1999*nof*	206	507
1999*f*	205	504
2000*nof*	201	650
2000*f*	201	651
2001*nof*	114	461
2001*f*	151	572
2002*nof*	238	860
2002*f*	238	859

Table 4.2: Median and maximum number of points per neighborhood, per year. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

Figure 4.1 shows the distribution of the number of neighbors for each year, using box plots. The filled circle inside the box represents the median, the upper and lower edges of the box represent the upper and lower quartiles and the distance between these two is a measure of the spread of the distribution. The location of the median with respect to the upper and lower edges of the box gives an indication about the shape and skewness of the distribution. The appendages of the box, or whiskers, give an indication of the spread and the shape of the tails of the distribution. Outer circles represent data outliers, that is, unusually large or small observations that are not included in the summarizing process of the box plot.

From this Figure it can be seen there is variability in the number of points found in each neighborhood, however, the median seems to fluctuate around a value of 200, the exception being 1995 which has a much lower median. The distribution of points per neighborhood appears to be skewed all years and most of them also have long tails towards the higher values up to a maximum of about 800 points and with a lower bound of 1 point per neighborhood.

An average of Global Moran's I was obtained for each year to give an idea of the general trend of the point pattern throughout the years. Also histograms were produced to show the distribution of Global Moran's I throughout the years and are shown in Figures 4.2 - 4.4. The average value of Moran's I is shown for the year in the top right corner of each histogram. The number of bars for the histograms was chosen according to the guidelines suggested in Doane (1976), Terrell and Scott (1985) and Scott (1979), that propose corrections to the 'optimal' number of classes defined by Sturges (1926) to be $K = 1 + \log_2(N)$, where N is the number of elements. Following these suggestions, all of the years except one, had an optimal number of classes of about eight; for consistency, histograms for all years were produced using eight classes or bins. It is worth noting that due to the nature of the data for each year, histograms are not drawn to the same scale on the x-axis. Each histogram is drawn on its own range thus care should be taken when comparing results across years, especially in those years that have both aerial and field data available. (This limitation is partially accommodated by the use of box plots, presented further in Figure 4.7.)

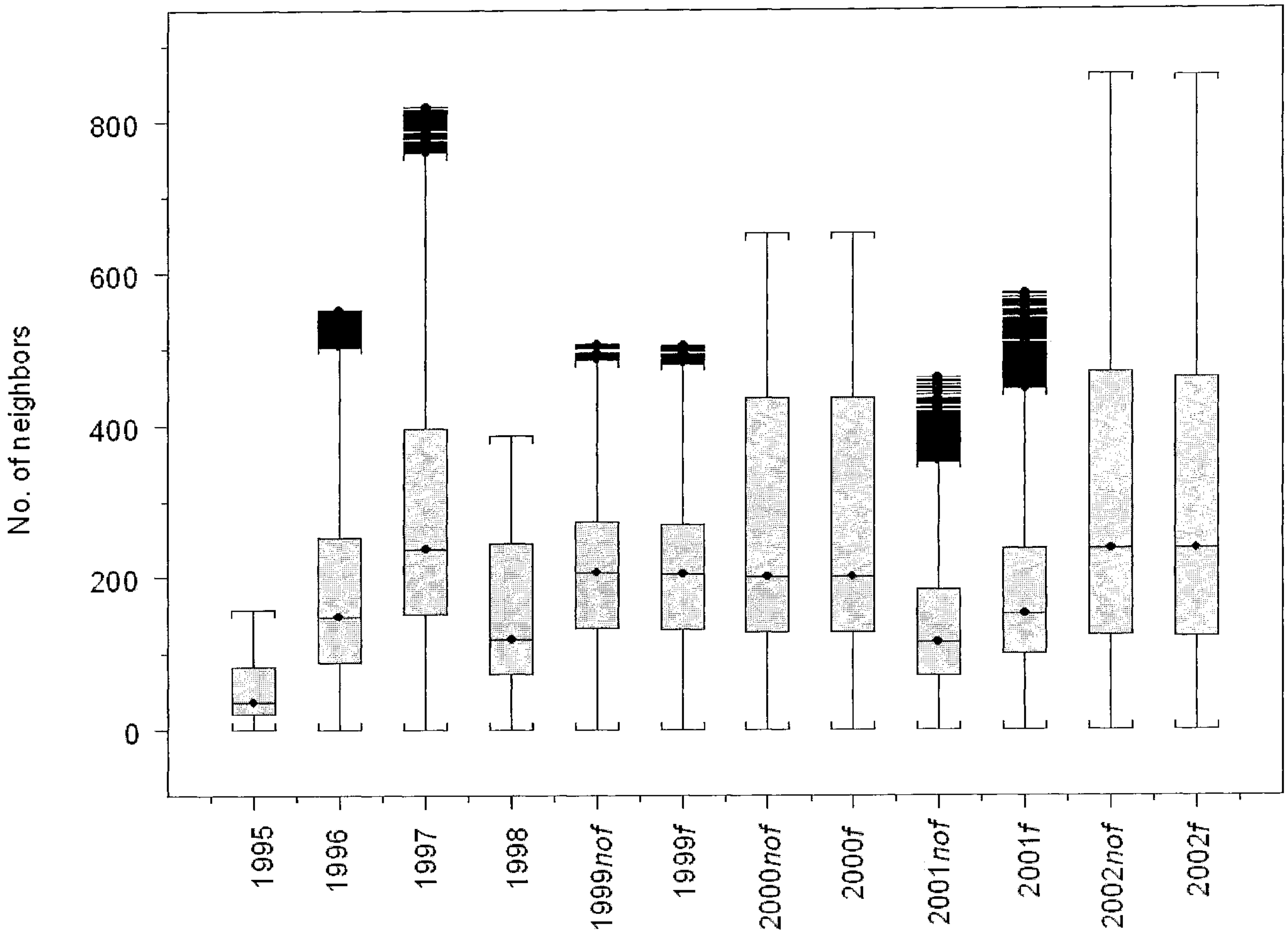


Figure 4.1: Box plots for neighbor distribution, from 1995 to 2002. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

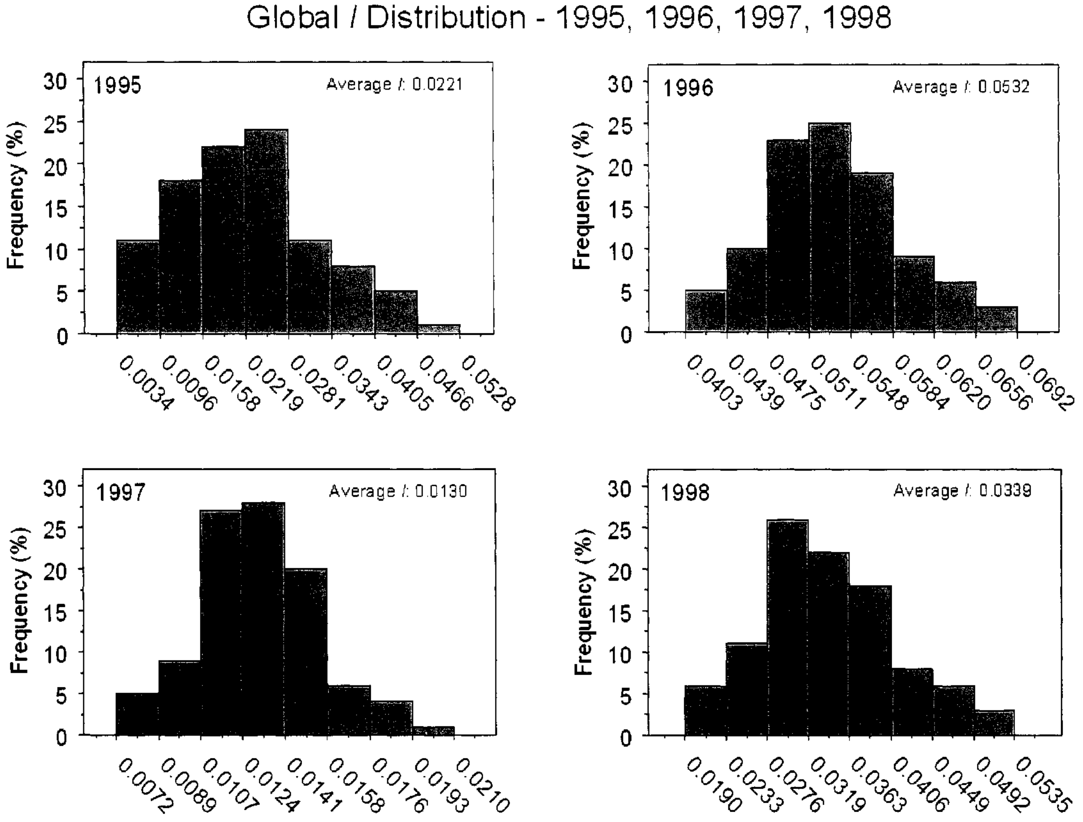


Figure 4.2: Moran's I distribution from 1995 to 1998.

A closer inspection of Figures 4.3 and 4.4 reveals the existence of certain differences in the inclusion of field data in the analysis. For 1999 and 2000 there is very little change in the average value of Moran's I, considering that inclusion of field data reduces this value from 0.0135 to 0.0124 for 1999, and from 0.0102 to 0.0093 for 2000. These are not significant changes and is not expected that these years would show such differences since there are very few points with field data collected (see Table 2.1). However, for 2001 and 2002 the values change considerably more with the inclusion of field data. For 2001 it changes from 0.0251 to 0.0376 and from 0.0195 to 0.0342 for 2002. Based on results from the years that have substantial field data collected (2001 and 2002), it is possible to say that global spatial autocorrelation is greater when field data are included. This behavior, however, is not observable for 1999 and 2000 since these years have very little field data available.

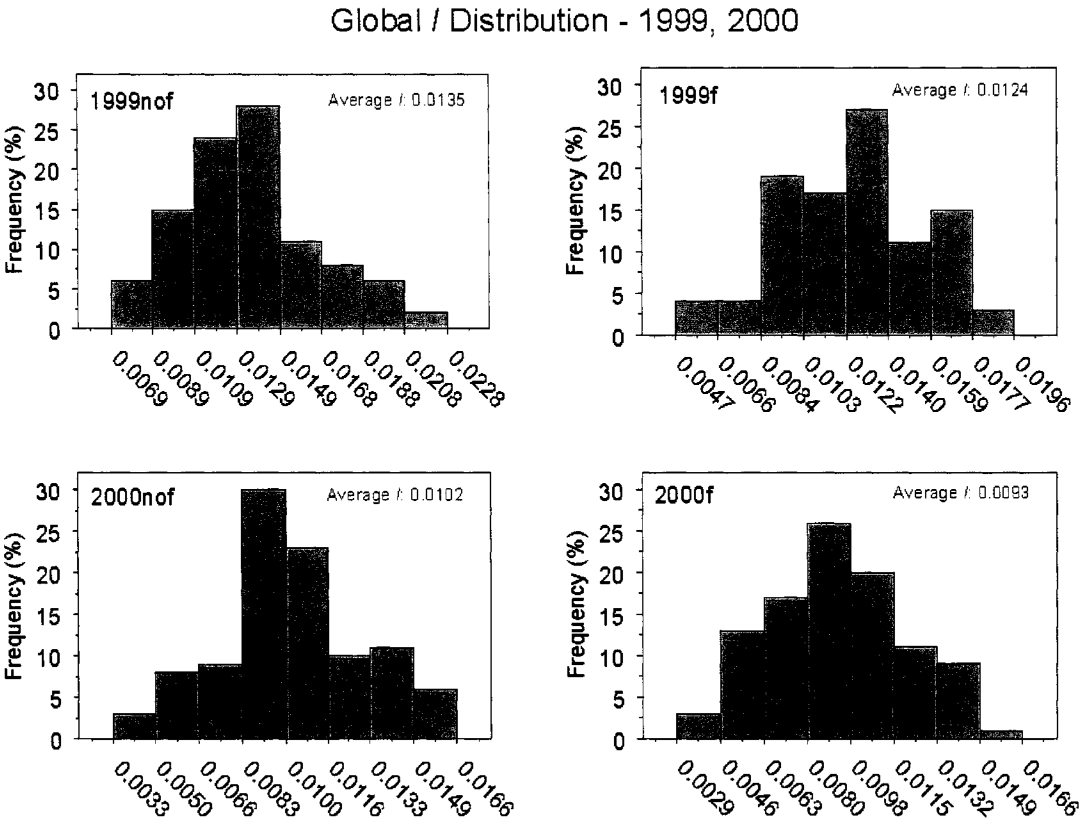


Figure 4.3: Moran's I distribution for 1999 and 2000.

It is interesting to note that, regardless of the inclusion of field data in the analyses, the averages obtained for Moran's I throughout the years seem to be quite low, compared to bounding values of the distribution (Bailey and Gatrell, 1995; Tiefelsdorf and Boots, 1995), since in general values closer to 1 would indicate strong positive spatial autocorrelation and closer to -1 strong negative spatial autocorrelation. However, this does not necessarily imply that there is no strong global spatial autocorrelation, since it is possible to get an idea of how significant the observed values of Moran's I are with respect to the expected value.

As it was explained in section 3.4.1, it is possible to generate a reference distribution for the data sets and test for significance. As was mentioned before, tests for spatial autocorrelation typically have a null hypothesis of no spatial autocorrelation, or spatial randomness, and some degree of structural stability across the space. Under these

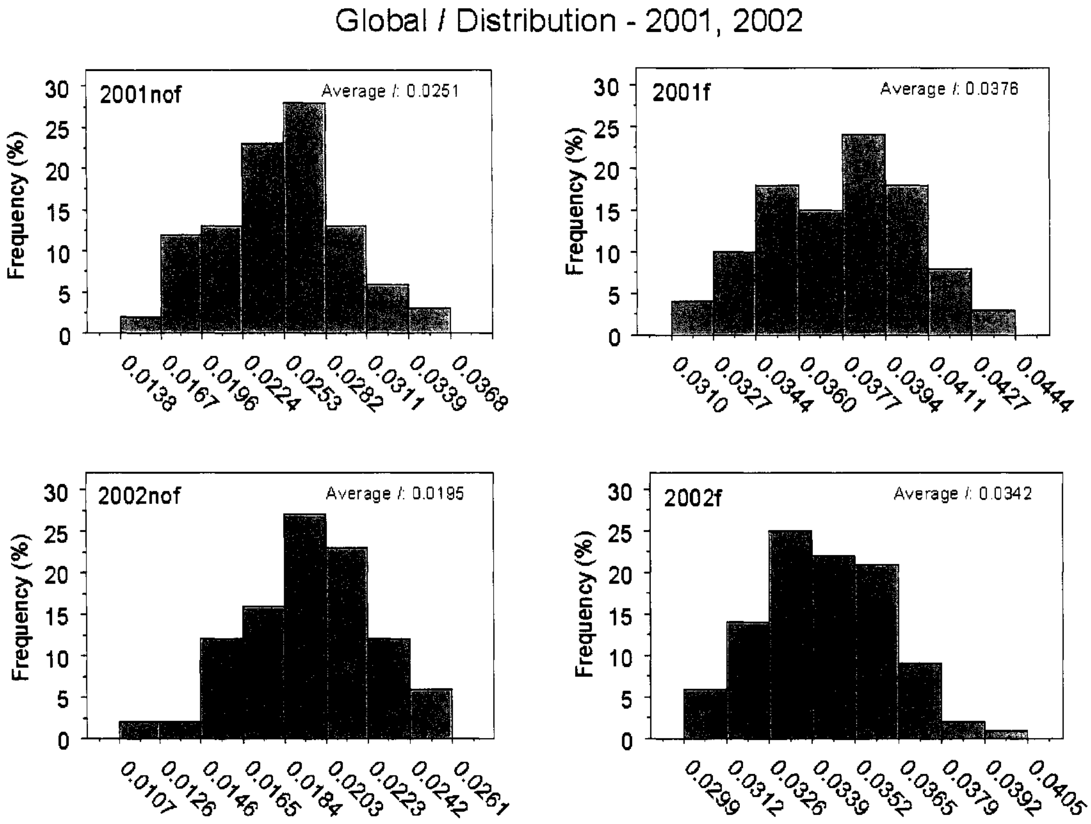


Figure 4.4: Moran's I distribution for 2001 and 2002.

assumptions, Moran's I has a near normal distribution with the following parameters (Bailey and Gatrell, 1995):

$$E(I) = \frac{-1}{(n-1)} \tag{4.1}$$

$$VAR(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2S_0^2}, \tag{4.2}$$

with the coefficients given by:

$$S_0 = \sum_{i \neq j}\sum w_{ij} \tag{4.3}$$

$$S_1 = \frac{1}{2}\sum_{i \neq j}\sum (w_{ij} + w_{ji})^2 \tag{4.4}$$

$$S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2, \tag{4.5}$$

where:

- i, j, k range from 1 to n, the number of data points, and
- w_{ij} are the elements of the weights matrix.

When these conditions hold, it is possible to test the observed values of Moran's I. A z-score can then be calculated with:

$$z \simeq \frac{I - E(I)}{\sqrt{VAR(I)}} , \tag{4.6}$$

and the 95% confidence interval (with $z \approx \pm 1.96$) can be used to search for 'extreme' values that would give an indication of strong spatial autocorrelation. However, throughout this research it has been recognized that positive spatial autocorrelation is expected (section 3.4.2), therefore the approach that has just been presented is not entirely suitable for this case. Nevertheless, it is possible to shed some light in this direction by making use of the same procedure outlined in section 3.4.1 to evaluate if the observed values of Moran's I are significant so it is possible to tell if there is, in fact, global spatial autocorrelation, and if there is, how pervasive it is throughout the years.

An exploratory approach was taken in order to fulfill this task. For each histogram, a representative from the highest bar is selected and tested to obtain a value for its significance level. If this turns out to be greater than or equal to 0.05 – that is, significant at the 95% interval – a representative of the previous class is selected and the same procedure applied until a non-significant class is found. If it is less or equal – that is, not significative at the 95% interval – a representative of the next higher class is tested for significance until a significant class is found. This exploratory technique allows to find an interval where Moran's I is significant for each year. A diagram of this procedure is presented in Figure 4.5.

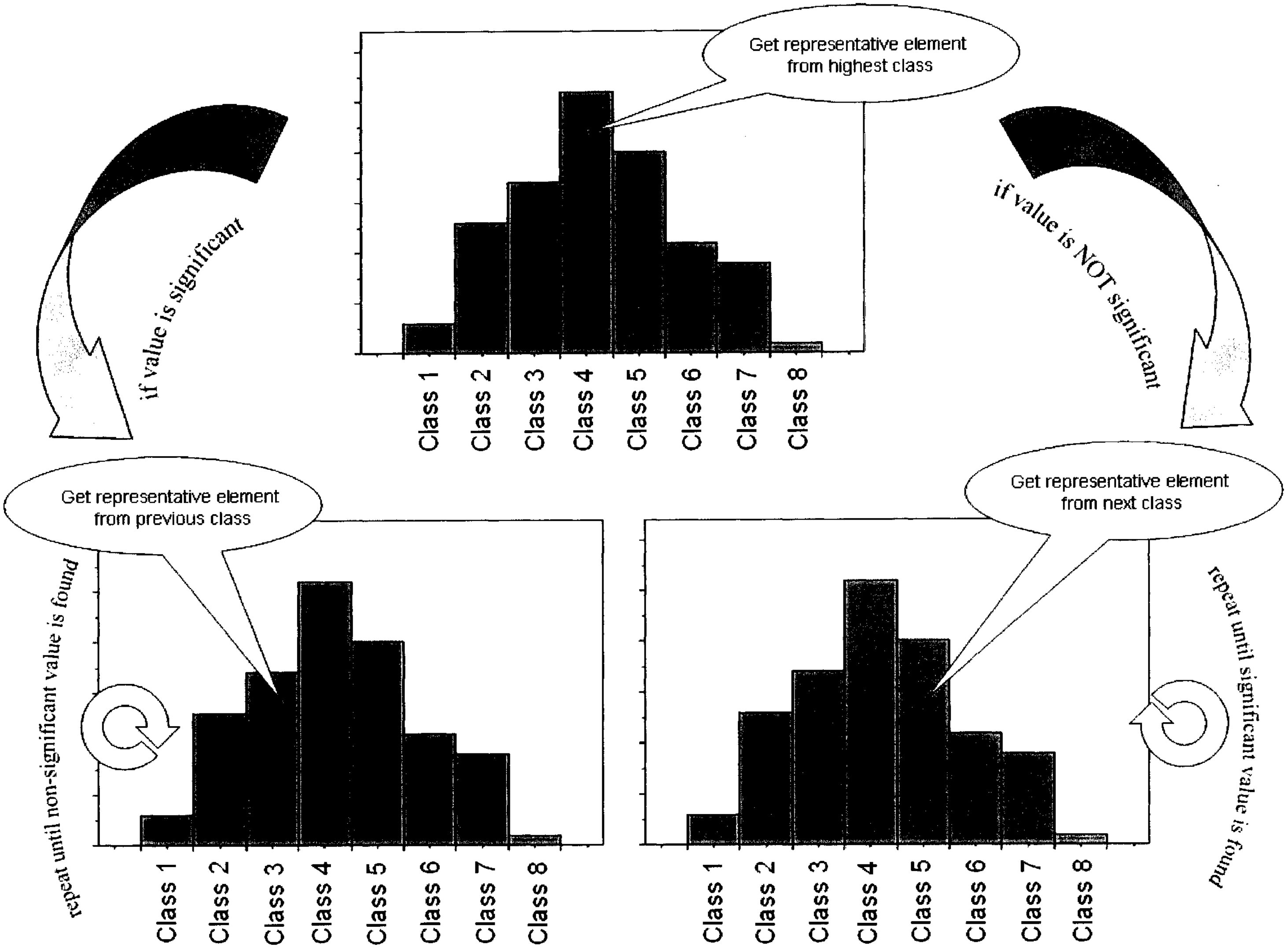

Get representative element from highest class
if value is significant
if value is NOT significant
Get representative element from previous class
Get representative element from next class
repeat until non-significant value is found
repeat until significant value is found
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8

Figure 4.5: Illustration of the procedure to identify values that have a significant Moran's I value.

Results for this exploratory analysis indicate that for all the years, values from the second bin onwards are significant. For 1995 and 2000 values contained in the first bin are not always significant but typical values for Moran's I values found in these bins are very low compared to the ones around the mean of the distribution (typically bin 4 of the histograms). Having said this, it is safe to claim that the observed values of Moran's I are in fact significant and therefore the data are autocorrelated at a global level. A graph showing the variations of Moran's I throughout the years and the impact of field data is shown in Figure 4.6.

Referring back to Table 2.1 it can be seen that from 1995 to 1998 no field data were collected and Moran's I has only one average value for each year. For 1999 and 2000 there are very few data collected and this is a reflected in the average values of Moran's I, since there does not appear to be much change. For the last two years, 2001 and 2002, there is a clear difference in Moran's I value indicating that the 'Aerial Adjusted + Field' data sets are showing stronger global spatial autocorrelation.

Ideally it would be desirable to create a diagram or a table containing information relating the "change" of Local Moran's I throughout the years. Unfortunately it is not clear how to compare the values of I_i at a specific site, j, in two different time periods, thus making it very difficult to go further in this direction (Tiefelsdorf, 2004). A better way to explore and compare the distribution of Moran's I throughout the years is by using box plots (Cleveland, 1993). They are useful in comparing all of the distributions with one another. Figure 4.7 shows box plots for the values of Moran's I for all years.

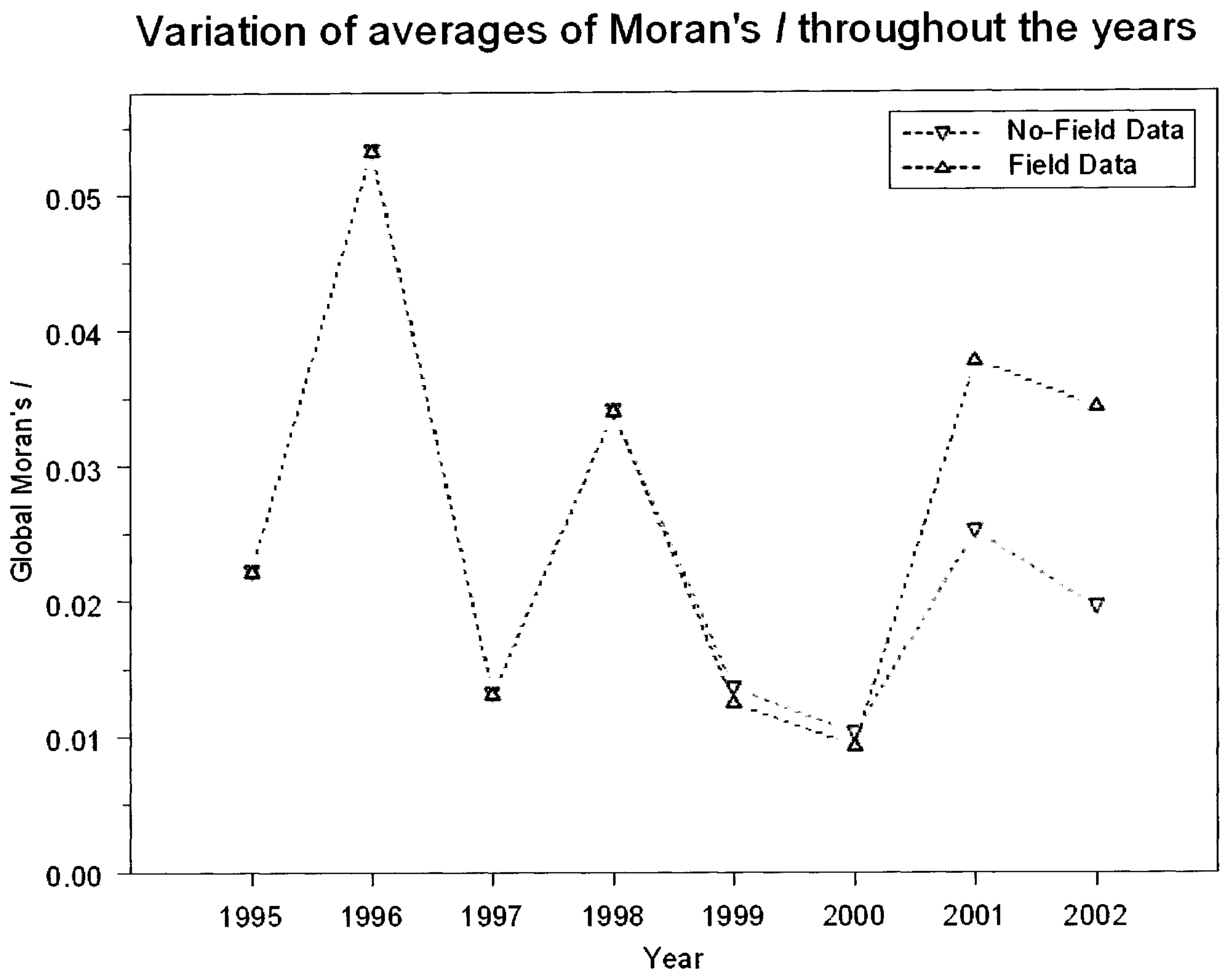


Figure 4.6: Moran's I variation throughout the years. Orange lines and downward triangular symbols are for years that have no field data collected ('Aerial Adjusted' data sets); green lines and upward triangular symbols are for years that have field data collected ('Aerial Adjusted + Field' data sets).

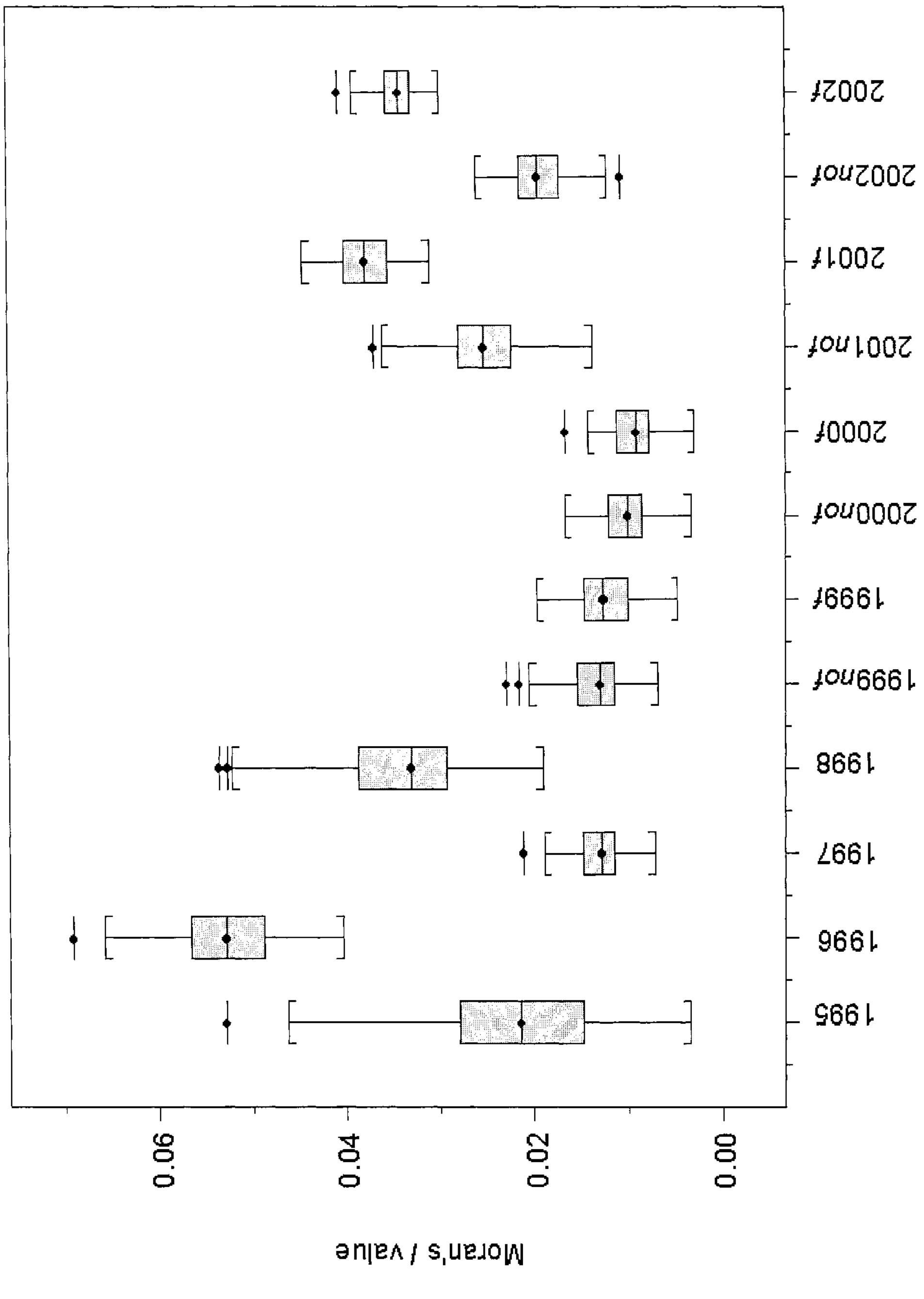


Figure 4.7: Box plots for Moran's I values, from 1995 to 2002. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

It can be seen from the box plots that the distributions are highly contained, except for 1995, which shows a much greater spread than the others, followed by 1998. All of the distributions seem to be skewed to either one side or the other and some, like 1998 or both in 1999, are clearly skewed. In some cases, however, it is hard to determine the direction in which they are skewed, such as 1995, 1996 and both in 2000, for example, which might suggest that in these cases there is a tendency for the distributions to approach a normal distribution. For three of the first four years the distributions have long tails. This behavior does not seem to hold for the next years as the appendages do not reach too far out from the box. Most of the distributions have outliers, the exceptions being $1999f$, $2000nof$ and $2001f$, but typically these appear to be close to the whiskers, except for the first couple of years.

The description of the box plots can be related to the spatial pattern of infested trees in the following way: those years that have a low Moran's I value show a lesser degree of global spatial autocorrelation than the others. It also portrays that the inclusion of field data is affecting the global distribution of infested trees as there is a clear difference in Moran's I value for 2001 and 2002. Box plots for 1995 and 1998 have large whiskers and larger boxes, suggesting that there is a wider variety of Moran's I values throughout the simulations that can be translated into a more diverse behavior in each one of the data sets simulated, in terms of the observed spatial pattern and overall clustering. Other years have a more compact distribution suggesting that the data sets show a similar behavior across the 100 simulations, meaning that there is a lot less variation in their overall clustering pattern.

As it was mentioned in the previous chapter, different levels of significance are used in order to have an informal method to test for stability of results. The results presented from now on in this section will be constrained to the lowest significance level, or the most liberal one, of $\alpha = 0.05$. It is important to note two things:

1. It has been shown before that autocorrelation is present in the data sets so care needs to be given to the importance assigned to detected hot spots, since it is plausible that certain locations have been marked as hot spots and may not actually be presenting high levels of infestation; that is, they might be 'false positives'.

2. There is a way to overcome this limitation of the method by taking into account the remaining significance levels of $\alpha = 0.01, 0.001$ and 0.0001. Due to time constraints it is not possible to carry out all of the analyses required and present the results for all significance levels. However, it is important to note that these remaining significance levels contain valuable information that can help assess in a more definite way the locations of hot spots, and remain as a subset of the analysis that is carried out here.

Since a hot spot, in this study, is understood to be a point that appears to be significant 50 or more times out of the 100 simulations, it is necessary to report that for 1995 only one high-low hot spot was found. This led to the exclusion of 1995 from further analyses and is shown in Figure 4.8 for illustrative purposes only. This provides the first important result, as this was not expected and contrasts with previous findings (Nelson, 2005) that showed that all of the years had several hot spots. This also provides clear evidence of the difference between detection methods used, since in this case it is not necessary to have a large number of hot spots, while in previous studies there always are. It is useful to remember why this is so: as it was mentioned in section 3.2, in Nelson (2005) hot spots were defined in terms of the relative upper 10% threshold of intensity values of KDE surfaces, thus forcing the existence of hot spots.

The rest of the years do show pockets of infestation and it is interesting to note that all of them show a distinction between locations of high-high (red dots) and high-low (blue dots) cluster types, that is, they are not intermixed. Figures 4.8 - 4.13 show maps depicting the location of these pockets of infestation. In these figures beige-colored regions show regions of infestation, red/brown areas indicate the regions that were designated as hot spot areas by means of KDE. These results are the same that were presented in Nelson (2005). In the same figures, red- and blue-colored points represent the locations of hot spots detected with LISA. Red dots indicate the location of high-high values and blue dots show high-low values (see Figure 3.3).

It is important to remember that infested areas and hot spots used from Nelson (2005) from 1995 to 1998 are based on the 'Adjusted Aerial' data sets, while from 1999 to 2002 they are based on the 'Adjusted Aerial-Field' data sets. This means that in this research comparison is being made in the following fashion:

- From 1995 to 1998, only 'Adjusted Aerial' data sets exist and comparison is carried out using these.
- From 1999 to 2002, only 'Adjusted Aerial + Field' hot spot surfaces exist, and comparison for both data sets ('Adjusted Aerial' and 'Adjusted Aerial-Field') are carried out against these existing surfaces.

For the maps of hot spots of infestations, an ordered sequence of appearances was expected, such as: $HL \rightarrow HH \rightarrow HL$, that could represent the initial stages of an outbreak, the peak of the infestation and then a late stage of the outbreak. This is not observed and it is possible that the definition of a hot spot itself is the cause of this, since when using both the statistical and the concentration criteria lots of points are stripped away from the final data set used to create the maps. If more points are kept maybe this behavior will show up. There also might be a relationship of hot spots with underlying environmental characteristics that do not allow for this behavior to be shown as expected.

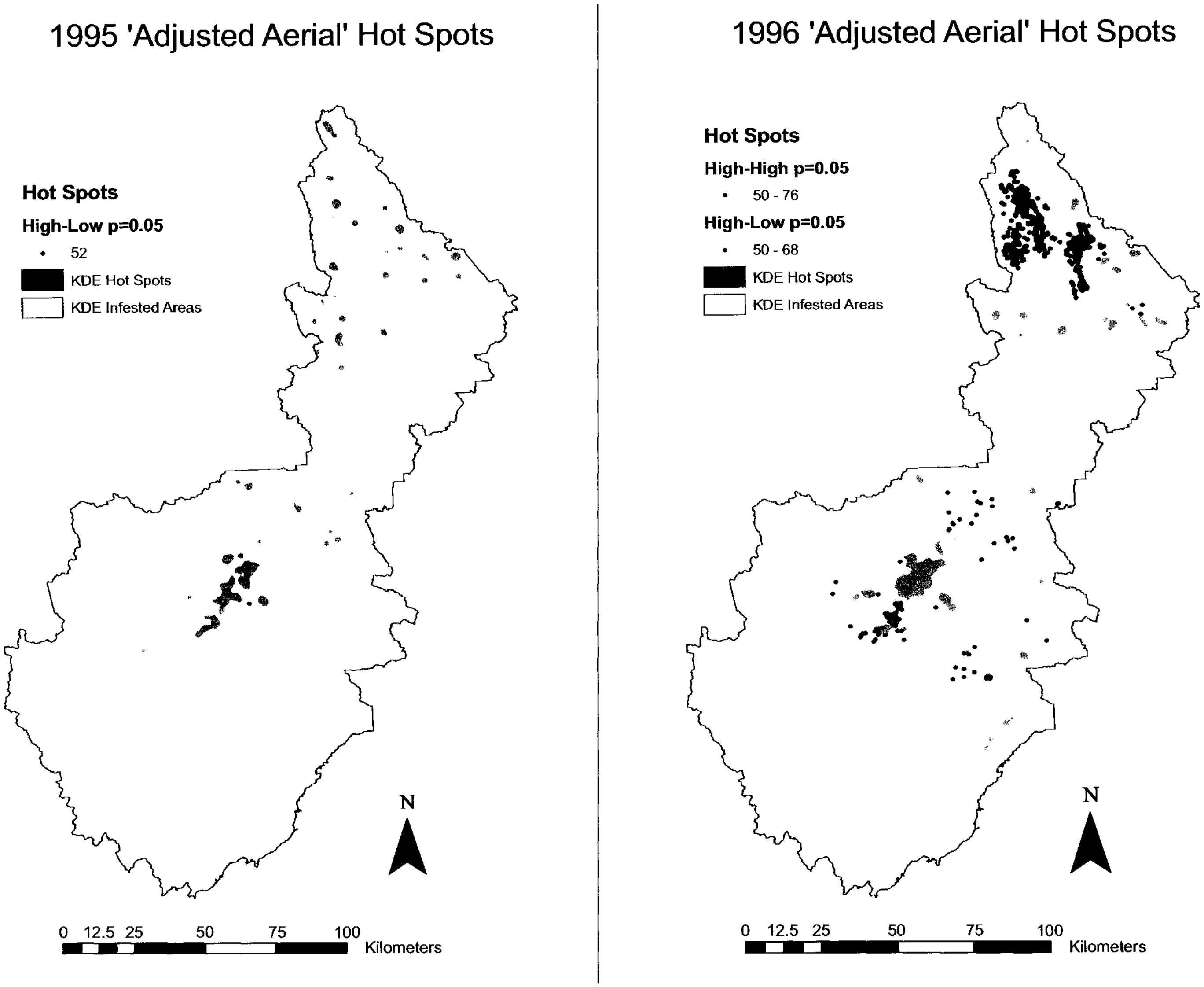


Figure 4.8: Significant Hot Spots for 1995 and 1996.

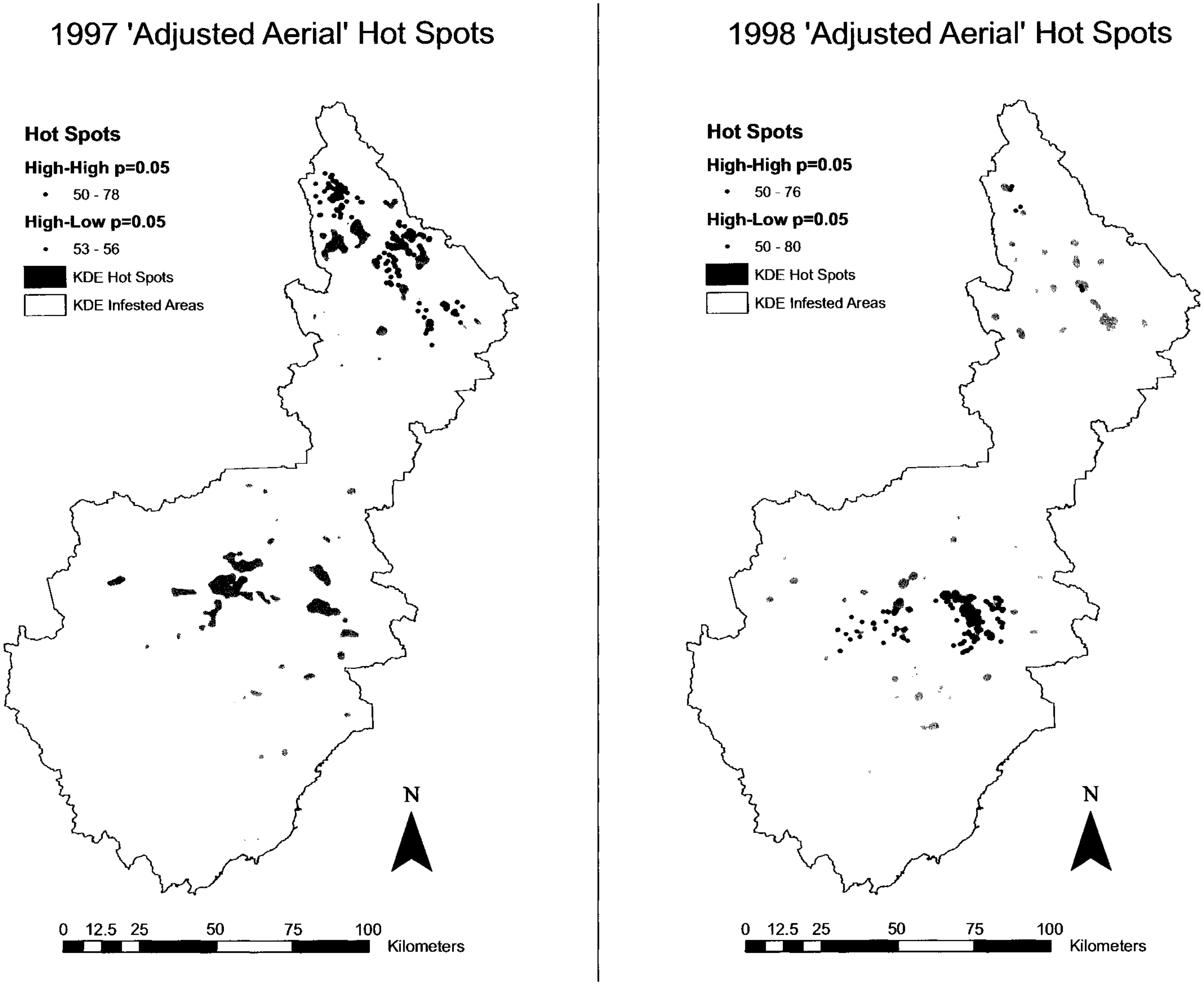


Figure 4.9: Significant Hot Spots for 1997 and 1998.

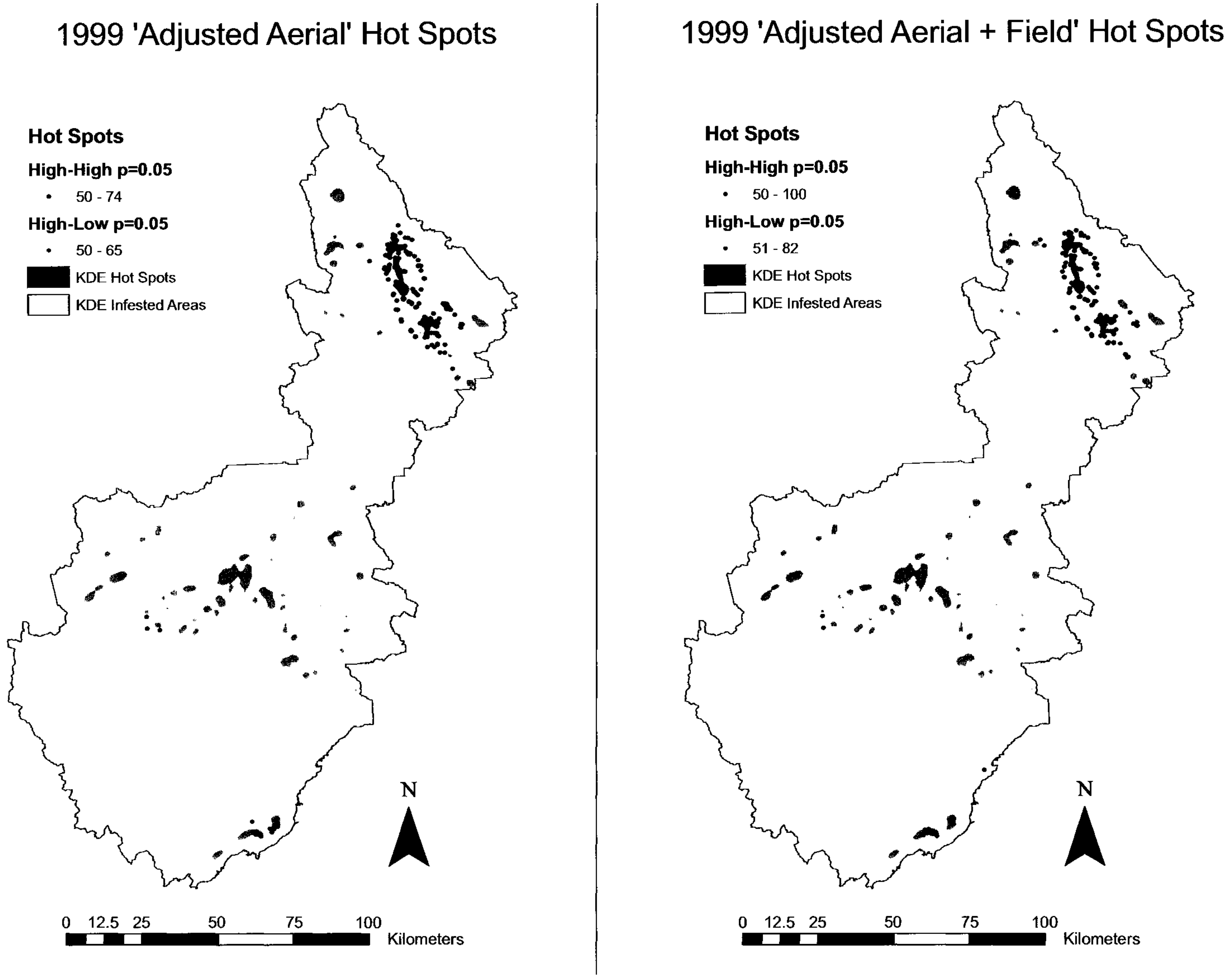


Figure 4.10: Significant Hot Spots for 1999.

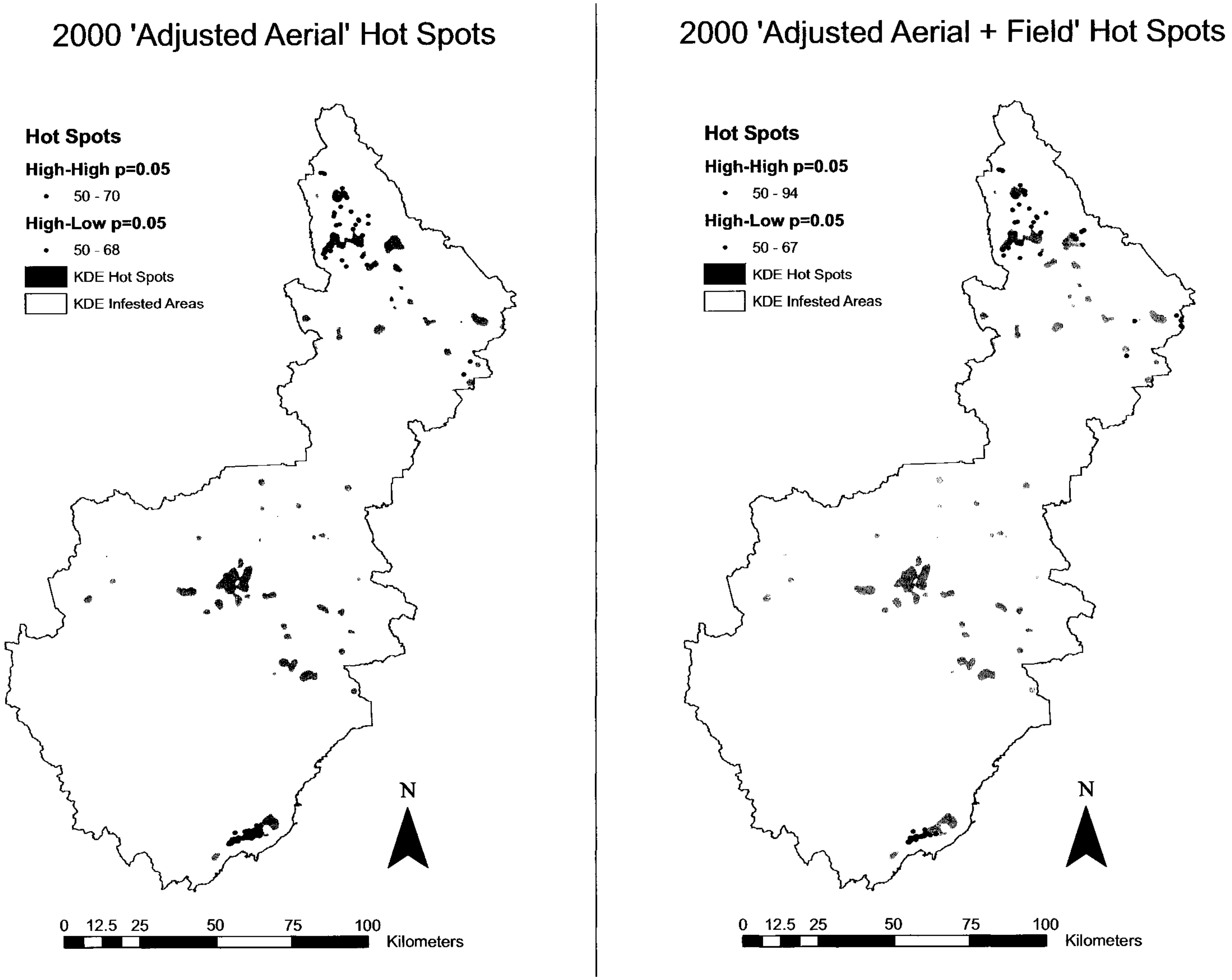


Figure 4.11: Significant Hot Spots for 2000.

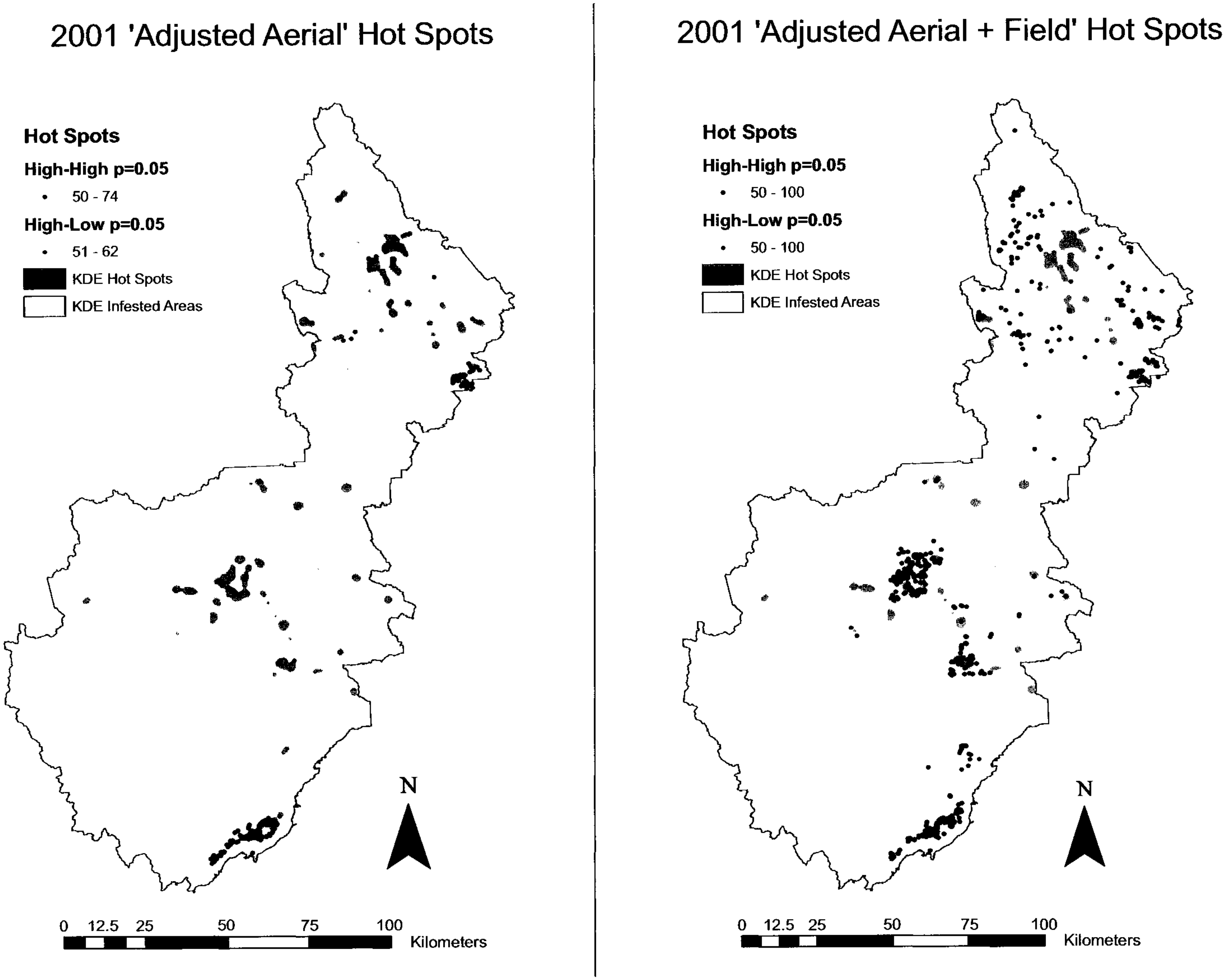


Figure 4.12: Significant Hot Spots for 2001.

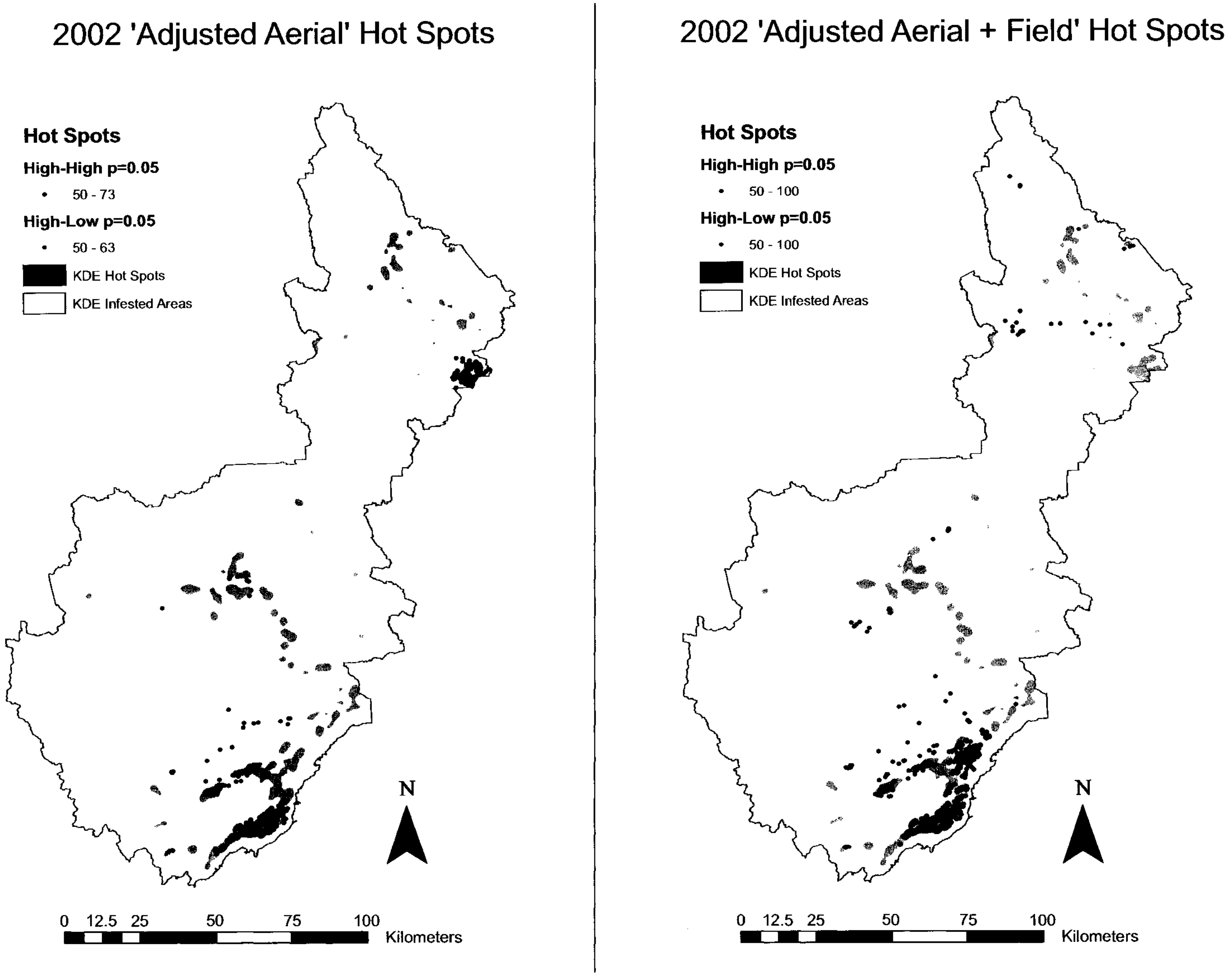


Figure 4.13: Significant Hot Spots for 2002.

The next procedural step is to compare detected hot spots with those obtained in previous works. Table 4.3 shows the number of detected hot spots that lie within those detected and reported in Nelson (2005). These results are shown as the number of significant hot spots, LISA HH or HL, that were classified as KDE hot spots. In other words, a point was either a high-high hot spot within the LISA framework as well as in the KDE one (LISA HH - KDE HS), or appeared to be high-high only with LISA, or was a high-low hot spot in LISA but high-high in KDE (LISA HL - KDE HS) or was high-low only in LISA. The number of occurrences were obtained by intersecting the high-high and high-low points with each year's hot spots surface and counting the number of points that are completely inside the surface. To make it easier to interpret these results, they are also presented as percentages obtained from the ratio between the number of corresponding and significant hot spots for both HH and HL.

Year	LISA HH	LISA HH - KDE	LISA HH - KDE (%)	LISA HL	LISA HL - KDE	LISA HL - KDE (%)
1996	687	487	70.89	92	47	51.09
1997	268	180	67.16	3	3	100.00
1998	228	211	92.54	41	21	51.22
1999*nof*	266	178	66.92	22	16	72.73
1999*f*	248	166	66.94	20	17	85.00
2000*nof*	88	79	89.77	49	27	55.10
2000*f*	40	30	75.00	42	25	59.52
2001*nof*	207	161	77.78	5	1	20.00
2001*f*	446	345	77.35	169	61	36.09
2002*nof*	499	465	93.19	50	25	50.00
2002*f*	646	545	84.37	119	44	36.97

Table 4.3: Number of LISA high-high and high-low hot spots that match KDE hot spots and infestation areas from 1996 to 2002. Results show the total number of *LISA HH* and *HL* points and the number of significant *LISA* hot spots that were classified as *KDE HS* and the associated percentage. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

It can be seen from Table 4.3 that there is a high percentage of coincidence in the high-high results, at least two thirds of the detected hot spots (above 66%) agree with those detected with KDE in all years. These findings indicate that those locations coinciding to be hot spots by both methods are very likely to be troublesome areas. For some years there is also a good correspondence between high-low spots with KDE hot spots, although there is more variability.

This can be understood in terms of the way KDE works when smoothing data. An illustration of points being high-high within LISA as well as with KDE is shown in Figure 4.14a). In this case, high values are surrounded by other high values and by smoothing the data, by means of KDE, a surface with high values is obtained. The case where a point is high-high within LISA but not a hot spot according to KDE is something similar to this illustration, but with surface values that are lower and do not reach the specified threshold set in order to be classified as a hot spot. If a point is identified to be high-low in LISA but hot spot in KDE, the KDE method has smoothed the data and returned a surface with high values. However, according to LISA, high values are surrounded by low ones. Figure 4.14b) shows an illustration of this case. Finally, when a point is high-low according to LISA, but not a hot spot according to KDE, what may be happening is that low and high values are smoothed out and the resulting surface values are not high enough in order to be classified as a hot spot, but it is detected as a high-low location by LISA. This is illustrated in Figure 4.14c).

When comparing the number of hot spots to total infested area, in terms of KDE, it is expected to find a ratio of 10%; this is a consequence of the relative threshold used (Section 3.2). It is interesting to compute this ratio for the LISA method by dividing the number of points that were classified as hot spots by the total number of points for each data set. These results are shown in Table 4.4.

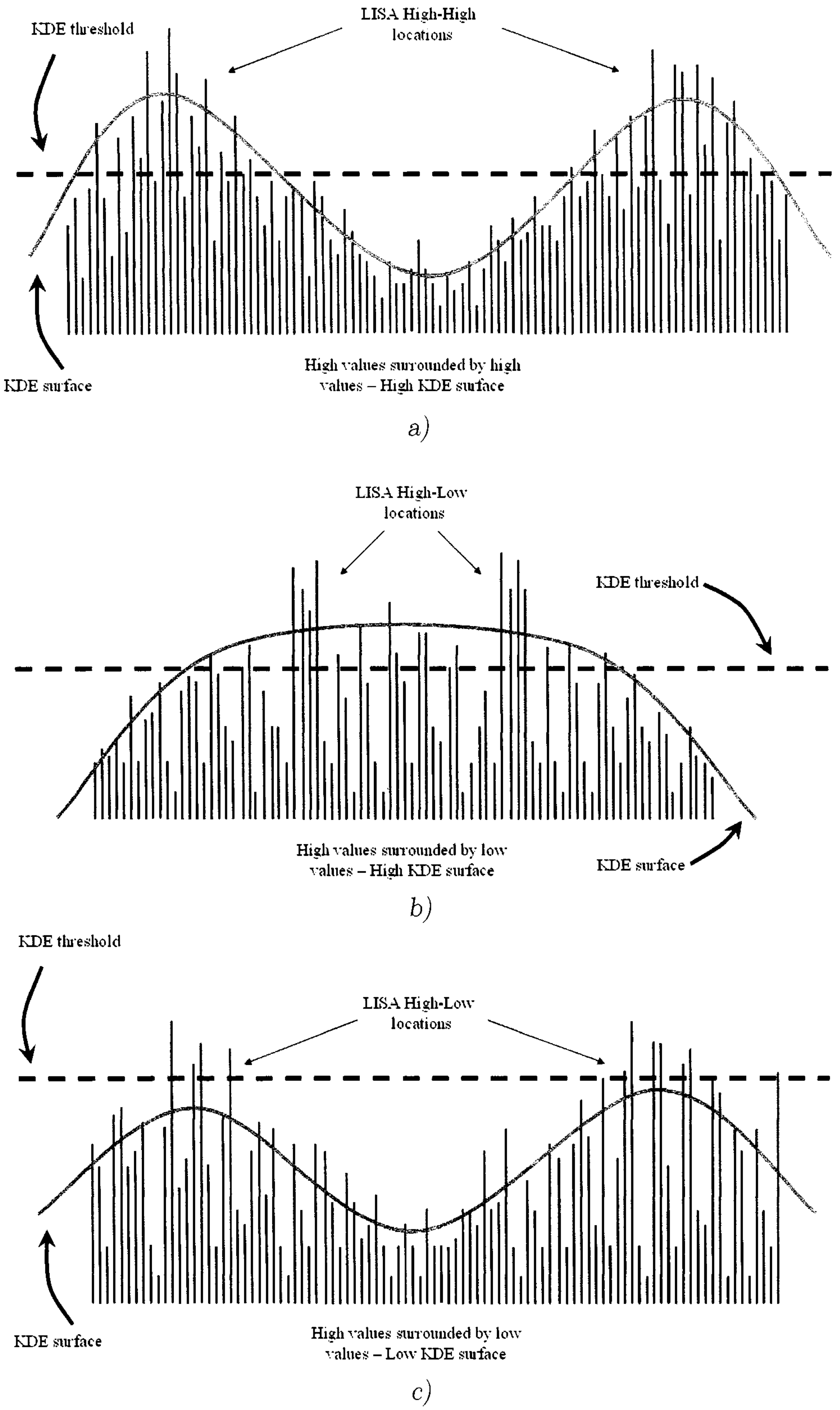

KDE threshold
LISA High-High locations
KDE surface
High values surrounded by high values – High KDE surface
a)
LISA High-Low locations
KDE threshold
High values surrounded by low values – High KDE surface
KDE surface
b)
KDE threshold
LISA High-Low locations
KDE surface
High values surrounded by low values – Low KDE surface
c)

Figure 4.14: Illustration of (a), LISA high-low and KDE hot spots (b) and LISA high-low and KDE infested areas (c). Red bars represent locations with high levels of infestation; blue bars represent locations with low levels of infestation.

Year	HH	HL	HH + HL	Points	HH %	HL %	HH + HL %
1996	687	92	779	6,076	11.31	1.51	12.82
1997	268	3	271	8,461	3.17	0.04	3.20
1998	228	41	269	2,418	9.43	1.70	11.12
1999*nof*	266	22	288	4,657	5.71	0.47	6.18
1999*f*	248	20	268	4,657	5.33	0.43	5.75
2000*nof*	88	49	137	5,310	1.66	0.92	2.58
2000*f*	40	42	82	5,310	0.75	0.79	1.54
2001*nof*	207	5	212	5,226	3.96	0.10	4.06
2001*f*	446	169	615	5,303	8.41	3.19	11.60
2002*nof*	499	50	549	8,308	6.01	0.60	6.61
2002*f*	646	119	765	8,401	7.69	1.42	9.11

Table 4.4: Percentage of points that have been classified as hot spots by means of LISA for each year. *HH* is the number of high-high hot spots, *HL* the number of high-low hot spots, *HH + HL* the total number of hot spots and *Points* the total number of points in a given data set. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

Comparing this table with Figure 4.7 quickly informs of the correspondence between those years having higher percentages of host spots detected and those that have higher values of global Moran's I, namely: 1996, 1998, 2001*f* and 2002*f*. It would be possible to think this result was predictable as it would be fair to think that since LISAs are proportional to the global statistic, those years showing a lower global Moran's I value should have a lower percentage of hot spots detected via Local Moran's I. It is important to emphasize that this is not necessaril8y the case (but it turned out to be this way in this case) since the proportionality constant can be smaller than unity thus making this assumption no longer valid.

Most of the years have a total hot spot percentage (*HH + HL*) below 10%, with 1996, 1998 and 2001*f* being the only exceptions. This is consistent with what was expected to occur since the use of an absolute threshold should have a lower outcome than the one used with the relative top 10%. However, it is important to remember that the analysis is being carried out only at the significance level of $\alpha = 0.05$ and one would expect to have an even lower percentage of hot spots, since under a 'pure chance' scenario 5% of the data would be expected to show this behavior. This is not reflected in the results as there is a large variability in the percentages for the detected hot spots $(HH + HL)$ ranging from 1.54% to 12.82%.

For those years for which field data were collected, there seems to be a very good correspondence between hot spots detected for the 'Adjusted Aerial' and the 'Adjusted Aerial-Field' data sets (see Figures 4.10 - 4.12), except for one small region in the northeastern portion in 2002 (Figure 4.13). A more detailed illustration of this situation is shown in Figures 4.15 and 4.16, where a comparison between the aforementioned region of both data sets is made.

It can be seen that for the 'Adjusted Aerial' data set there appears to be a concentration of high-high hot spots, suggesting the existence of high values of infested trees surrounded by other high values. However, for the 'Adjusted Aerial-Field' data set there are no hot spots in the area. In order to investigate this discrepancy, points that had purely simulated data and points that had field data available were identified. Figure 4.16 shows this information in the following way:

- Locations marked with a green tree symbol have field data collected, the values of which vary from 1 to 300 trees and are kept constant throughout the 100 simulations;

- Locations marked with a red cross have values that are always simulated and are assigned a value by randomly drawing its values from a gamma distribution;

- Locations marked with a black dot represent locations that have field data collected and have been corrected to have a value of 0. These values are kept the same throughout the 100 simulations.

As it can be interpreted from the above classifications and from the data shown in Figures 4.15 and 4.16, there appears to be an artificial way of obtaining hot spots in

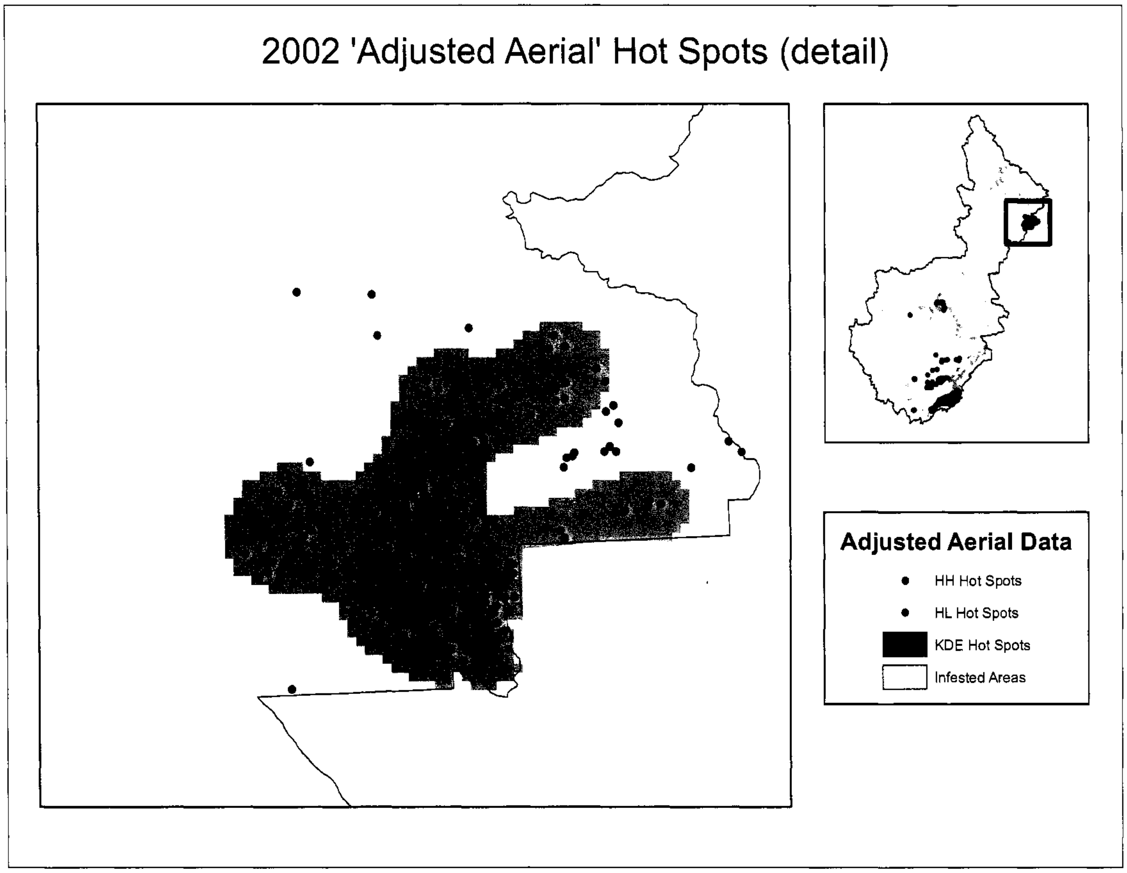


Figure 4.15: Northeastern region of the study area for the 'Adjusted Aerial' data set.

this area as field data does not indicate the existence of very high levels of infestation. Purely simulated data, however, shows that there is, indeed, a cluster of high values in the area, but this cannot be confirmed using field data. Furthermore, there appears to be a very low number of fixed values corresponding to field data (green tree symbols) in the region under scrutiny. There is, however, a large number of purely simulated data (red cross symbols) that could be affecting the outcome and are potentially involved in creating the observed difference. It is not clear what mechanism may be producing such differences, but it is likely that the existence of several locations with field data collected (both green symbols and black dots) may have the effect of suppressing high values from occurring in the area. It is also possible that the

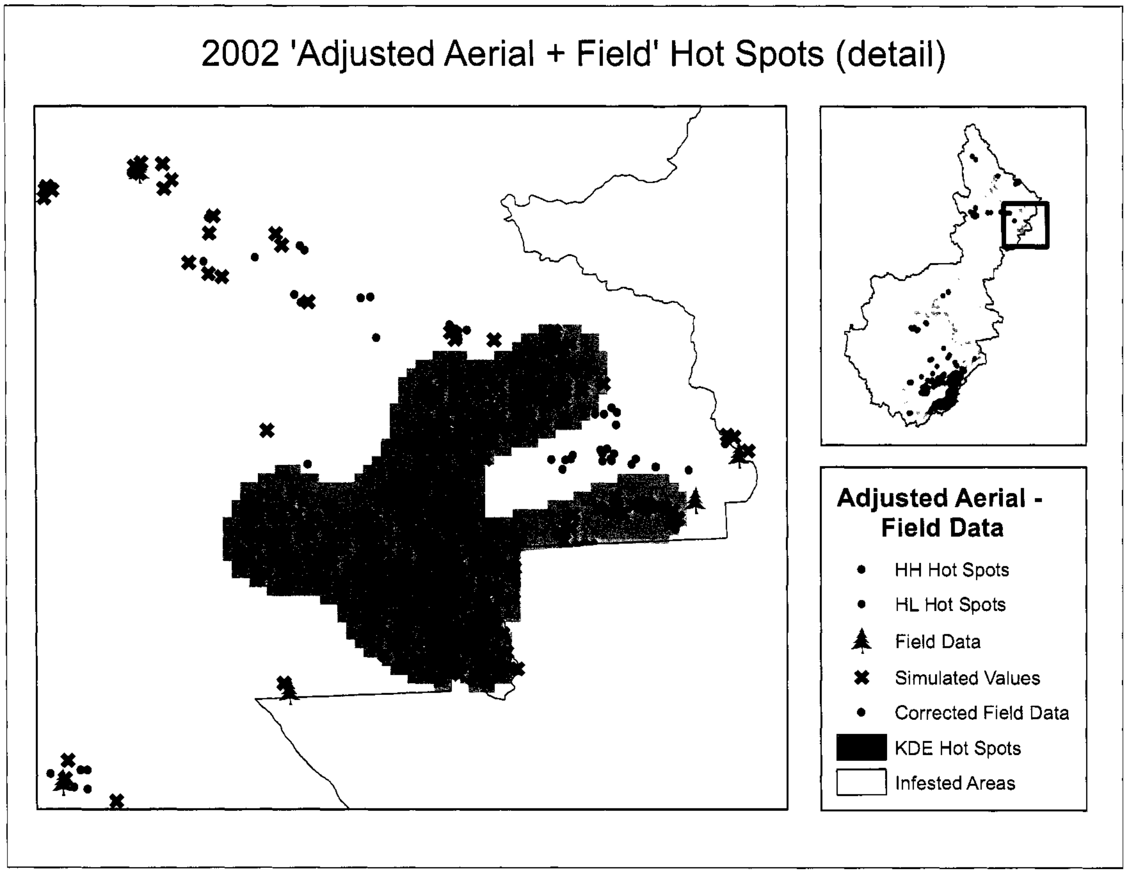


Figure 4.16: Northeastern region of the study area for the 'Adjusted Aerial-Field' data set.

aerial survey has other sources of error that have not been taken into account, such as human error introduced by a particular surveyor, difficulties to survey the area in that particular year, the identification of several clusters of infested trees that are not associated to mountain pine beetle or that are not pines. It is, however, interesting to note that for the previous year, 2001, there appears to be good agreement on detected hot spots in the same region, for both data sets.

As was mentioned before, it is not possible to work out all the analyses required and present results for all significance levels and all years. Nevertheless, it is useful to present some results that would give some indication of what can be expected should these analyses be carried out. Table 4.5 shows how the percentage of overall hot spots ($HH + HL$) changes when different levels of significance (α) are used. It can be seen that with more strict significance levels, this percentage decreases.

Year	Points	HS ($\alpha = 0.05$)		HS ($\alpha = 0.01$)		HS ($\alpha = 0.001$)		HS ($\alpha = 0.0001$)	
		Points	%	Points	%	Points	%	Points	%
1996	6,076	779	12.82	688	11.32	580	9.55	457	7.52
1997	8,461	271	3.20	198	2.34	159	1.88	75	0.89
1998	2,418	269	11.12	234	9.68	215	8.89	202	8.35
1999*nof*	4,657	288	6.18	182	3.91	44	0.94	7	0.15
1999*f*	4,657	268	5.75	160	3.44	76	1.63	35	0.75
2000*nof*	5,310	137	2.58	41	0.77	12	0.23	0	0.00
2000*f*	5,310	82	1.54	26	0.49	10	0.19	0	0.00
2001*nof*	5,226	212	4.06	176	3.37	90	1.72	24	0.46
2001*f*	5,303	615	11.60	387	7.30	220	4.15	147	2.77
2002*nof*	8,308	549	6.61	444	5.35	347	4.18	196	2.36
2002*f*	8,401	765	9.11	617	7.34	475	5.65	382	4.55

Table 4.5: Percentage of points that have been classified as hot spots by means of LISA for each year, for different significance levels, α. *HS* is the total number of hot spots and *Points* the total number of points in a given data set. Years without a suffix (*nof* or *f*) only contain 'Adjusted Aerial' data.

Furthermore, aiming to know a little bit more of the expected behavior of the rest of the available data and significance levels, hot spots were studied and intersected with KDE surfaces to produce the equivalent of Table 4.3 for both of the 2002 data sets and these results are shown in Table 4.6. It can be seen from Tables 4.5 and 4.6 that as one becomes more strict with the significance level, the overall number of hot spots decreases for each year. Also, there are more high-high than high-low spots and while the percentage of coincidence for high-high and KDE hot spots increases, it decreases for high-low and KDE hot spots.

The differences presented previously show the potential of using a variety of methods to carry out hot spots detection and compare the outcomes, as well as the usefulness of field data collection. An ongoing study that will focus on hot spot

	Year	LISA HH	LISA HH - KDE	LISA HH - KDE (%)	LISA HL	LISA HL - KDE	LISA HL - KDE (%)
$\alpha = 0.05$	2002*nof*	499	465	93.19	50	25	50.00
	2002*f*	646	545	84.37	119	44	36.97
$\alpha = 0.01$	2002*nof*	428	404	94.39	16	5	31.25
	2002*f*	552	478	86.59	65	22	33.85
$\alpha = 0.001$	2002*nof*	341	324	95.01	6	1	16.67
	2002*f*	451	420	93.13	24	2	8.33
$\alpha = 0.0001$	2002*nof*	196	192	97.96	0	0	0.00
	2002*f*	371	359	96.77	11	0	0.00

Table 4.6: Number of LISA high-high and high-low hot spots that match KDE hot spots and infestation areas for different significance levels for 2002. Results show the total number of *LISA HH* and *HL* points and the number of significant *LISA* hot spots that were classified as *KDE HS* and the associated percentage.

detection using a different local statistic, with the same data sets, will help shed some more light into the validity of the findings reported here. Until these results become available and are compared to the ones presented here and the ones available in Nelson (2005), it is only possible to say that the proposed combination of data model and detection technique used in this work gives results that are comparable to those from Nelson (2005). It is interesting to note, however, that two very different data models and detection approaches turn out to give very similar results.

At this stage it is not quite possible to say much more about how sensitive mountain pine beetle hot spot identification is to the data model used or the detection technique. In order to find out more about this dependence, it would be necessary to use either: *a)* the same detection technique with two different data models or *b)* two different detection techniques to the same data model. The limitation in this case is that LISA cannot be applied to a KDE surface but, as mentioned above, the use of another local statistic for hot spot detection will provide further insights about how sensitive hot spot detection is to the data model and detection technique.

§ Chapter 5

Conclusions

Results presented in the previous chapter indicate that global spatial autocorrelation is indeed present in the data sets as can be inferred from the low, yet significant Moran's I values obtained for each year. In terms of the spatial pattern of infested trees, this means there is a correlation at a global scale of the locations beetles prefer to attack. As it was mentioned before, large and mature trees are preferred for they provide the best conditions for mountain pine beetle survival. This is also an indication of the way beetles seem to spread to colonize and kill neighboring tree stands, since suitable trees in the vicinity of an infested area are likely to be attacked.

The use of Local Indicators of Spatial Autocorrelation (LISA) has indeed proved useful to determine the locations of those regions that show unusually high levels of infested trees. There is a strong similarity between high-high hot spots detected with LISA and KDE, although there is a significant difference in the way they are defined. In the case of high-low hot spots that are detected with LISA, there is more variability in its correspondence with KDE hot spots. It is likely that local variations in the spatial pattern were not picked up by KDE but are revealed with the use of the LISA approach. Still, for most of the years there is a very good correspondence between LISA and KDE hot spots.

The use of the more 'liberal' significance level of $\alpha = 0.05$ gives very interesting results relating to the locations of hot spots. For the purpose of comparing how sensitive detection is, this significance level seems to indicate little sensitivity. However, as was stated in the previous chapter, it would be important to fully compare the obtained results with those distilled from the use of more strict significance levels of $\alpha = 0.01, 0.001$ and 0.0001. It has been shown that becoming less liberal will have an effect on the number of detected hot spots by decreasing its number. In this sense, this approach can be very useful for detecting the most pervasive hot spots, those regions of space that are very strongly infested. The use of more strict significance levels will likely result in a higher percentage of correspondence between detected LISA and KDE high-high hot spots, since by keeping only those locations that are significant at higher levels they would be more likely to be clearly placed inside the KDE HS patches. It is also expected that the percentage of overall hot spots $(HH + HL)$ to decrease when compared to the corresponding column in Table 4.4. This due to the fact that the number of points in each data set remains the same, but it is expected that the more strict the significance level is, the less number of hot spots are detected.

The availability of data and the ability to consider different levels of significance is very useful when trying to compare and assess the obtained results. It appears that the way of defining hot spots by using LISA is more flexible than KDE definition: it permits the existence of different significance levels that are helpful in assessing the validity of results and it is capable of not indicating any hot spots if the data does not show enough significant values to be considered. It also allows to identify local variations that would otherwise be ignored using KDE that could potentially help into gaining more information about other features of mountain pine beetle processes, such as dispersal and host selection. Specifically, it might be possible to study the evolution of high-low locations throughout the years with more detail in order to see if these local variations are transformed into high-high regions. It is worth mentioning that the use of more or less restrictive threshold levels (*e.g.*, 10%, 5%, etc.) in the KDE scenario could provide something similar to what can be obtained with the use of different significant levels with the LISA approach.

It would certainly be useful to use the rest of the data that were produced for this research – that is, carry out the analyses of the remaining significance levels – to undertake a comparison between the corresponding results and those presented here, to see if they are in fact a subset of what has been obtained. It would also be very interesting to use a different local statistic instead of Moran's I. Currently work is being carried out in this direction, on the same data sets, by a group at the University of Victoria, British Columbia, using the Getis statistic.

It is important to note that the inclusion of field data does in fact make a difference in the results. The most dramatic example of this situation is 2002, in which according to simulated data a small region in the northeast appears to be heavily infested, but the use of field data suggests that this behavior may be due to an artificial mechanism of allotting high simulated values along this region. This finding certainly is valuable towards understanding more about the nature of spatial error in the data sets. It is worth remembering that error has only been incorporated into the data sets based on its frequency distribution. However, it is very likely that it will have a spatial structure associated with it, one that so far has been ignored. In this sense, if this spatial pattern is indeed present, it could help explain why a particular area showed significant differences for the 2002 data sets. Furthermore, it would greatly improve the present conception of spatial uncertainty associated to mountain pine beetle aerial and field surveys. It would also be helpful in adjusting the way simulations are obtained, to properly reflect the nature of spatial error by somehow allowing for more variability in those regions that show to be more heavily impacted by spatial uncertainty.

An interesting approach would be to carry out the exploration of environmental and climatic characteristics in terms of the differences in the hot spots detected by LISA and KDE surfaces. This would provide insights into the 'preferred' selection of hosts by the mountain pine beetle as well as information on various types of dispersal. It would be important to try to assess the climatic and environmental conditions where there is both agreement and disagreement between these two methods. It would be critical to include geographical and physical information from the landscape such as tree age, tree height, slope, aspect and any other characteristics that could give insights into the process of mountain pine beetle dispersal and host selection.

As with any research, this one has taken on a specific approach but there are other possibilities that yet remain to be explored. For example, in order to investigate the dependence of hot spots on scale, it would be useful to repeat the analysis by using increments of the minimum distance that was used to define adjacency.
Also, instead of using a minimum distance to define neighborhoods, a 2 $km.$ distance could be fixed for every point and this would allow a better way of comparing LISA results with KDE, since in Nelson (2005) a smoothing radius of 2 $km.$ was used for the Kernel Density Estimated surfaces.

References and Bibliography

Amman, G. (1973). Population changes of the mountain pine beetle in relation to elevation. *Environmental Entomology*. 2: 541-547.

Amman, G., McGregor, M., Schmitz, R. and Oakes, R. (1988). Susceptibility of lodgepole pine to infestation by mountain pine beetles following partial cutting of stands. *Canadian Journal of Forest Research*. 18: 688-695.

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

Anselin, L. (2004). *GeoDa*.
`https://geoda.uiuc.edu/default.php`

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive spatial data analysis*. Harlow, Essex, England, Prentice Hall.

Bentz, B, Amman, G and Logan, L. (1993). A critical assessment of risk classification systems for the mountain pine beetle. *Forest Ecology and Management*. 61: 349-366.

British Columbia Ministry of Forests (2003). Timber supply and the mountain pine beetle infestation in British Columbia. Victoria, Forest Analysis Branch.

Bland J. M. and Altman, D. G (1995). Multiple significance tests: the Bonferroni method. *BMJ* 310:170.

Caldas de Castro, M. and Singer, B. H. (2006). Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis* 38(2): 180-208.

Cleveland, W. S. (1993). *Visualizing Data.* New Jersey, Hobart Press.

Diggle, P. (1985). A kernel method for smoothing point process data. *Applied Statistics.* 34: 138-147.

Doane, D. P. (1976). Aesthetic Frequency Classifications. *The American Statistician.* 30: 181-183.

Fotheringham, A. S., C. Brunsdon, and Charlton, M. (2000). *Quantitative geography: perspectives on spatial data analysis.* London; Thousand Oaks, Calif., Sage Publications.

Gatrell, A. C. (1994). Density estimation and the visualization of point patterns. *Visualization in Geographic Information Systems.* H. Hernshaw and D. Unwin. Chichester, John Wiley & Sons.

Geiszler, D., Gallucci, V. and Gara, R. (1980), Modelling the dynamics of mountain pine beetle aggregation in a lodgepole pine stand. *Oecologia.* 46: 244-253.

Getis, A., and Boots, B. (1978). *Models of Spatial Processes.* Cambridge, Cambridge University Press.

Haining, R. (1993). *Spatial data analysis in the social and environmental sciences.* New York, Cambridge University Press.

Jacoby, W. G. (1997). *Statistical graphics for univariate and bivariate data.* Series: Quantitative Applications in the Social Sciences, No. 117. London; Thousand Oaks, Calif., Sage Publications.

Leva, J., Uijt de Haag, M. and Dyke, K. (1996). Performance of standalone GPS. *Understanding GPS: Principles and Applications.* E.D. Kapplan, Boston, Artech House Publishers.

Logan, J., and Bentz, B. (1999). Model Analysis of Mountain Pine Beetle (Coleoptera: Scolytidae) Seasonality. *Environmental Entomology.* 28: 924-934.

Logan, J., White, P., Bentz, B., and Powell, J. (1998). Model analysis of spatial patterns in mountain pine beetle outbreaks. *Theoretical Population Biology.* 53: 236-255.

Manjunath, G., Simunic, T., Krishnan, V., Tourrilhes, J., Das, D., Srinivasmurthy, V. and A. McReynolds. (2004). Smart Edge Server Beyond a Wireless Access Point, *WMASH '04: Proceedings of the 2nd ACM international workshop on Wireless Mobile Applications and Services on WLAN Hotspots* pp. 41-50. ACM Press, New York, USA.

Mitchell, R. and Preisler, H. (1991). Analysis of spatial patterns of lodgepole pine attacked by outbreak populations of the mountain pine beetle. *Forest Science.* 37: 1390-1408.

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

Nelson, T. (2005). Spatial and Spatial-temporal anaysis of mountain pine beetle infestations at a landscape level. PhD Dissertation, Wilfrid Laurier University, Waterloo, Ontario.

Ord, J. K., and Getis, A. (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science.* 41: 411-432.

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

O'Sullivan, D. and Unwin, D. (2003). *Geographic information analysis.* Hoboken, N.J., J. Wiley.

Powell, J., and Rose, J. (1997). Local consequences of a global model for mountain pine beetle mass attack. *Dynamics and Stability of Systems.* 12: 3-24.

Powell, J., Kennedy, B., White, P, Bentz, B., Logan, J., and Roberts, D. (2000). Mathematical elements of attack risk analysis for mountain pine beetle. *Journal of Theoretical Biology.* 204: 601-620.

Preisler, H. and Mitchell, R. (1993). Colonization patterns of the mountain pine beetle in thinned and unthinned lodgepole pine stands. *Forest Science.* 39:528-545.

Province of British Columbia (1996). *Gridded DEM specifications.* Victoria, Ministry of Sustainable Resource Management.

Safranyik, L., Shrimpton, D. and Whitney, H. (1974). *Management of lodgepole pine to reduce losses from the mountain pine beetle.* Victoria, Environment Canada, Forestry Service.

Safranyik, L., Silversides, R., McMullen, L. and Linton, D. (1989). An empirical approach to modeling the local dispersal of the mountain pine beetle (*Dendroctonus ponderosae Hopk.*)(Col., Scolytidae) in relation to sources of attraction, wind, direction and speed. *Journal of Applied Entomology.* 108: 498-511.

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

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika.* 66: 605-610.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* New York, Chapman Hall.

Sokal, R., Oden, N. and Thomson, B. (1998). Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society.* 65: 41-62.

Sturges, H.A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, p. 65.

Terrell, G. R. and Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association.* 80: 209-214.

Tiefelsdorf, M. (2000). *Modelling Spatial Processes.* Springer-Verlag, Berlin, Germany.

Tiefelsdorf, M. and Boots, B. (1995). The exact distribution of Moran's *I*. *Environment and Planning A.* 27: 985-999.

Tiefelsdorf, M. (2004). A Local and Global Test for Spatial Pattern Coherence among Sets of Regression Residuals. Under review by *Environment and Planning A.*

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234-240.

Turchin, P. and Thoney, W. (1993). Quantifying dispersal of southern pine beetles with mark-recapture experiments and a difussion model. *Ecological Applications.* 3: 187-198.

Turner, M.G., Gardner, R.H. and O'Neil, R.V. (2001) *Landscape Ecology in Theory and Practice: Pattern and Process*, New York, Springer-Verlag.

Voss, S. and George, S. (1995). Multiple significance tests. *BMJ*, 310:1073.

Wartenberg, D. and Greenberg, M. (1992). Methodological problems in investigating disease clusters. *The science of the total environment.* 127: 173-185.

Wood, C. S., and Unger, L. (1996). “Mountain pine beetle: a history of outbreaks in pine forests in British Columbia, 1910 to 1995”, Natural Resources Canada Forest Service, Victoria, BC.