

Covariate construction of nonconvex windows for spatial point
pattern data

by

Kabelo Mahloromela

14194237

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Department of Statistics

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

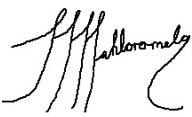
January 2020



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration

I, *Kabelo Mahloromela*, declare that this mini-dissertation (100 credits), which I hereby submit for the degree Magister Scientiae in Mathematical Statistics at the Univeristy of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: 

Date: January 2020

Summary

In the field of spatial statistics, window selection for point pattern data is a complex process. In some cases, the point pattern window is given a priori when a local phenomena is studied. In other cases, a researcher may choose this region using some objective means that reflects their view that the window may be representative of a larger region, or based on a probability sampling method. The common approaches used are the smallest rectangular bounding window and convex windows due to the obvious use of the Euclidean distance. The chosen window must however cover the true domain of the sampled point pattern data. Choosing a window too large results in estimation and inference in areas which are empty of observed data, but for which it has not been confirmed that observations could have occurred there. These holes in the domain could be regions where for some geographic (or other) reason the phenomena of interest does not occur.

In this mini-dissertation a review of methods for spatial convex and nonconvex window estimation is provided, and an algorithm is proposed for selecting the point pattern domain without the restriction of convexity, allowing for a better fit to the true domain, and based on spatial covariate information. The effect of the window choice on spatial intensity estimates is illustrated by giving particular attention to the technique of smoothed kernel intensity estimation. The proposed algorithm is applied in the setting of rural villages in Tanzania's Mara province. As a spatial covariate, remotely sensed data based on the elevation of a point pattern is used in the form of a Digital Elevation Model (DEM) GTOPO30, specific to village house locations in this setting. Mathematical morphological operators are also used to extract physiographic features from the DEM and are included here as a preprocessing step in the spatial window domain modelling. Analysis was conducted using the R software [76].

This research, entitled *Covariate construction of nonconvex windows for spatial point patterns*, was approved by the Faculty of Natural and Agricultural Science Research Ethics committee at the University of Pretoria under the reference NAS339/2019.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Inger Fabris-Rotelli, for providing her continued support and guidance. Her patience, positivity, motivation, inspiration, dynamism, innovative thinking (and the occasional provision of a delicious snack) have been invaluable in allowing me to complete this research and expand my repository of skills and experience.

I would like to thank and show my appreciation to my co-supervisor Christine Kraamwinkel for her valued inputs, contributions, encouragemnts, and the cups of coffee that were also invaluable to this work.

The completion of this research would also have not been accomplished without sponsorship and financial support from STATOMET and the DST/NRF SARChI Chair. This work is based upon research supported by the South Africa National Reasearch Foundation and South Africa Medical Research Council (South Africa DST-NRF-SAMRC SARChI in Biostatistics, Grant number 114613). Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

I am extremely grateful to my friends and colleagues in the Department of Statistics at the University of Pretoria. Some special mentions include Lindo and Cleo. Their empathy, great sense of humour and friendship have kept me balanced throughout this research.

Last but not least, I would like to thank my family, my parents Ntombifuthi Mahloromela and Masilo Mahloromela, my sister Keneilwe Mahloromela and other friends who offered encouragement and support, and who, despite not being versed in the subject matter and not knowing what I was talking about half the time, bore the brunt of my ramblings and patiently listened when I spoke about my research.

Contents

1	Introduction	13
1.1	Representative spatial window domains	16
1.2	Spatial measures and Euclidean distance	18
1.3	Outline	21
2	Window selection methods	23
2.1	Set theory	24
2.1.1	Convex sets	24
2.1.2	Convex hull	25
2.1.3	Open and closed sets	26
2.1.4	Compact set	26
2.1.5	σ -algebra	26
2.1.6	Lebesgue measure	27
2.2	Homogeneous Poisson Process	27
2.3	Convex set estimation	28
2.3.1	Dilation of the convex hull about its centroid	28
2.3.2	Convex domain estimation using inside and outside observations	30
2.4	Nonconvex Voronoi set estimate	32
2.5	Concluding remarks	32

<i>CONTENTS</i>	5
3 Covariate construction of nonconvex sets	34
3.1 Testing dependence of a point pattern from spatial covariates	36
3.1.1 Quadrat defined by a covariate	36
3.1.2 Berman's test	37
3.1.3 Cumulative Distribution Function test (CDF)	37
3.2 Window construction via covariates	38
3.3 Concluding remarks	42
4 Intensity estimation	43
4.1 Intensity function	43
4.2 Quadrat counts	44
4.3 Kernel smoothed intensity estimation	48
4.4 Intensity as a function of a covariate	54
4.5 Intensity estimation on nonconvex domains	56
4.5.1 Estimation in empty space	56
4.5.2 Euclidean measure of proximity	61
4.6 Concluding remarks	63
5 Application	65
5.1 Data description	65
5.2 Digital Elevation Model	68
5.3 Morphological segmentation of physiographic features from a DEM	70
5.3.1 Mathematical morphology	70
5.3.2 Mathematical morphological operators	70
5.3.3 Peak extraction in DEM using mathematical morphological operators	74
5.3.4 Gaussian blurring	77

<i>CONTENTS</i>	6
5.4 Nonconvex covariate constructed domain for locations of households	78
5.4.1 Landscape and terrain features	81
5.5 Discussion	82
6 Conclusion	89
Bibliography	91

List of Figures

1.1	Marked point pattern of the locations of forest fires in the Castilla-La Mancha region of Spain between 1998 and 2007 with the cause of fire as marks [4].	14
1.2	Distribution of buzzard nesting territories in two upland regions, Snowdonia and Migneint-Hiraethog, in North Wales, UK [20]. Open and closed dots indicate nesting sites outside and inside the study region marked by dashed boundaries respectively. Horizontal hatching is used to indicate terrain unsuitable for nesting.	17
1.3	Distribution of sparrowhawk nesting territories in two areas, Upper Speyside and Eksdale, Britain [69]. Closed dots indicate the nesting sites.	17
1.4	Kernel smoothed intensity estimates of a generic point pattern of the locations of fish in a lake (red dots), estimated on a rectangular window domain and fitted with a Gaussian kernel. The point pattern is created by simulating points on a polygon of Canada's Great Bear Lake in Northwest Territories marked by a solid boundary.	18
1.5	Google Earth plot of the geographical locations of households in Magatini village in Mara province, Tanzania. Green markers are used to denote household locations. The black line shows the Euclidean distance between two selected points and the yellow line shows the distance along the base on the mountain between the same points.	20
1.6	Minkowski distance measures	20
2.1	Spatial set estimates for a random sample of 40 points.	24
2.2	Convex and nonconvex planar sets	25
2.3	Convex hull of a set of points in \mathbb{R}^2	25
2.4	Realization of a homogeneous planar Poisson process on a square window with $\lambda = 3$. . .	28

2.5	Ripley-Rasson estimate of 200 simulated points on an elliptical domain.	30
2.6	The estimates of D when inside and outside observations are observed for two domains. . .	31
2.7	Voronoi estimate of nonconvex D . The nonconvex set D is outlined with a solid black line, and the estimate of the set \hat{D} is shaded in grey.	32
3.1	Seasonal spatial distribution of the wind vector data in 2003-2013 over the East China Sea. Each point of the wind vector data 10m over the sea surface are averaged according to the latitude and longitude in the same month from 2003 to 2013. The colours indicate the average wind speed and the arrows indicate wind direction.	35
3.2	Point pattern plot, in (a), denoting the locations of 3605 trees in a tropical rain forest in a 1000 by 500 meter study area in Barro Colorado Island, and equal area tessellation of the study area, in (b), to the quartiles of terrain elevation [4].	36
3.3	A toy example of a point pattern plot in (a) and spatial covariate surface in (b).	39
3.4	Illustration portraying M_i regions on toy example point pattern.	40
3.5	Illustration portraying the selected points \mathbf{u} (represented by closed dots), \mathbf{x}_i (represented by an open dot) and M_i enclosed in the square boundary.	41
3.6	Point pattern plot over covariate constructed nonconvex window.	42
4.1	Point pattern simulated from a homogeneous Poisson process with $\lambda = 2$ over a square window with a side length of 10.	45
4.2	Point pattern simulated from an inhomogeneous Poisson process with $\lambda(\mathbf{x})$ taking the form of Equation 4.2, over a square window with a side length of 10.	46
4.3	Quadrat counts and corresponding intensity plots with square quadrats of various sizes for simulated pattern depicted in Figure 4.1.	46
4.4	Quadrat counts and corresponding intensity plots with square quadrats of various sizes for simulated pattern depicted in Figure 4.2.	47
4.5	Quadrat counts and corresponding intensity plots with quadrats of various shapes with equal area for simulated pattern depicted in Figure 4.1.	47
4.6	Quadrat counts and corresponding intensity plots with quadrats of various shapes for simulated pattern depicted in Figure 4.2.	48

4.7	Examples of bivariate kernel functions	49
4.8	Kernel smoothed intensity estimates and corresponding 3D perspective plots are shown in panel (a) and (b) respectively with bandwidths of various sizes for the simulated homogeneous pattern depicted in Figure 4.1.	51
4.9	Kernel smoothed intensity estimates and corresponding 3D perspective plots are shown in panel (a) and (b) respectively with bandwidths of various sizes for the simulated inhomogeneous pattern depicted in Figure 4.2.	52
4.10	Contours of bivariate Gaussian kernel function for varying choices of the bandwidth matrix.	54
4.11	Point pattern (points in pattern shown with black dots) and bivariate Gaussian kernel smoothed intensity plot are shown in panel (a) and (b) respectively. Points in pattern are shown with red dots in panel (b). Contour lines are also indicated in the kernel smoothed intensity plot in panel (b).	54
4.12	Irregular nonconvex window domains with varying areas, where (a), (b), (c), and (d) have areas of 508.6584, 3.697251, 2.15776, and 56.80648 square units respectively.	58
4.13	Point pattern plots of simulated points in a rectangular window, overlaid with the true window domain.	59
4.14	Kernel smoothed intensity plots of simulated point patterns.	60
4.15	Relative intensity plots of simulated point patterns.	60
4.16	Point pattern simulated from a Poisson process with $\lambda = 25$ over an irregular, nonconvex window,	61
4.17	Illustration of the effect of using the Euclidean distance, over nonconvex domains, in estimating the kernel smoothed intensity function shown for a single point event from Figure 4.16. A Gaussian kernel is used.	62
4.18	Illustration of the effect of using the shortest path distance, over nonconvex domains, in estimating the kernel smoothed intensity function shown for a single point event from Figure 4.16. A Gaussian kernel is used.	63
5.1	Point pattern plots (left) and terrain elevation (right) for villages Iseresere and Nyamakobiti in Tanzania's Mara province on a rectangular window.	66

5.2	Point pattern plots (left) and terrain elevation (right) for villages Magatini, Majimoto and Hekwe in Tanzania's Mara province on a rectangular window.	67
5.3	DEM elevation data plots for Bokore village in Tanzania's Mara Province.	69
5.4	Illustration of an example of a 3×3 square structuring element around a target pixel and the neighbourhood defined by the structuring element	71
5.5	Illustration of an erosion of a greyscale image by a 3×3 square structuring element	72
5.6	Illustration of a dilation of a greyscale image by a 3×3 square structuring element	73
5.7	Illustration of a geodesic dilation of size 1 using a marker f and mask g , and a 3×3 square structuring element in the dilation step.	75
5.8	Illustration of ultimate erosion operation performed by the successive erosion (left) on the image until objects vanish and reconstructing each eroded image (middle) using an erosion of smaller size. The eroded sets (right) are formed by subtracting the reconstructed images with the corresponding eroded images.	76
5.9	Gaussian kernel with mean $(0, 0)$ and $\sigma = 1$	77
5.10	Illustration of Gaussian blurring; (a) shows the input image of two parrots and (b) shows the resultant image after Gaussian blurring is applied to the input image with $\sigma = 6$	78
5.11	Point pattern plots (left) and corresponding terrain slope (right) for villages Iseresere and Nyamakobiti in Tanzania's Mara province, on nonconvex window constructed using covariate data	79
5.12	Point pattern plot (left) and corresponding terrain slope (right) for villages Magatini, Majimoto and Hekwe in Tanzania's Mara province, on nonconvex window constructed using covariate data	80
5.13	Google Earth satellite image of Iseresere village in Tanzania's Mara province, with household locations indicated with red circles.	81
5.14	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province.	84
5.15	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).	85
5.16	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).	86

5.17 Relative intensity plots for villages Iseresere and Nyamakobiti estimated on a rectangular window and overlaid with the nonconvex constructed window. 87

5.18 Relative intensity plots for villages Magatini, Majimoto and Hekwe estimated on a rectangular window and overlaid with the nonconvex constructed window. 88

List of Tables

4.1	Examples of univariate kernel functions [48, 97] commonly utilized	50
-----	--	----

Chapter 1

Introduction

Spatial point pattern analysis reached prominence in geographical sciences during the 20th century [36]. It involves methods for explaining patterns that may be expressed with location. The aim of an analysis may be to expand basic understanding of spatial processes or to extrapolate results to regions where observations have not been made [5]. Data comprising of the location of features are termed geospatial or spatial data. Instances of spatial data occur frequently in a vast field of scientific disciplines, which include ecology [105, 111], seismology [71, 112], geography [31] and spatial epidemiology [87, 106], to name a few. In the sphere of seismology, a researcher may be interested in collecting data on the regional distribution of earthquakes to investigate whether there are predictable patterns over this space [5]. An example of the use of spatial analysis in spatial epidemiology was by an English physician named Dr. John Snow who studied London's cholera outbreak in 1854 [106].

A spatial point pattern is the mapped locations of objects or events over a region of interest [25, 45]. Point locations may correspond to all possible events (mapped) or a subset of these events (sampled point pattern) [4, 45]. Stated more formally, a spatial point pattern can be defined as a realization of a finite random subset of a given bounded region $S \subset \mathbb{R}^2$ for a spatial point process [65, 36]. One way of representing the spatial point process mathematically is by a set of random variables, $N(A)$ contained in the area $A \in \mathbb{R}^2$, where $N(A)$ denotes the number of points or events in sub-regions A contained in the entire study region [36]. Points in a spatial point pattern are observed over a region W , termed the window, which conventionally represents the study area. The window is strictly contained in the region S where the point process is defined [65]. The window W may assume any 2-dimensional geometric shape, such as a rectangular window or a complex polygonal region [36]. Additional information on the points may be recorded and are termed marks [4]. A mark is a variable corresponding to the point locations in a spatial point pattern that represents a qualitative or quantitative attribute of the point [4]. Figure 1.1 [4] depicts an example of a marked point pattern denoting the locations of forest fires in the Castilla-La

Mancha region of Spain between 1998 and 2007. The cause of fire denotes a qualitative mark and contains the factor levels lightning, accident (for accidents or negligence), intentional (for fires started intentionally) and other (for unknown and other causes).

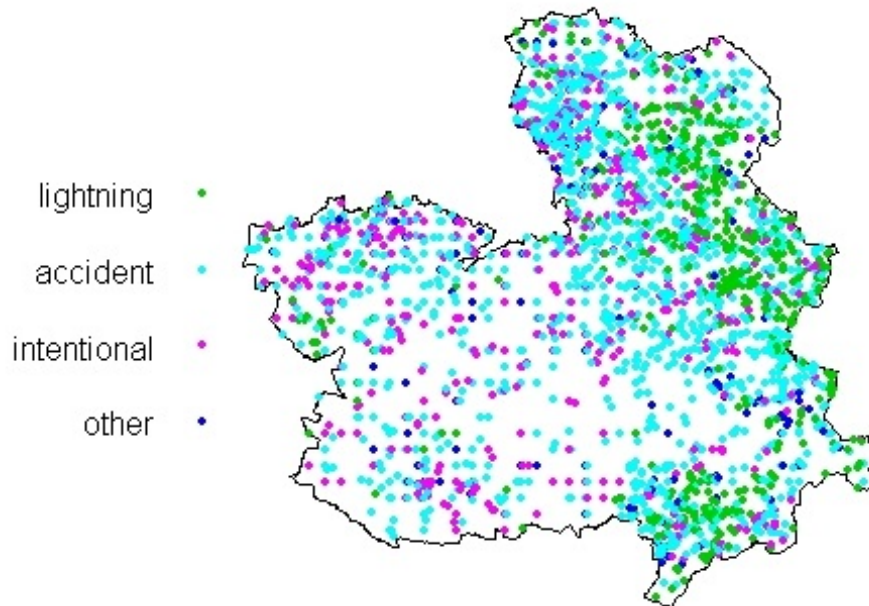


Figure 1.1: Marked point pattern of the locations of forest fires in the Castilla-La Mancha region of Spain between 1998 and 2007 with the cause of fire as marks [4].

Observations on other spatial variables, termed covariates, may be defined over the window domain and may aid in the improvement of modelling and interpretation. For example, a point pattern denoting the locations of trees in a forest could have basal area measurements corresponding to each point record as a mark and terrain elevation, defined on the entire study window, as a spatial covariate. When covariate information is available, an investigation into the dependence of a point pattern on covariate data should be conducted and this dependence quantified. Formal statistical tests have been developed to achieve this. The tests we considered in this mini-dissertation are outlined in Section 3.1 and include Berman's test for dependence of a point process on a spatial covariate and the cumulative distribution function test [4]. Testing for global and local dependence of point patterns on covariates in parametric models can also be done using the method proposed in [68]. Mapped point pattern data [45] may thus comprise of,

- an observation window W ,
- the point record of the points in the observation window are each in the form $\{\mathbf{x}_i, m_{i1}, m_{i2}, \dots, m_{ip}\}$, where \mathbf{x}_i denotes the coordinates of the point location and $m_{i1}, m_{i2}, \dots, m_{ip}$ are marks collected at \mathbf{x}_i , and

- covariate information as a spatial measurement $Z(u_j)$, $u_j \in W$, for $j = 1, \dots, l$, where the points u_j form a lattice extracted from continuous data over W and do not coincide with the points of the point pattern. It is customary for $Z(u_j)$ to describe a continuous regionalised variable, defined at every point in W , whereas the marks are only given and defined for the points in the point pattern.

Information regarding the window W and the point locations $\{\mathbf{x}_i\}$ in this window, are required for an appropriate analysis of the point pattern.

For a sampled point pattern, a consequence of constructing a boundary around the study region is that edge effects are occasionally erroneously defined, resulting in so-called "edge-effects". Edge effects arise when there is unknown interdependence between points outside the boundary and those inside the boundary [25]. Points in the study region that are close to the boundary may have neighbours situated outside the boundary. As a result, distance measures and summary statistics involving these points are not observed. If the influence of this interdependence is not considered a spatial point pattern analysis may produce erroneous results and bias statistical estimations. There exist several general methods for the correction of edge effects and for corrections specific to spatial point pattern statistical measures (i.e. the K -function). Among these are the guard method and the toroidal edge correction [54, 86, 99, 115], the border method [24] and the weighting method [54]. The guard method involves defining the window W within a specified distance from the boundary of a larger domain W^* to allow for a guard area. In the toroidal method for edge correction, the region W , is considered a torus so that points on opposite edges of the window are regarded as close to one another. This method is only applicable to rectangular study regions and is implemented by replicating the study region eight times and surrounding the original study region with the eight replicates. The border method proposed by Diggle [24] corrects for edge effects by removing all sample points that are closer to the boundary than to their nearest neighbor. Gatrell *et al* [36], have also considered edge correction methods for kernel estimates one of which involves dividing the kernel estimate with an edge correction term. The techniques proposed in this mini-dissertation give focus only to complete point pattern sets, thus edge correction methods are beyond the scope of what is considered. A more comprehensive discussion of these methods and others not discussed here can be reviewed in [24, 36, 37, 43, 54, 74, 99, 115]. Most edge correction methods are mainly considered for rectangular and circular study regions and become more complex and computationally involved for complex polygons.

Attributes of a spatial point process can be characterized by their first-order and second-order properties [36]. First order properties describe how the average of the process varies over different locations in space. Second-order properties describe how the process values are correlated in space [36]. A measure used to describe the first-order properties of a point process is the intensity function $\lambda(\mathbf{x})$ [36], defined as the average number of points per unit area and is denoted by,

$$\lambda(\mathbf{x}) = \lim_{dx \rightarrow 0} \left\{ \frac{E(N(d\mathbf{x}))}{dx} \right\},$$

where $d\mathbf{x}$ is a small region around the point $\mathbf{x} \in \mathbb{R}^2$, $E(\cdot)$ is the expected value operator, dx is the area for this region, and $N(d\mathbf{x})$ refers to the number of events in the region $d\mathbf{x}$. The intensity can be estimated using kernel density estimation with the general form.

$$\hat{\lambda}(\mathbf{x}) = \sum_{i=1}^n \frac{1}{h^2} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right),$$

where $K(\cdot)$ denotes a unimodal weighting function that intergrates to one over its entire domain, h represents the bandwidth, $\hat{\lambda}(\mathbf{x})$ is the estimate for the kernel density of the spatial point pattern at point \mathbf{x} , and \mathbf{x}_i is the observed location of the i -th event.

1.1 Representative spatial window domains

In some standard applications of spatial point pattern data analysis, the selection of the study window W is a nontrivial task. The study window is important as it gives information about where observations were made and not made as well as where observations should be predicted [4]. In some cases, there is knowledge about W when a local phenomenon is studied, such as pores of a metallic foam in a pipe, fungal spots on leaves, fish in a lake, and cell centers in a small biological organ [45]. In other cases, depending on the purpose of an analysis, a researcher may choose this region using some objective means that reflects their view that W may be representative of a larger region, or W is based on a probability sampling method [25]. Techniques discussed in published literature typically consider analysis on a rectangular or circular domain. In real world applications however, sampled points may be defined over an unknown irregular window that has a disconnected or a nonconvex domain. These more general windows result from factors that may make it impossible to observe points in certain regions. Consider the following examples.

Figure 1.2 depicts the distribution of buzzard nests in two upland regions, Snowdonia and Migneint-Hiraethog, in North Wales, UK [20]. Open and closed dots are indicative of nesting sites outside and inside the study region respectively. The two upland regions are separated by a dashed line and horizontal hatching is used to indicate terrain unsuitable for nesting. In areas unsuitable for nesting, points may not be observed. A similar example can be found in a paper by Newton *et al* [69], based on a study of spacing of territories of *Accipiter ninus*, the sparrowhawk, in twelve areas of Britain. Figure 1.3 illustrates the locations of the nesting territories for these regions. The locations of the nesting territories were found by systematically searching all of the woods in the study area over a period of 3-8 years. The spacing of the territories were examined by plotting them on a map. Woodlands suitable for nesting were often separated by large areas of unsuitable woodland, farmland or other open country. The position of the border of the study areas presented a problem as some areas in the woodland were mainly adjacent to rivers and streams, so estimates of density and habitat proportions were influenced by the amount of open

land included on either side. The boundaries of the study regions were marked arbitrarily by making a line around the outermost nesting territories at a distance from their centers equal to half the mean inter-territory distance in continuous nesting habitat.

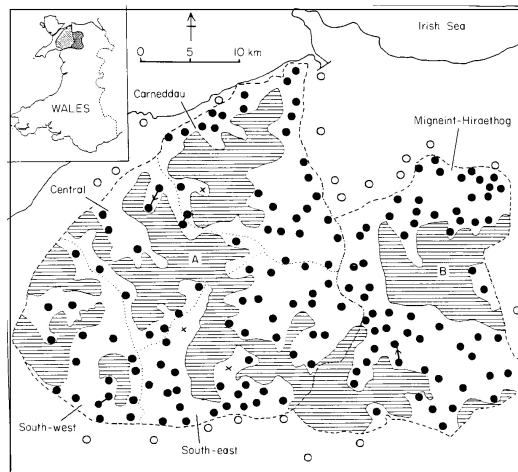


Figure 1.2: Distribution of buzzard nesting territories in two upland regions, Snowdonia and Migneint-Hiraethog, in North Wales, UK [20]. Open and closed dots indicate nesting sites outside and inside the study region marked by dashed boundaries respectively. Horizontal hatching is used to indicate terrain unsuitable for nesting.

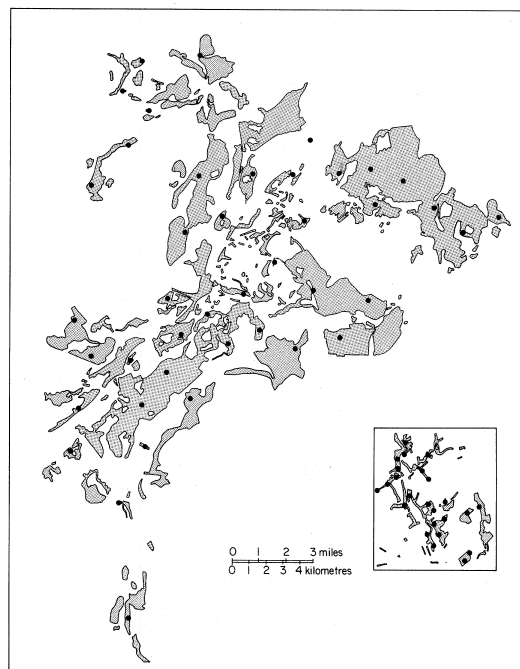


Figure 1.3: Distribution of sparrowhawk nesting territories in two areas, Upper Speyside and Eksdale, Britain [69]. Closed dots indicate the nesting sites.

Selection of the window W must be made carefully; choosing a window too large results in estimation in areas which are empty of observed data, but for which it has not been confirmed that observations could have occurred there. The empty areas could be regions where for some geographical (or other) reason the phenomenon does not occur. How W is chosen may also affect statistical measurements based on a spatial analysis, such as the intensity function [100]. Since the intensity function is a point measure of the locations in W , if W is misspecified we get intensity estimation in areas of the domain for which a point occurrence is non-observable. We illustrate this with the example in Figure 1.4. The figure depicts the kernel smoothed intensity estimate of a generic point pattern where point locations (red dots) denote fish in a lake. The point pattern was created by simulating points in a polygon of the Great Bear Lake in Northwest Territories, Canada¹. The simulation and analysis was done using the `spatstat` package in R [4, 76]. The bold irregular boundary represents the separation between lake water and dry land. The estimation is done over a rectangular window domain and fitted with a Gaussian kernel. As seen in the figure, for such a point pattern, estimation of intensity in this window domain will result in the incorrect allocation of positive intensity of fish over areas of dry land.

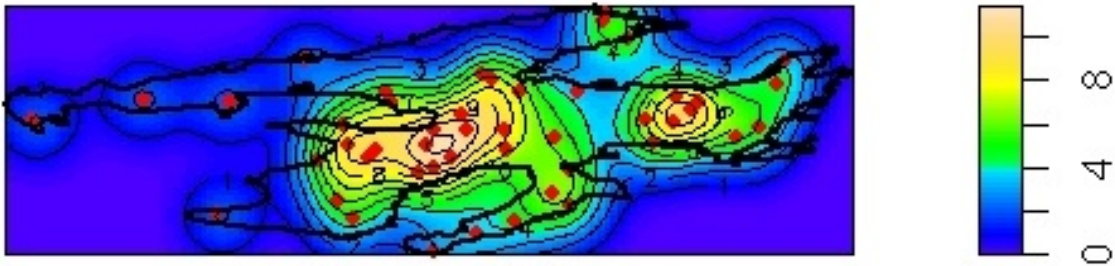


Figure 1.4: Kernel smoothed intensity estimates of a generic point pattern of the locations of fish in a lake (red dots), estimated on a rectangular window domain and fitted with a Gaussian kernel. The point pattern is created by simulating points on a polygon of Canada's Great Bear Lake in Northwest Territories marked by a solid boundary.

1.2 Spatial measures and Euclidean distance

In statistical analysis of spatial point patterns, Euclidean metrics are commonly used as a measure of distance between two points. The distance is calculated as the length of the line segment connecting two points. The expression for the Euclidean distance between two points, $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$, in \mathbb{R}^2 is given as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \quad (1.1)$$

¹<https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-lakes/>

This distance measure makes two implicit assumptions, (1) the distance is calculated along the shortest physical path formed by the line segment connecting two points; (2) the space has no variation in direction and is completely uniform [55]. These assumptions make it possible to use expressions that only require knowledge of the coordinates of the end point locations, thus the actual path between two points is avoided [55]. In instances as discussed above, this would mean the Euclidean does not represent the “true” distances between points on the physical surface represented by the window. Figure 1.5² shows Google Earth plots of the geographical locations, latitude and longitude, of households in the Magatini village in the Mara province of Tanzania³. In Figures 1.5(b) and (c), we see that mountainous regions make it non-viable for villagers to travel along the black coloured line segment, indicative of the Euclidean distance, and that a more representative path would follow along the yellow line given over a chosen nonconvex window. In addition to this, mountainous areas make it impracticable to build households, thus for the given point pattern, intensities should not be registered over these regions.

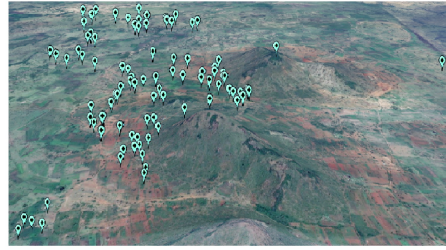
Different techniques have been proposed in order to determine distance measures between points [55]. One common method is the adjustment of the standard expression of the Euclidean distance formula in Equation (1.1) by replacing the powers 2 and $\frac{1}{2}$ with the values p and $\frac{1}{p}$. This distance is termed the Minkowski distance [55] and is defined as

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}. \quad (1.2)$$

For $p = 1$ and 2, the distance is the Manhattan and Euclidean distances respectively. If p tends to infinity, the distance is known as the Chebyshev distance given by the expression $d(\mathbf{x}, \mathbf{y}) = \max\{|x_i - y_i|\}$ for all i . Figure 1.6 is a visual depiction of the Minkowski distance measures, namely that of the Euclidean, Manhattan and Chebyshev distances. Adjusting the expression for the Euclidean distance may aid in distance measures between points that are more representative of the physical path between them on the surface represented by the window.

²Google earth V 7.3.2.5776. (July 21, 2017). Mara province, Tanzania. 1°37'52.38"S, 34°17'16.81"E. Eye alt 7.19km. Maxar Technologies 2020, CNES/Airbus 2020. <http://www.earth.google.com>[January 22, 2020], Google earth V 7.3.2.5776. (July 21, 2017). Mara province, Tanzania. 1°37'51.25"S, 34°17'17.33"E. Eye alt 7.19km. Maxar Technologies 2020, CNES/Airbus 2020. <http://www.earth.google.com>[January 22, 2020], Google earth V 7.3.2.5776. (July 21, 2017). Mara province, Tanzania. 1°37'16.36"S, 34°17'27.04"E. Eye alt 2.85km. CNES/Airbus 2020. <http://www.earth.google.com>[January 22, 2020]

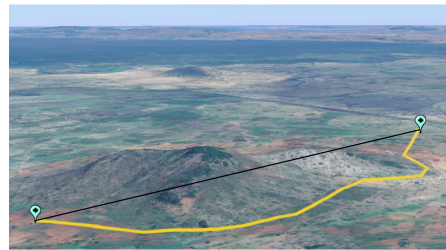
³<http://www.gla.ac.uk/researchinstitutes/bahcm/staff/katiehampson/>, <http://www.katiehampson.com/#intro>



(a)



(b)



(c)

Figure 1.5: Google Earth plot of the geographical locations of households in Magatini village in Mara province, Tanzania. Green markers are used to denote household locations. The black line shows the Euclidean distance between two selected points and the yellow line shows the distance along the base on the mountain between the same points.

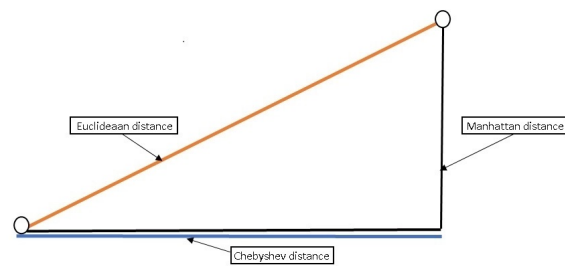


Figure 1.6: Minkowski distance measures

Literature detailing the selection of spatial window domains typically rely on an assumption of convexity of the domain. Ripley and Rassin [84] proposed a method for window selection via the reconstruction of

a compact convex set D from a realization of a homogeneous planar Poisson process observed over this unknown set D . Given a set of observations from a realization of a homogeneous planar Poisson point process within a compact convex set D , of unknown intensity, they present a technique to determine the compact convex set. Their proposed solution is the dilation of the convex hull about its centroid. This proposed method has possible applications in finding the edge of a natural forest given the locations of trees. Moore in [67] presents a method for solving a similar problem; that is, estimate D , a compact convex set in \mathbb{R}^p , given independent observations x_1, \dots, x_n sampled uniformly from unknown D . Rasson *et al* [78, 79, 82] consider how to extend these methods to estimate convex sets with observations that are inside and outside the convex set. There is other literature that address determining a convex hull or convex set from a random set of points; these include and are not limited to [29], [21], and [73]. The common approaches for window selection are (1) the smallest rectangular bounding box, illustrated in Figure 1.4, and (2) the convex hull. We propose here an approach to obtain a nonconvex window using additional geographical information.

1.3 Outline

The objectives of this mini-dissertaion are as follows:

- To model irregular nonconvex spatial window domains for spatial point patterns using spatial covariate information.
- Evaluate how the proposed noncovex window aids in improving intensity estimation and prediction.

To this end, we have organized this document as follows:

Chapter 2 is dedicated to a discussion of window selection techniques that involve the estimation of an unknown set when points realized from the unknown set (and points outside the set) are observed. The chapter begins with an overview of the definitions and concepts based on set theory, measure theory, and point processes, that are mentioned in the discussion of these methods.

Chapter 3 details statistical tests for investigating the dependence of the distribution of a point pattern on spatial covariates. We then propose an algorithm that uses the known effects of a covariate on the point occurrence of points in a spatial point pattern to construct a nonconvex spatial window domain.

Chapter 4 includes a discussion on nonparametric intensity estimation methods, with particular focus given to the method of kernel smoothed intensity estimation. We discuss how the kernel method for intensity estimation can be extended to allow for spatial covariate effects and discuss the effects of using the standard kernel smoothed intensity estimate as an estimate of intensity on irregular nonconvex window

domains and how these can be adapted.

In Chapter 5 the nonconvex window construction algorithm proposed in Chapter 3 is applied to households in a rural setting in Tanzania's Mara province, and use elevation data from a Digital Elevation Model as spatial covariates. A discussion on mathematical morphological operators used in the processing of the Digital Elevation Model is also detailed.

Chapter 6 is devoted to consolidating the results, concluding remarks, and discussing prospects for future research.

Chapter 2

Window selection methods

Selecting a spatial window S using points in a point pattern is analogous to set estimation. Set estimation tackles the statistical problem of estimating an unknown, often compact, set $S \subset \mathbb{R}^d$ from the observed values x_1, x_2, \dots, x_n of a randomly selected sample of points X_1, X_2, \dots, X_n drawn from this set [6]. The upper case letters denote a random variable, and the lower case letters the observed value of the random variable. Literature on the subject typically consider the case where S is assumed to be convex [29, 67, 82, 90]. The convex hull of the sample is the estimate of S under this assumption and is unique (see Section 2.1.2 for discussion). In the case where S is not assumed to be convex, there is no unique estimator [6]. An example proposed by Devroye and Wise [22] is

$$\hat{S}_n = \bigcup_{i=1}^n B(X_i, \varepsilon_n),$$

where $B(x, r)$ denotes the closed ball centered at x with radius $r > 0$, given by

$$B(x, r) = \{y : y \in \mathbb{R}^d \text{ and } \|x - y\| \leq r\}$$

and ε_n is a sequence of smoothing parameters which must slowly tend to zero in order to get a consistent estimation¹ as $n \rightarrow \infty$. Figure 2.1 [19] shows the convex hull and the Devroye-Wise estimator for the same random sample of 40 points. In the case of the Devroye-Wise estimator in panel (b), we observe that the estimated domain has holes and some areas of partial disconnection in contrast to the convex hull estimator in panel (a) which is connected.

In this chapter we review methods for spatial window set estimation. Sections 2.1 and 2.2 give an overview of some definitions and concepts based on set theory, measure theory and point processes that

¹A consistent estimate in statistics is an estimate which converges to the true value of the parameter being estimated as the sample size increases indefinitely [7].

are referenced in Section 2.3. Methods for convex and nonconvex set estimation are covered in Sections 2.3 and 2.4 respectively.

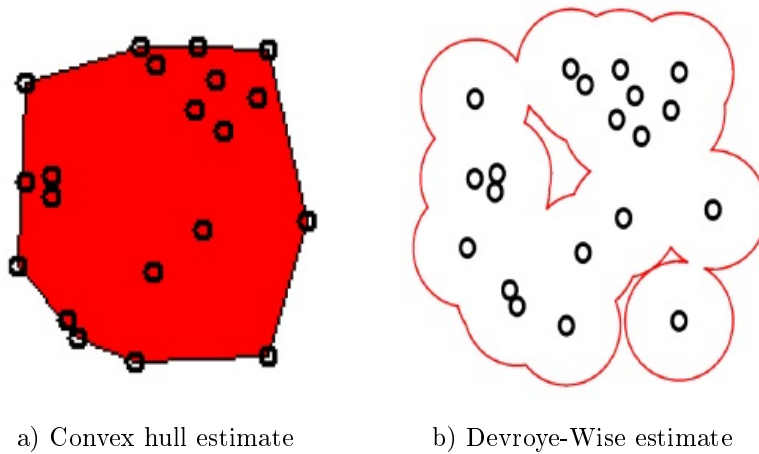


Figure 2.1: Spatial set estimates for a random sample of 40 points.

2.1 Set theory

This section introduces on terminology and concepts used for convex set estimation that will be considered in Section 2.2.

2.1.1 Convex sets

A set $C \subseteq \mathbb{R}^p$ is convex if for all points $x_1, x_2 \in C$, the line segment joining x_1 and x_2 is entirely in the set C . In other words, C is said to be convex if it contains the line segment,

$$\alpha x_1 + (1 - \alpha)x_2, \tag{2.1}$$

where $\alpha \in [0, 1]$ [21, 58, 88]. The linear sum in Equation (2.1) is termed a convex combination of x_1 and x_2 . When α is either 0 or 1, the points are x_2 and x_1 respectively. For intermediate values of α , the points lie on the connecting line segment. A convex set is thus a connected set [58], meaning mathematically that there exists a path contained in the set between any two points. Figure 2.2 shows examples of convex and nonconvex sets. The empty set is trivially convex. The intersection of a number of convex sets is also convex [58, 88].

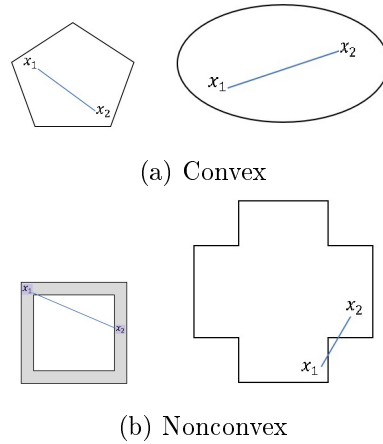


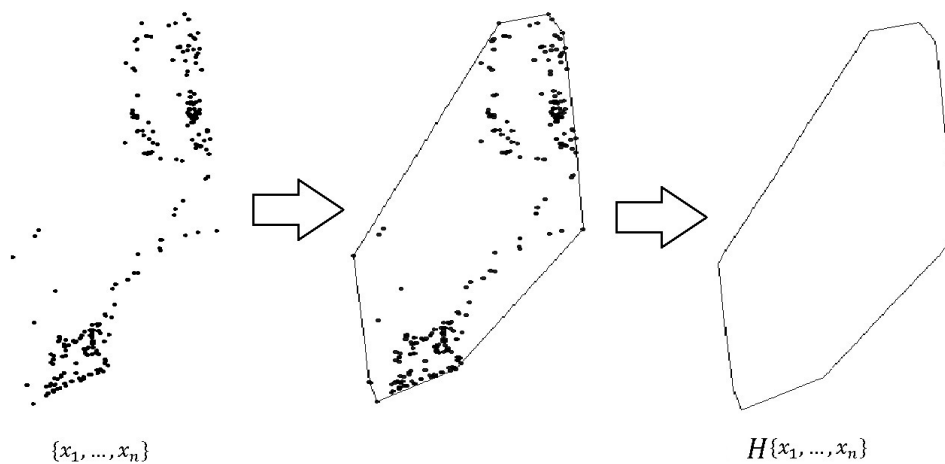
Figure 2.2: Convex and nonconvex planar sets

2.1.2 Convex hull

For a set $A \in \mathbb{R}^p$, whose elements comprise of points, the convex hull is defined as the intersection of all convex sets which contain the set A . That is, the convex hull of A is the smallest convex set that contains the set A , where smallest here means that there is no other convex set contained in the convex hull that also contains A [21, 34, 88]. This implies that the convex hull is also unique. For a point pattern, this would be the smallest convex window that envelops all the points. For n points x_1, \dots, x_n in \mathbb{R}^d the convex hull is given by

$$H\{x_1, x_2, \dots, x_n\} = \left\{ y \in \mathbb{R}^d : y = \sum_{i=1}^n \alpha_i x_i, \alpha_i \geq 0 \forall i, \sum_{i=1}^n \alpha_i = 1 \right\},$$

where the restrictions on the α_i 's, $\alpha_i \geq 0$ and $\sum_{i=1}^n \alpha_i = 1$, ensure that the set is the smallest. Figure 2.3 depicts the convex hull of a set of points in \mathbb{R}^2 .

Figure 2.3: Convex hull of a set of points in \mathbb{R}^2 .

2.1.3 Open and closed sets

An open set is a set that does not contain limiting points (i.e. boundary points). Mathematically, a set A is open if for every point, $a \in A$, there exists an $\varepsilon > 0$ such that the ball

$$B(a, \varepsilon) = \{y \in \mathbb{R}^d : \|a - y\| < \varepsilon\}$$

is contained in A [58, 104]. A closed set is defined as the complement of an open set [58]. An example of an open set is the unit circle in \mathbb{R}^2 centered at the origin without its boundary points,

$$A = \{y = (y_1, y_2) \in \mathbb{R}^2 : y_1^2 + y_2^2 < 1\}.$$

The union of set A and the set of its boundary points $B = \{y = (y_1, y_2) \in \mathbb{R}^2 : y_1^2 + y_2^2 = 1\}$ is a closed set.

2.1.4 Compact set

A set A is compact [58] if for any collection of open subsets, B , of A such that

$$A \subseteq \bigcup_{b \in B} b,$$

there is a finite subset B^* such that

$$A \subseteq \bigcup_{b \in B^*} b,$$

Compactness queries whether the set A can be formed with a finite number of open subsets of B . A compact set is closed and bounded.

2.1.5 σ -algebra

The σ -algebra of a set X [91], $\sigma(X)$, is nonempty collection of subsets of X such that

1. $\emptyset \in \sigma(X)$,
2. if $X \in \sigma(X)$, then the complement $X^c \in \sigma(X)$, and
3. if A_1, A_2, A_3, \dots is a sequence of elements in $\sigma(X)$, then $A_1 \cup A_2 \cup A_3 \cup \dots \in \sigma(X)$

The collection of Borel sets \mathcal{B} is the smallest σ -algebra that contains all open sets [91].

2.1.6 Lebesgue measure

The formulation of the Lebesgue measure [75] is as follows. An open box [75] $B \subseteq \mathbb{R}^d$ is a set of the form

$$B = \prod_{k=1}^d I_k,$$

where $I_k = (a_k, b_k)$, $k = 1, \dots, d$ are open intervals and the product symbol represents the Cartesian product. Its d -dimensional volume [75], denoted by $|B|$, is the product of the lengths of the open intervals I_k , expressed as

$$|B| = \prod_{k=1}^d (b_k - a_k).$$

The d -dimensional outer measure [75] of a set $A \subset \mathbb{R}^d$ is expressed as

$$m^*(A) = \inf \left\{ \sum_k |B_k| : \{B_k\} \text{ covers } A \right\}.$$

A set B^* is a cover for A if

$$A \subseteq \bigcup_{b \in B^*} b.$$

The set $\{B_k\}$, a collection of sets whose union contains A , forms a cover for A . A set A is Lebesgue measurable [75] if for each $S \subset \mathbb{R}^d$,

$$m^*(S) = m^*(S \cap A) + m^*(S \cap A^c).$$

If A is Lebesgue measurable its Lebesgue measure, denoted by $m(A)$, is $m^*(A)$.

The Lebesgue measure extends on the notion of length and area to more abstract sets [75, 91]. Intuitively, this measure can be interpreted as the size of the set. If X is a set in \mathbb{R}^d , for $d = 1, 2$ and 3 , the Lebesgue measure corresponds with common measures of length, area, and volume respectively.

2.2 Homogeneous Poisson Process

The set estimates described in Section 2.3 are derived under the assumption that observed points are realisations of a homogeneous Poisson process. As a precursor to the discussion in this section we give the definition of a homogeneous Poisson process.

A homogeneous Poisson process [51] is the simplest stochastic model used for planar point patterns. The point process X is a homogeneous Poisson process if it satisfies the following properties.

1. There exists a constant $\lambda > 0$ such that the number of points in set B for process X , $X(B)$, for each bounded Borel set B , is a Poisson distributed random variable with parameter $\Lambda = \lambda m(B)$, where

m denotes the Lebesgue measure in \mathbb{R}^2 . The parameter Λ denotes the expected number of points in the region B .

2. The random variables $X(B_1), \dots, X(B_n)$ are independent events for disjoint bounded Borel sets B_1, \dots, B_n .

The parameter λ is the intensity of X , the expected number of points per unit area. The process, X , is inhomogeneous when the intensity is a nonconstant positive function $\lambda(A)$ of the unit area A . Figure 2.4 depicts a realization from a homogeneous Poisson process with intensity 3. The simulation was done using the `spatstat` in R [4, 76].

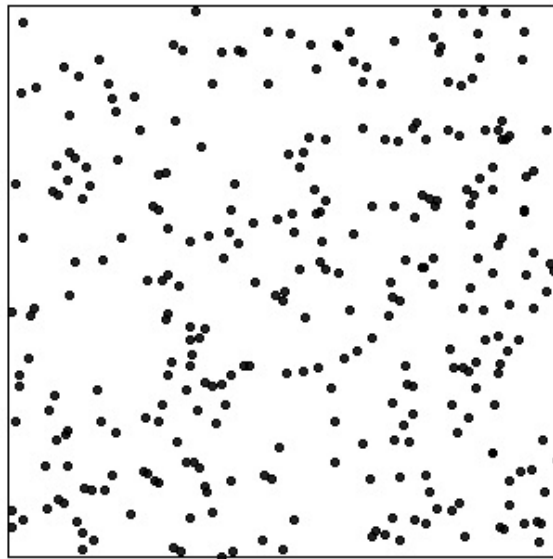


Figure 2.4: Realization of a homogeneous planar Poisson process on a square window with $\lambda = 3$.

In the next section we review methods for convex set estimation constructed from a set of observed points realized inside (and/or outside) this unknown set.

2.3 Convex set estimation

In this section we review methods for convex set estimation.

2.3.1 Dilation of the convex hull about its centroid

Ripley and Rasson in [84] construct a compact convex set D from a realization of a homogeneous planar Poisson process observed over this unknown set D . Given a set of observations from a realization of a

homogeneous planar Poisson point process within a compact convex set D , of unknown intensity λ , they present a technique to determine the compact convex set.

Let x_1, x_2, \dots, x_N denote a realization of points observed in $D \in \mathfrak{D}$, where \mathfrak{D} is the class of compact convex subsets of \mathbb{R}^d of positive measure. Given N and under the assumption of a homogeneous Poisson process, it follows that the x_i 's are independent uniformly distributed random vectors in D . That is, we observe

$$X_1, X_2, \dots, X_N \stackrel{i.i.d.}{\sim} U(D), N \sim \text{Poisson}(\lambda m(D))$$

where $m(\cdot)$ is the Lebesgue measure on the Borel subsets of \mathbb{R}^d and X_1, X_2, \dots, X_N, N are all independent. Consequently the joint density of X_1, X_2, \dots, X_N is given by

$$p_D(x_1, x_2, \dots, x_N) = \prod_{i=1}^N p_D(x_i) = \prod_{i=1}^N \frac{I_D(x_i)}{m(D)} = \frac{\prod_{i=1}^N I_D(x_i)}{m(D)^N} = \frac{I_D(x_1, x_2, \dots, x_N)}{m(D)^N}$$

for $D \in \mathfrak{D}$, where $I_D(\cdot)$ is an indicator function taking on the value of 1 if the set of points are in D and 0 otherwise. The point set $\{x_1, x_2, \dots, x_N\}$ is in D if and only if the convex hull of the points denoted by $H\{x_1, x_2, \dots, x_N\}$ is contained in D . It thus follows that

$$p_D(x_1, x_2, \dots, x_N) = \frac{I_D(H\{x_1, x_2, \dots, x_N\})}{m(D)^N}.$$

$H\{x_1, x_2, \dots, x_N\}$ is therefore a sufficient statistic and the maximum likelihood estimate of D .

The estimation of D involves first considering the set $s(H\{\mathbf{x}\}) = H\{\mathbf{x}\} - g(H\{\mathbf{x}\})$, where $g(H\{\mathbf{x}\})$ is the centroid of the convex hull $H\{\mathbf{x}\}$ where $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$. In the case when the Lebesgue measure $m(D)$ of D is known, the maximum likelihood estimator of D is expressed as a translation of

$$\left(\frac{m(D)}{m(H\{\mathbf{x}\})} \right)^{\frac{1}{2}} s(H\{\mathbf{x}\}).$$

For unknown $m(D)$, a constant c is found such that $E[m(cs(H\{\mathbf{x}\}))] = m(D)$,

$$c = ((n+1)/[n+1 - E[V_{n+1}]])^{1/2},$$

where V_{n+1} is the number of vertices in the convex hull of $\{x_1, x_2, \dots, x_{n+1}\}$, that is, the number of vertices if one more point were drawn from D . Difficulty in the computation of $E[V_{n+1}]$ warrants the use of an estimator such as the observed value v_n of V_n , that is to use $c = \left(\frac{n}{n-v_n} \right)^{\frac{1}{2}}$. The final estimator is a dilation² of the convex hull about its centroid

$$\hat{D} = g(H\{\mathbf{x}\}) + cs(H\{\mathbf{x}\}).$$

²A dilation (also termed a homothety) [44, 56, 61] is defined here as a transformation that enlarges or reduces the size of a geometric figure by a given scaling factor k around a center point O . It is a function that sends every point X on the plane to a point X' such that $\overrightarrow{OX'} = k\overrightarrow{OX}$. In this case, the center point is the centroid of the convex hull $g(H\{\mathbf{x}\})$, the enlargement/reduction scale factor is c and the geometric figure is $s(H\{\mathbf{x}\})$. This should not be confused with the mathematical morphological dilation operator given in Section 5.3.2.

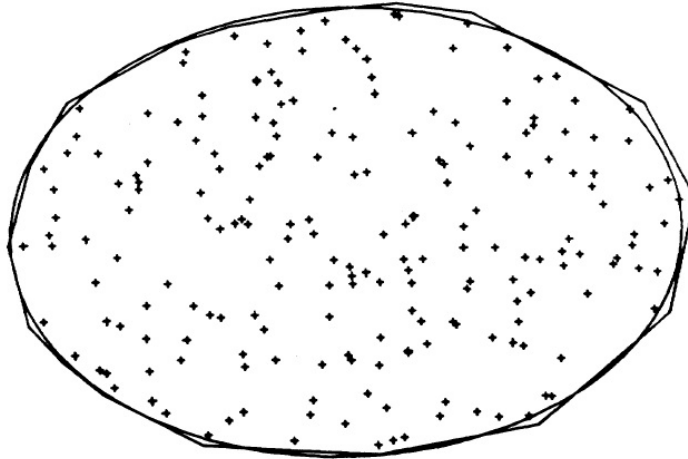


Figure 2.5: Ripley-Rasson estimate of 200 simulated points on an elliptical domain.

Figure 2.5 obtained from [84] depicts the result of this method for 200 points simulated on an elliptical domain. The result is derived assuming that the area of the true set is unknown. The smooth line depicts the boundary of the true elliptical domain D , and the kinked line the boundary of the Ripley-Rasson estimate \hat{D} of D .

2.3.2 Convex domain estimation using inside and outside observations

Remon [82, 81] proposes a method to estimate an unknown convex domain using observations inside and outside the domain. An application in which inside and outside observations are available is pattern recognition [41, 103]. In this context, the aim may be to delineate the boundary of convex bodies on an image that has pixels of differing texture [103]. In [82], Remon presents this method as a generalization of the Ripley-Rasson solution discussed in Section 2.3.1.

The problem being solved is formulated as follows. A realization of $n + m$ points from a Poisson process X is observed within a fixed window $F \in \mathbb{R}^d$. Contained in F is a compact convex domain D . From the $n + m$ observations, n are inside D and m are outside D . D^c denotes the complement of D . It is assumed that the Poisson process is homogeneous on F with density λ . The aim is then to estimate the unknown convex domain D . Suppose $(x_1, y_1, \dots, x_{n+m}, y_{n+m})$ is a realization of X , a homogeneous Poisson process, and Y , a labelling variable that takes on the value of 1 if $x_i \in D$ and 2 otherwise. Let $y_i = 1$ for $i = 1, \dots, n$ and $y_i = 2$ for $i = n + 1, \dots, n + m$ without loss of generality.

Given the observed values n and m , it follows that the likelihood function for (x, y) is

$$\begin{aligned} \mathcal{L}(x; y) &= \left\{ \frac{1}{[m(D)]^n [m(D^c)]^m} \prod_{i=1}^n \mathbf{I}_D(x_i) \prod_{i=n+1}^{n+m} \mathbf{I}_{D^c}(x_i) \right\} \\ &= \left\{ \frac{1}{[m(D)]^n [m(D^c)]^m} \mathbf{I}_D(H\{x_1, \dots, x_n\}) \mathbf{I}_{D^c}(J\{x_{n+1}, \dots, x_{n+m}|x_1, \dots, x_n\}) \right\} \end{aligned}$$

where $H\{A\}$ is defined again as the convex hull of A , and

$$J\{x_{n+1}, \dots, x_{n+m}|x_1, \dots, x_n\} = \bigcup_{i:y_i=2} \{x_i + \lambda(x_i - b) \in \mathbb{R}^d | \lambda \geq 0, b \in H(x_1, \dots, x_n)\}$$

$(H\{.\}, J\{.\})$ is a minimal sufficient statistic for the estimation of D . $J\{x_{n+1}, \dots, x_{n+m}|x_1, \dots, x_n\}$ is a consistent estimate of D^c . The boundary between D and D^c is the set of points,

$$\{x_0 : \text{such that likelihood of } x_0 \text{ belonging to either } D \text{ or } D^c \text{ is equal}\}.$$

That is,

$$S_1(x_0) = S_2(x_0)$$

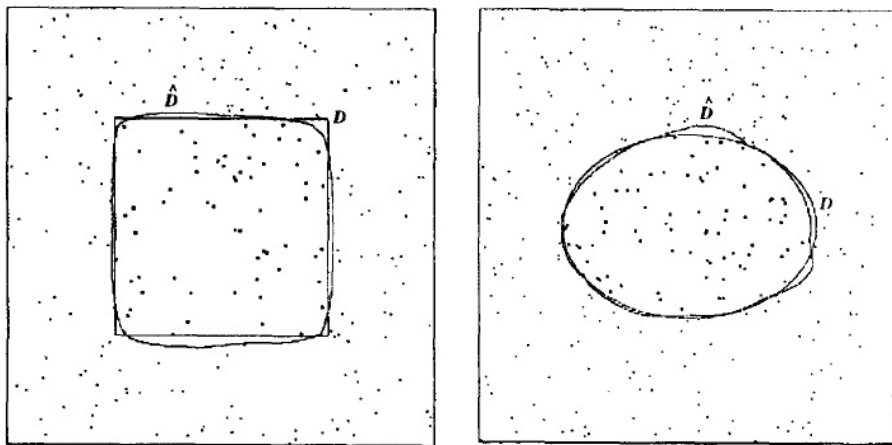
where

$$S_1(x_0) = m(H\{x_1, \dots, x_n, x_0\}) + m(J\{x_{n+1}, \dots, x_{n+m}|x_1, \dots, x_n\})$$

and

$$S_2(x_0) = m(H\{x_1, \dots, x_n\}) + m(J\{x_{n+1}, \dots, x_{n+m}, x_0|x_1, \dots, x_n\}).$$

Figure 2.6 obtained from [81] depicts the results for a square (a) and an ellipsoidal (b) domain. The observation sizes are 250 with $n = 59$ from D in the case of the quadratic domain, and 300 with $n = 68$ in the case of the ellipsoidal domain. The intensity in D and D^c is assumed constant and equal, thus the points inside and outside the domains are generated from a Poisson process with the same intensity. The true areas are given by $m(D) = 0.25$ and $m(D) = 0.2$ respectively. The respective areas for the symmetric difference $D \Delta \hat{D} = D \cup \hat{D} \setminus D \cap \hat{D}$ is noted as 0.018 and 0.011 and gives an indication of the size of the area in \hat{D} that is misclassified as belonging to D .



(a) Square domain

(b) Ellipsoidal domain

Figure 2.6: The estimates of D when inside and outside observations are observed for two domains.

2.4 Nonconvex Voronoi set estimate

The problem outlined in this section involves the estimation of a nonconvex set D using points observed inside and outside this unknown set. The formulation of the problem is given as in Section 2.3.2, the difference being that the unknown set D need not be convex. The Voronoi estimate [83] \hat{D} is the set of all points x_0 that satisfy the criterion

$$\min_{1 \leq i \leq n} d(x_0, x_i) \leq \min_{n+1 \leq i \leq n+m} d(x_0, x_i)$$

where d denotes the Euclidean distance in \mathbb{R}^2 . The Voronoi cell V_i related to the point x_i is given by

$$V_i = \{x_0 \in \mathbb{R}^2 : d(x_0, x_i) \leq \min_{j \neq i} d(x_0, x_j)\}$$

then

$$\hat{D} = \bigcup_{1 \leq i \leq n} V_i.$$

The solution is non-smooth and this irregularity cannot be reduced by increasing the size of observed points. Figure 2.7 obtained from [83] illustrates the result of the Voronoi estimate.

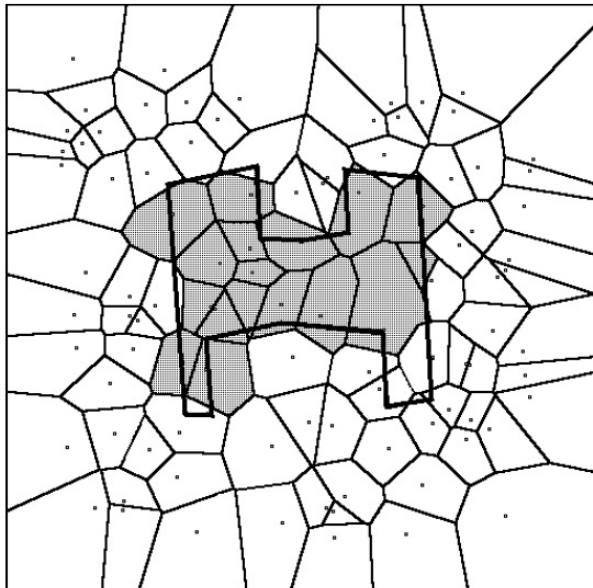


Figure 2.7: Voronoi estimate of nonconvex D . The nonconvex set D is outlined with a solid black line, and the estimate of the set \hat{D} is shaded in grey.

2.5 Concluding remarks

In this chapter, we gave attention to set estimation techniques that involved the estimation of a set when a realization of points from this set (and points outside the set) are observed, which has an analogue, in

\mathbb{R}^2 , to the task of constructing a window domain W from an observed point pattern. In instances detailed in Section 2.2, estimates were derived under assumptions of a homogeneous Poisson process and a convex domain. In Section 2.4 we considered a Voronoi solution to nonconvex set estimation that did not rely on a convexity assumption of the set domain. To our knowledge, these are the only approaches for spatial window estimation detailed in literature. The assumptions used to derive the results may not necessarily hold in some spatial applications (i.e. as discussed in Chapter 1). The solutions presented also do not incorporate the effect of covariates that may influence where points on the domain can be observed. In the next chapter we present an algorithm for choosing the window domain using additional spatial covariate information.

Chapter 3

Covariate construction of nonconvex sets

In Chapter 2, we considered window selection methods that involved the estimation of a set when a realization of points from this set (and points outside the set) are observed. However, in Chapter 1 it was noted that these sets may not be suitable for applications in the real world where points may be observed over window domains that have a complex polygon structure that could contain holes, be disconnected, or nonconvex. Generalization of these sets to such window domains may cause spurious estimation and prediction for regions where point occurrences were not observed, but for which it has not been confirmed that observations could have occurred there. The regions empty of observed data could be areas where for some geographical (or other) reason, expressed as the function of a spatial covariate of an underlying process, the phenomenon does not occur.

The main objective of point pattern analysis is to investigate small-scale spatial interactions between objects [4]. There are a large number of cases, however, in which the locations and marks of objects are impacted by extrinsic factors. In the case where these factors are spatially correlated, the resulting point pattern may show spatial correlation in the locations or marks even without intrinsic spatial interactions among the objects. It is thus important to investigate the effects of spatial covariates on the point pattern. The covariates based on some phenomena may make the point occurrence in regions of the domain very small or close to zero. In this chapter, we present an algorithm for choosing the window domain using spatial covariates.

We begin first by stating the definition of a spatial covariate. A spatial covariate is a spatial measurement defined at each point location in the window domain. It customarily describes a continuous regionalised variable defined at every point in the window. This may be denoted by $Z(u_j)$, $u_j \in W$, for $j = 1, \dots, l$,

where the points u_j form a lattice and do not coincide with the points of the point pattern. A few examples of spatial covariates include temperature, wind speed, soil pH level and elevation. Figure 3.1 depicts a plot of the spatial distribution of average wind speed (i.e. the spatial covariate) over the East China Sea for spring, summer, autumn and winter¹. The colours indicate wind speed and the arrows wind direction.

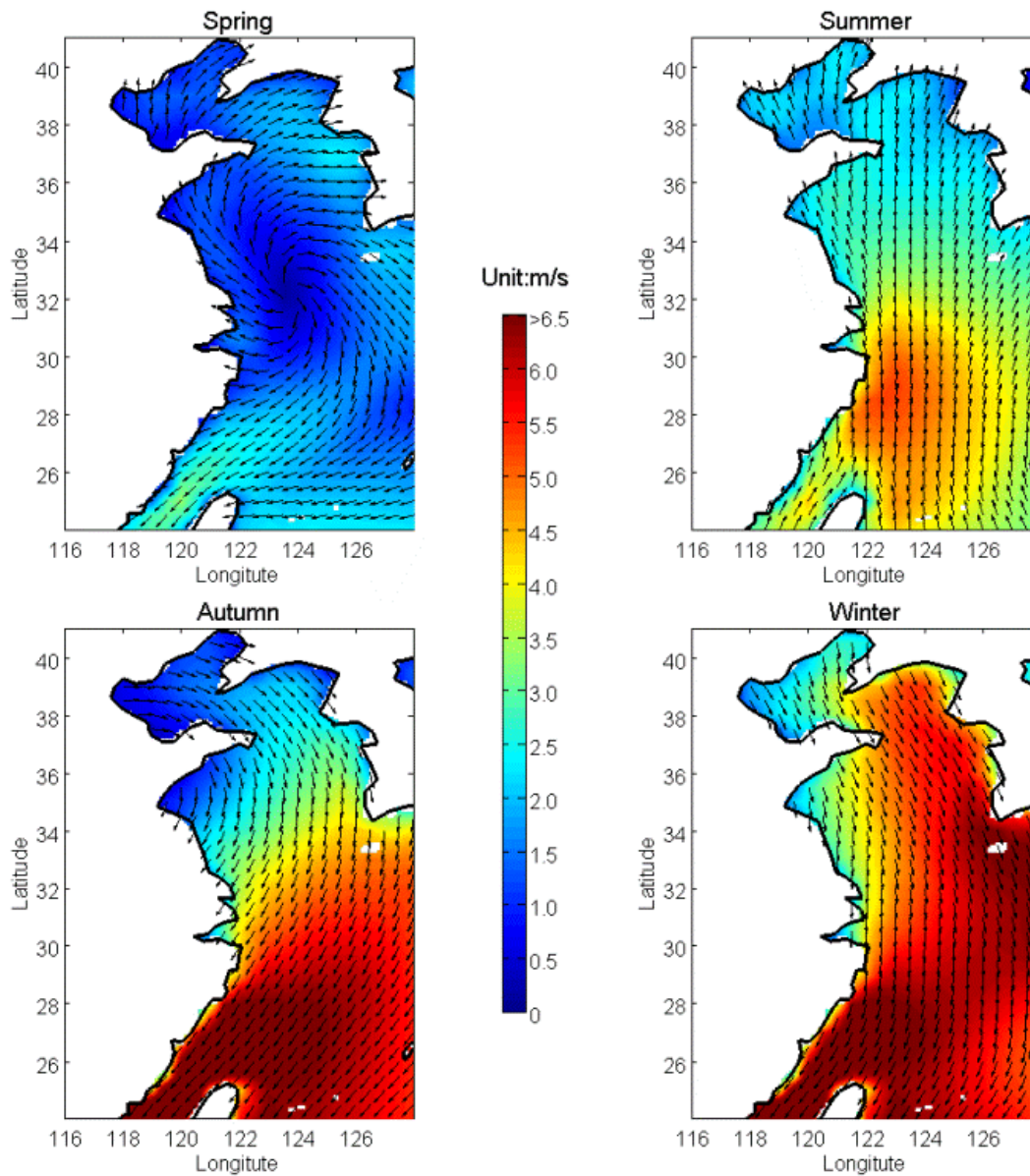


Figure 3.1: Seasonal spatial distribution of the wind vector data in 2003-2013 over the East China Sea. Each point of the wind vector data 10m over the sea surface are averaged according to the latitude and longitude in the same month from 2003 to 2013. The colours indicate the average wind speed and the arrows indicate wind direction.

¹<http://article.sciencepublishinggroup.com/html/10.11648/j.ijema.20160403.15.html>. Date Accessed: 03 March 2017

3.1 Testing dependence of a point pattern from spatial covariates

When covariate information is available, an investigation into the dependence of a point pattern on covariate data should be conducted and this dependence quantified. In this section we consider some methods and formal statistical tests for investigating the dependence of a point pattern on spatial covariates.

3.1.1 Quadrat defined by a covariate

One method of getting a rough understanding of how a covariate relates to a point pattern is to use quadrat methods [4]. The window W is divided into irregular quadrats of equal area defined by the bin ranges determined from the covariate quantiles. The number of points in each quadrat is then counted. If the point pattern process is uniform, that is, a homogeneous Poisson process, the expectation is that the number of points in each quadrat should be nearly equal to each other. Consider the tropical rainforest dataset shown in Figure 3.2. The dataset has the point record of the tree locations and is accompanied by the terrain elevation over arbitrary points in the study window. The study region is split into irregular subregions according to the quartiles of the elevation values. Figure 3.2 depicts the result of this tessellation created in R using the `spatstat` package [4, 76]. Since the regions have equal area, the number of points in each quadrat should be approximately equal if there is uniform density in the trees. The counts however indicate a strong preference for higher elevation suggesting that the elevation covariate influences the underlying pattern. This method gives a visual idea of how a covariate relates to a point pattern, but is not a formal test.

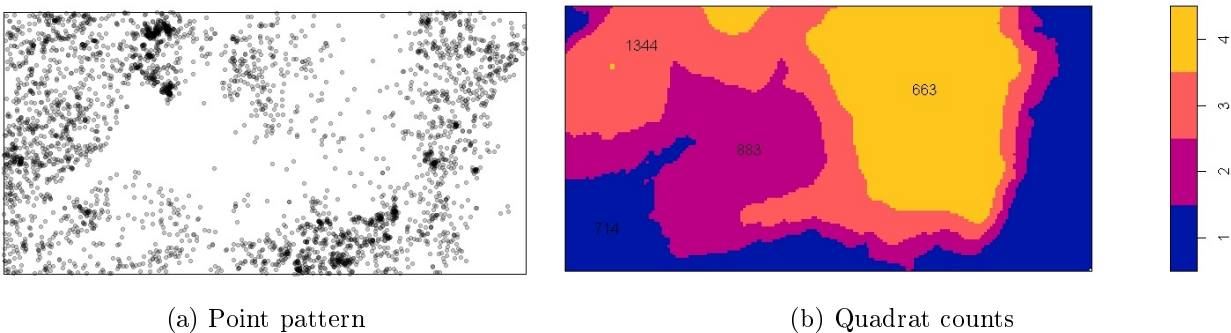


Figure 3.2: Point pattern plot, in (a), denoting the locations of 3605 trees in a tropical rain forest in a 1000 by 500 meter study area in Barro Colorado Island, and equal area tessellation of the study area, in (b), to the quartiles of terrain elevation [4].

3.1.2 Berman's test

Berman's test for the dependence of a point process on a spatial covariate [4] is a test that evaluates the hypothesis that the pattern was generated by a homogeneous Poisson process independent of the covariate against the alternative that the pattern is an inhomogeneous Poisson distribution with an intensity that depends on the covariate. The test statistic given by

$$Z_1 = \frac{S - \mu}{\sigma}$$

is used to perform the test and is based on $S = \sum_i Z(\mathbf{x}_i)$ the sum of the covariate values at the observed locations \mathbf{x}_i , where μ is the predicted mean value of S under n realizations of the null model and σ^2 the associated variance. Under the null hypothesis the test statistic is approximately standard normal distributed. Significant deviation of the test statistic Z_1 from the null distribution can be assessed by comparing the value observed for Z_1 to the standard normal percentiles.

3.1.3 Cumulative Distribution Function test (CDF)

The CDF test [4] is an adapted goodness-of-fit test for spatial point patterns. The test compares the distribution of the covariate at the observed data points with that of the values of the covariate at all the other spatial locations in the observation window. The idea here is that if the point process is completely random, then the observed points are a random sample of spatial locations in the window, so the values of the covariates at the data points should be a random sample of the values of the covariate at all spatial locations in the window. The cumulative distribution functions $\hat{F}(z)$ and $F_0(z)$ are compiled from the covariate values evaluated at the observed data points and every spatial location in the window respectively. $\hat{F}(z)$ denotes the empirical cumulative distribution function of the covariate values at the observed data points and is expressed as

$$\hat{F}(z) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(z_i \leq z),$$

where $\mathbf{1}(\cdot)$ is an indicator function. $F_0(z)$ is the cumulative distribution function of the covariate values at every location u in the spatial window W . In practice, the covariate values are evaluated at the center of every pixel in a grid. Thus,

$$F_0(z) = \frac{|\{u \in W : Z(u) \leq z\}|}{|W|} \approx \frac{\#\{\text{pixels } u : Z(u) \leq z\}}{\#\text{pixels}}.$$

The Kolmogorov-Smirnov statistic

$$D = \max|\hat{F}(z) - F_0(z)|$$

is then used to measure the discrepancy between $\hat{F}(z)$ and $F_0(z)$. The null distribution of the test statistic depends only on the number of data points under the assumptions that F_0 is continuous and that the point process is a homogeneous Poisson process.

3.2 Window construction via covariates

The methods for set estimation detailed in Chapter 2 typically give estimators that only depend on point observations from the unknown set. Estimators derived may condition on some additional assumption about the homogeneity of the underlying point process or a convexity restriction on the point process domain. The algorithm we now propose considers the construction of a nonconvex window domain that is built by incorporating the effect of the covariate on the occurrence of a point. The aim is to select a window domain filtered of all the areas for which the chance of a point occurrence is known to be equal to or close to zero, and to determine this based on covariate information. To this end, suppose x_1, \dots, x_N denote the set of unique points observed from the unknown, not necessarily convex, domain $W \subset \mathbb{R}^2$. Let M_j denote a moving window centered at a location u_j contained in W^* , where W^* is a large convex area that extends over the true domain W . Without loss of generality we suppose here that W^* is the smallest bounding rectangular window that contains all the observed points and that the moving window M_j is a square region given by

$$M_j = \{\mathbf{u} \mid \|\mathbf{u}_j - \mathbf{u}\|_1 \leq \frac{1}{g}d, g \in \{1, \dots, m\}\},$$

that is, M_j is a square area centered at \mathbf{u}_j . The size of M_j is chosen as a function of the minimum distance $d = \min_{i \neq k} \{\|\mathbf{x}_i - \mathbf{x}_k\|\}$ of the point observations; chosen such that the number of points in M_i (i.e. the area centered at the i -th observed point) is one. If \mathbf{u}_j is chosen so that they are regularly spaced, then using the moving window corresponds to superimposing a grid over W^* with a grid cell size of $\frac{1}{g}d$. The moving window will visit each cell and compute a measure based on the covariates in the cell. Let $Z(u)$ denote a spatial covariate evaluated at $u \in W^* \subset \mathbb{R}^2$ and f be a function that computes a localised measure for subsets of W^* based on the spatial covariate $Z(u)$. The investigation of the dependence of the point pattern on the spatial covariate should precede the implementation of this algorithm. Obviously, it is desirable to use covariates believed to impact the distribution and abundance of the object of interest, or that is correlated to them. The tests in Section 3.1 should be used to confirm this. Let $F = \{f \mid \tau(f)\}$ be the set that satisfies some predicate $\tau(\cdot)$ based on f . The true domain W is then obtained as,

$$W = \bigcup_{\forall j, f_j \in F} M_j$$

where f_j is the function value evaluated using the covariates at locations in M_j .

The steps are presented in Algorithm 1. The idea here is to use the moving window to search over overlapping grid cells in the larger domain W^* and based on some decision rule that conditions on the covariates in the neighborhood of the observed data points, filter out the regions that do not meet the specified criterion.

Algorithm 1 Covariate construction of nonconvex window

1. Calculate $d = \min_{i \neq k} \{\|\mathbf{x}_i - \mathbf{x}_k\|\}$
 2. Create $M_i = \{\mathbf{u} \mid \|\mathbf{x}_i - \mathbf{u}\|_1 \leq \frac{1}{g}d, g \in \{1, \dots, m\}\}$
 3. Select points \mathbf{u} in M_i and corresponding $Z(\mathbf{u})$
 4. Calculate f using selected covariates $Z(\mathbf{u})$ in M_i and let these computed values be denoted by f_i
 5. Let $F = \{f \mid \tau(f)\}$ satisfy a chosen predicate $\tau(\cdot)$ such that all $f_i \in F$
 6. Create $M_j = \{\mathbf{u} \mid \|\mathbf{u}_j - \mathbf{u}\|_1 \leq \frac{1}{g}d, g \in \{1, \dots, m\}\}$
 7. Select points \mathbf{u} in M_j and corresponding $Z(\mathbf{u})$
 8. Calculate f using selected covariates $Z(\mathbf{u})$ in M_j and let these computed values be denoted by f_j
 9. True domain: $W = \bigcup_{j, f_j \in F} M_j$
 10. Iterate through the regions at the edges and reduce their size each time to reduce roughness of window
-

Algorithm 1 is illustrated with the toy example shown in Figure 3.3. The figures are created using R [76]. The point pattern could denote for example the locations of households in a rural settlement and the terrain elevation used as a spatial covariate. The spatial covariate surface is generated using Gaussian mixtures in this figure. In this setting, the point occurrence of a household would only be observed in regions that have viable building land.

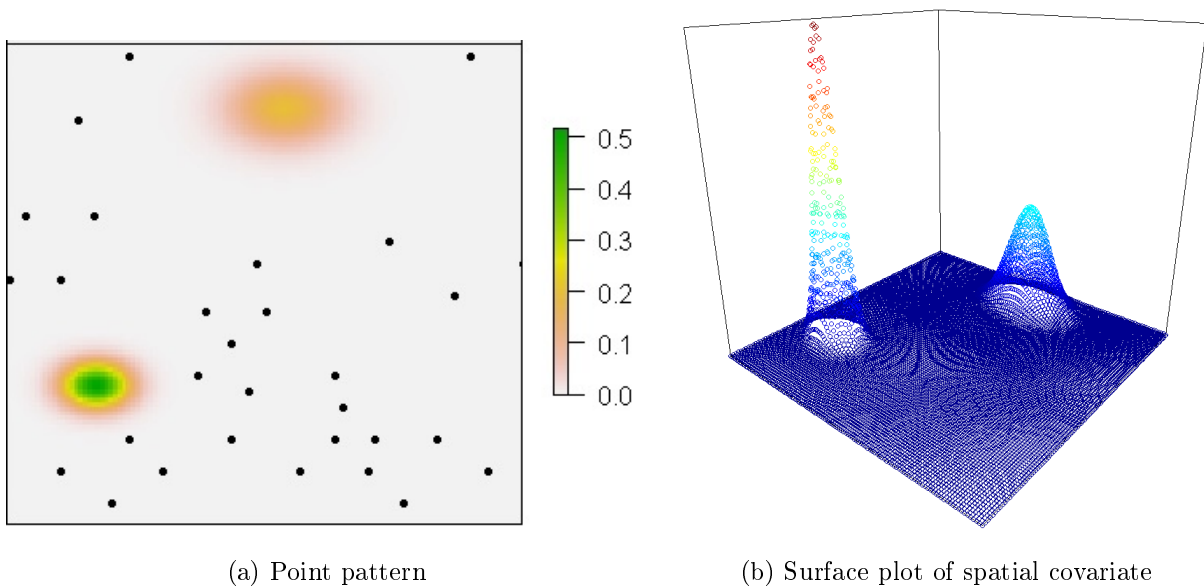


Figure 3.3: A toy example of a point pattern plot in (a) and spatial covariate surface in (b).

A feasible region for building could be characterized by the average change in elevation over areas where the points in the pattern were observed. In regions not suitable for building (i.e. areas for which the point occurrence of a household would not be observed), we expect that the average change in elevation will be more variable and thus outside the range of values observed over the points in the pattern.

In implementing the first step of Algorithm 1, we calculate the minimum Euclidean distance between the points in the observed pattern. In the next step, a square moving window is created that is centered over the points in the pattern and has a side length proportional to the minimum distance between two points. The only requirement in choosing the size of the moving window is that it should only contain a single point (its center) when centered on each observed point in the pattern. Choosing any value less than the minimum distance between two points observed in the pattern and greater than zero will achieve this. For our purposes, we arbitrarily choose half the minimum distance as the side length for the square window. Thus the window is given by

$$M_i = \{\mathbf{u} \mid \|\mathbf{x}_i - \mathbf{u}\|_1 \leq \frac{1}{2}d\}.$$

Figure 3.4 depicts an illustration of this step.

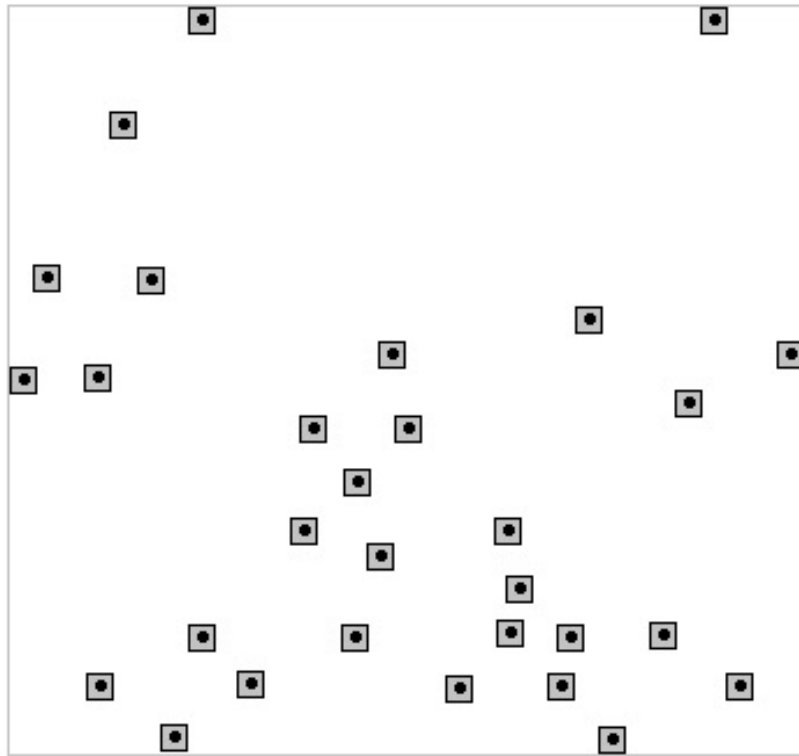


Figure 3.4: Illustration portraying M_i regions on toy example point pattern.

In the next step, points \mathbf{u} shown in Figure 3.5 are selected in each M_i as

$$\mathbf{u} = \mathbf{x}_i \pm \frac{d}{4}(s_1, s_2)$$

where $s_1, s_2 = 1, 3$. The points are selected such that they cover the area spanned by the moving window. The spatial covariate values $Z(\mathbf{u})$ corresponding to the locations \mathbf{u} and at the point \mathbf{x}_i are then determined.

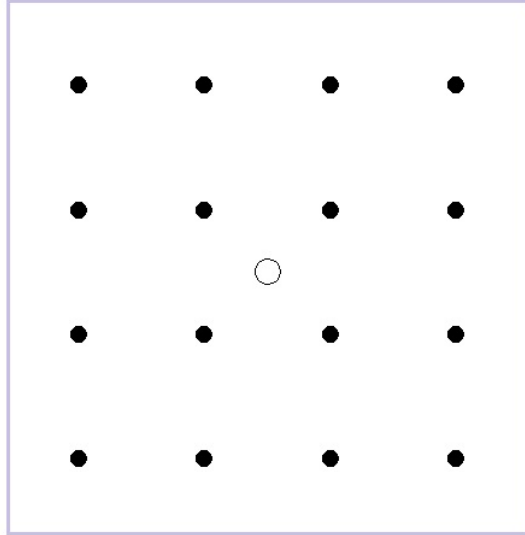


Figure 3.5: Illustration portraying the selected points \mathbf{u} (represented by closed dots), \mathbf{x}_i (represented by an open dot) and M_i enclosed in the square boundary.

The successive step entails evaluating a function of these spatial covariate values which is selected here as the sample variance and is given by

$$f_i(z_1, \dots, z_J) = \frac{1}{J-1} \sum_{j=1}^J (z_j - \bar{z})^2$$

where z_1, \dots, z_J denotes the observed values of the covariates corresponding to the J selected points \mathbf{u} in M_i and \bar{z} is the average of these values. We then specify a selection criterion that determines whether regions that do not coincide with the observed pattern should be included or excluded from the true domain. In the setting of our toy example these are the areas for which the variability in the covariate is outside the range given by the variability in covariate values for the M_i regions. Thus a region is included if the value of the function of the covariates in that area is an element of

$$F = \{f \mid \min_i \{f_i\} \leq f \leq \max_i \{f_i\}\}.$$

Steps 6-8 entail repeating the process of steps 2-4 for the areas that do not coincide with the points observed in the pattern. The true domain is then constructed as the union of these areas for which the

value of the f is contained in the set F . The result of this process is shown in Figure 3.6. The roughness of this window can be improved by iterating through the regions at the edges and reducing their size each time.

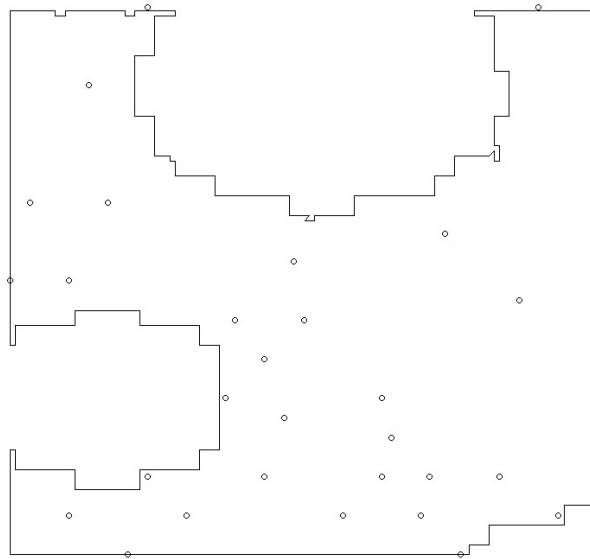


Figure 3.6: Point pattern plot over covariate constructed nonconvex window.

3.3 Concluding remarks

In this chapter, we presented an algorithm for selecting a nonconvex window for a point pattern when additional covariate information is available. The algorithm works by using a moving window to search over a larger domain than that of the true window and, using a function or feature of the spatial covariate in regions over points observed in the pattern, construct a nonconvex window. Test for the dependence of the point pattern on the spatial covariate should be evaluated and confirmed using the methods in Section 3.1. Covariates that impact the distribution and abundance of the points of interest are desired. The investigation of the dependence of the point pattern on the spatial covariate should precede the implementation of this algorithm. Testing for global and local dependence of point patterns on covariates in parametric models can also be done using the method proposed in [68]. The effect of how the proposed nonconvex window may improve the accuracy of spatial measures is the subject of the next chapter. Particular attention is given to kernel smoothed intensity estimates, a descriptive measure that captures the first order properties of a point pattern. The definition of first-order properties in the context of point pattern analysis was introduced in Chapter 1, we visit this definition again in the next chapter.

Chapter 4

Intensity estimation

In Chapter 1, we introduced and briefly discussed first-order properties for a spatial point pattern and how these can be expressed using the intensity function. In this chapter we expand on this discussion. All analysis, simulations and figures in this chapter are created using the `spatstat` package in R [4, 76].

4.1 Intensity function

The intensity of a point pattern is a descriptive measure analogous to the average of a population of numbers [4]. A query into the properties of the intensity of a process is often the aim of a scientific investigation and is a crucial part of point pattern analysis [51, 66]. For example, in forestry surveys the stand density, the number of trees growing per unit area, is an important quantity to be estimated [38]. A standard scientific question is whether the intensity varies spatially; is inhomogeneous or homogeneous. The spatial variation in intensity could indicate an inclination for point objects to cluster in certain regions of the study area. In an ecological setting for example, the intensity could demonstrate a preference of a species of animal for a certain habitat [109, 110, 111]. In other settings it may indicate the productivity of crops [89, 107], or areas of high crime prevalence [59, 39].

The intensity characterizes the first-order properties [36] of a spatial point process. First-order properties describe how the average of the process varies over different locations in space [36]. The intensity function $\lambda(\mathbf{x})$ [36] at position $\mathbf{x} \in W$, where W is the spatial domain of the observed point pattern, is defined as the average number of points per unit area and is denoted by the expression,

$$\lambda(\mathbf{x}) = \lim_{dx \rightarrow 0} \left\{ \frac{E(N(d\mathbf{x}))}{dx} \right\} \quad (4.1)$$

where $d\mathbf{x}$ is a small region around \mathbf{x} , $E(\cdot)$ is the expected value operator, dx is the area for this region,

and $N(\mathbf{dx})$ refers to the number of events in the region \mathbf{dx} . A point pattern that has constant intensity $\lambda(\mathbf{x}) = \lambda$ over varying locations is termed first-order homogeneous whereas inhomogeneous point patterns have non-constant intensity functions varying with \mathbf{x} . In the case of a homogeneous point process, the intensity has an unbiased empirical estimator [4],

$$\hat{\lambda} = \frac{\text{number of points}}{\text{area of } W},$$

where W denotes the spatial domain of the observed point pattern. The intensity $\lambda(\mathbf{x})$ is a point definition quantity, that in practice, can only be estimated by the available data. Values of the intensity at certain locations in W , are directly estimated from neighbouring data and the resultant estimates presented. The intensity function may be estimated using parametric or nonparametric methods. In the case of a parametric approach, an underlying parametric model is proposed and the parameter estimates for the model determined via optimal criteria such as maximum likelihood [16, 65]. The optimisation is done in the global scale over the whole region W being considered. As an example, the likelihood function for an inhomogeneous Poisson point process [16] can be expressed as

$$l(\boldsymbol{\theta}) = \left\{ \prod_{i=1}^n \lambda_{\boldsymbol{\theta}}(\mathbf{x}_i) \right\} \exp \left(- \int_W \lambda_{\boldsymbol{\theta}}(\mathbf{u}) v(d\mathbf{u}) \right)$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are the data locations, n is the number of observed data points in the spatial domain W , $\lambda_{\boldsymbol{\theta}}(\cdot)$ is the intensity function and $\boldsymbol{\theta}$ a vector of parameters. The parameter estimates $\hat{\boldsymbol{\theta}}$ will be the set of values that then maximise the given likelihood function.

Nonparametric estimation is data-driven and the estimation is only concerned with local scale variations. A nonparametric approach of estimation may be preferred over a parametric approach of estimation since it requires few modelling assumptions that may not necessarily hold in practice. Two main techniques for nonparametric intensity estimation are based on quadrat density and kernel density. We give a discussion below of intensity estimation via quadrat densities and dedicate the remainder of Chapter 4 to smooth estimates for the intensity function based on kernel functions. The reader is referred to [16, 65] for a more extensive discussion of parametric methods.

4.2 Quadrat counts

Quadrat intensity estimates [4] give a simple estimate of the intensity function. The spatial domain or study area W is divided into L subregions $\{s_l : l = 1, \dots, L\}$, termed quadrats. The quadrats may have equal area, but this is not necessarily required. Partial quadrats may also arise at the boundary of the observation window resulting from a non-rectangular or irregular shaped study window. The point intensity is computed for each quadrat by dividing the number of points in the quadrat by the quadrat's

area, that is,

$$\hat{\lambda}(s_l) = \frac{\text{number of points in } s_l}{\text{area of } s_l}.$$

Quadrats can take the form of any shape, for example, rectangles, hexagons, triangles, etc, but at each application are all the same shape usually. The quadrat size has an impact on the estimated intensity and should be chosen carefully. Small quadrats may yield spiked intensity maps with many empty quadrats, the results of which may be uninformative. Large quadrats produce smoother intensity maps but may mask subtle changes in the distribution of spatial intensity. Consider Figures 4.1 and 4.2. Figure 4.1 shows a point pattern simulated from a homogeneous Poisson process with $\lambda = 2$ and Figure 4.2 a point pattern simulated from an inhomogeneous point pattern with intensity of the form,

$$\lambda(\mathbf{x}) = 30 \exp\left(-\frac{1}{4}(x_1 - x_2)\right). \quad (4.2)$$

Both are simulated over a square window domain with a side length of 10.

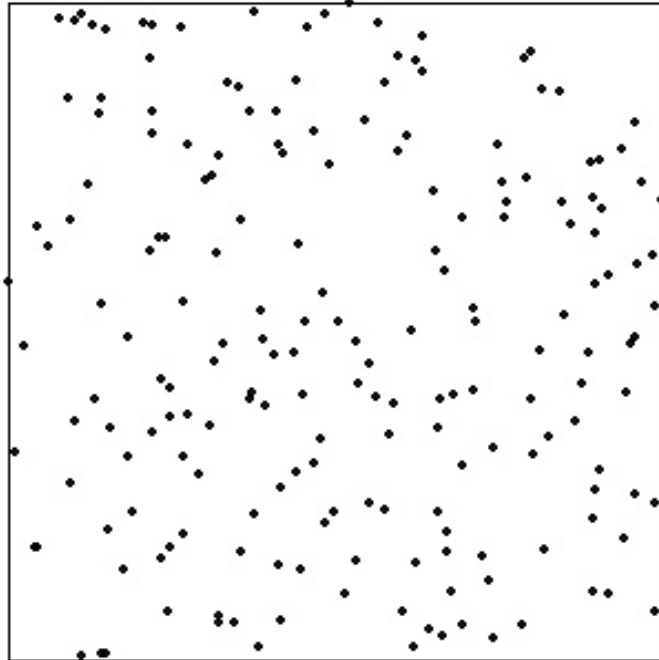


Figure 4.1: Point pattern simulated from a homogeneous Poisson process with $\lambda = 2$ over a square window with a side length of 10.

The quadrat intensity estimates with varying quadrat sizes for the homogeneous and the inhomogeneous point patterns are shown in Figures 4.3 and 4.4 respectively. Figures 4.5 and 4.6 depict the respective quadrat intensity estimates with varying options for the quadrat shapes.

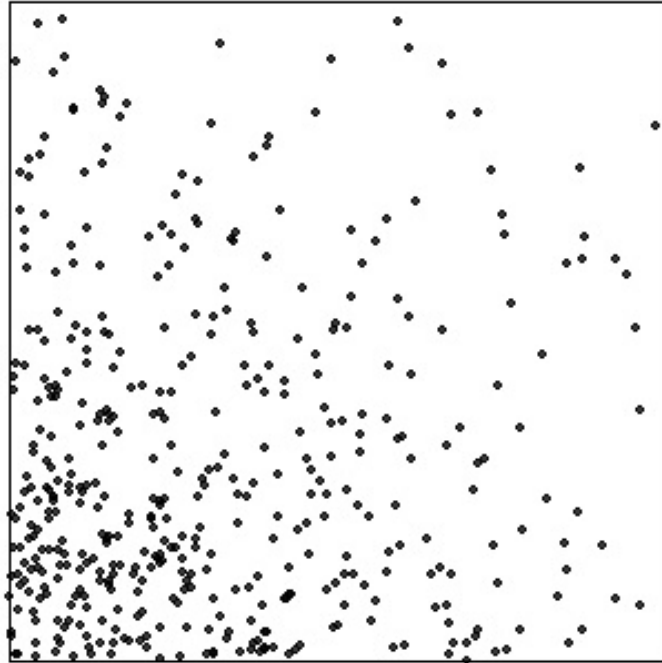
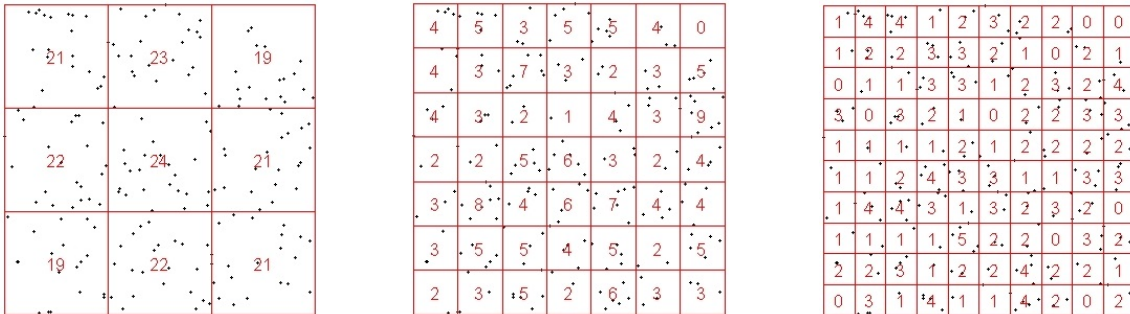
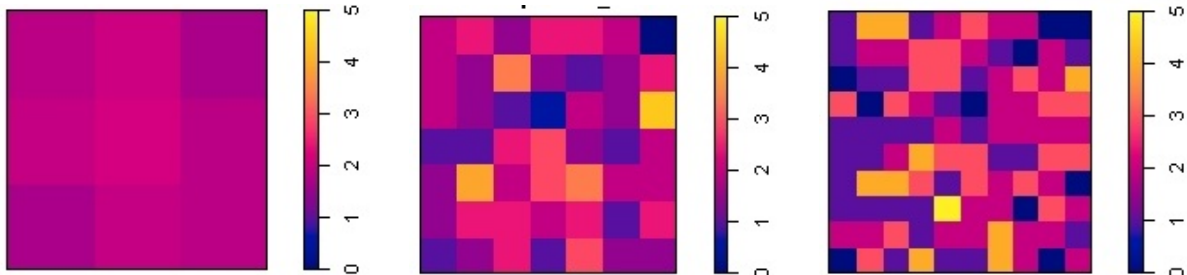


Figure 4.2: Point pattern simulated from an inhomogeneous Poisson process with $\lambda(\mathbf{x})$ taking the form of Equation 4.2, over a square window with a side length of 10.

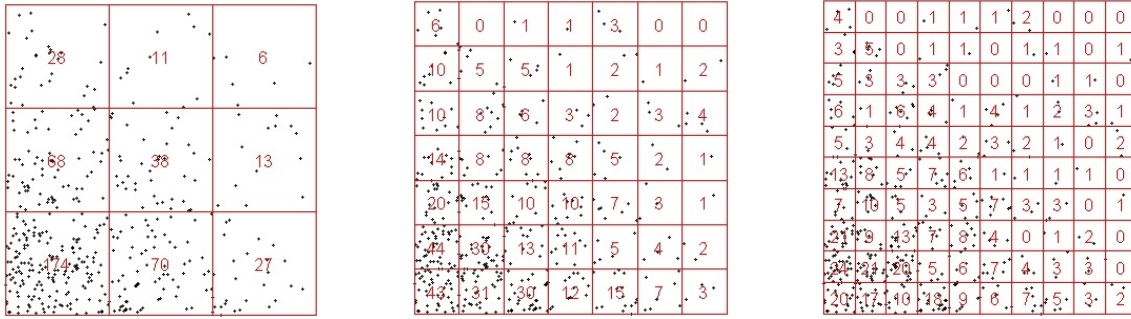


(a) Quadrat counts for 3×3 , 7×7 and 10×10 grids

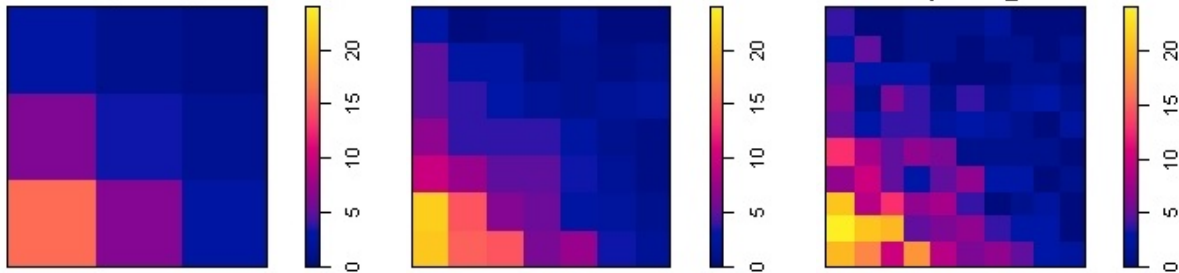


(b) Intensity estimates for 3×3 , 7×7 and 10×10 grids

Figure 4.3: Quadrat counts and corresponding intensity plots with square quadrats of various sizes for simulated pattern depicted in Figure 4.1.

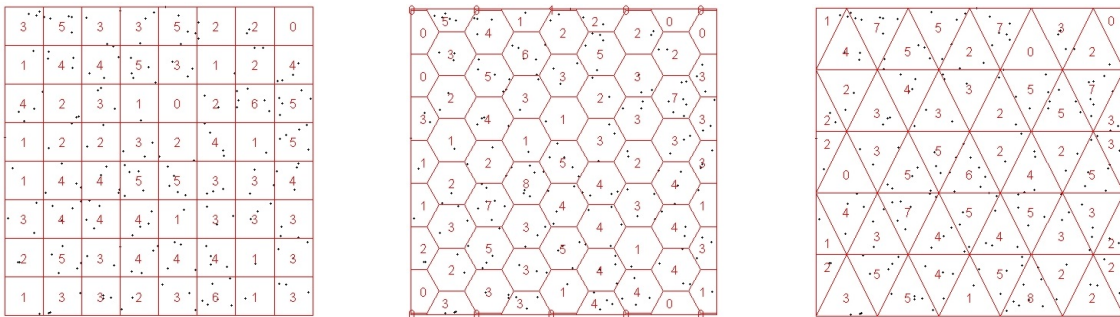


(a) Quadrat counts for 3×3 , 7×7 and 10×10 grids

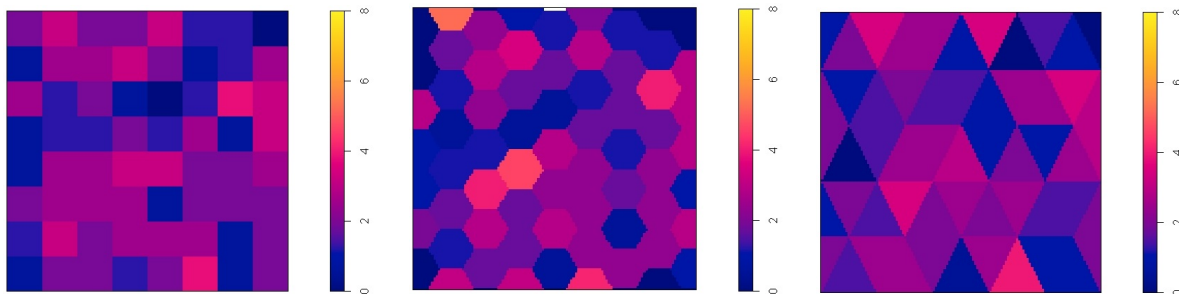


(b) Intensity estimates for 3×3 , 7×7 and 10×10 grids

Figure 4.4: Quadrat counts and corresponding intensity plots with square quadrats of various sizes for simulated pattern depicted in Figure 4.2.

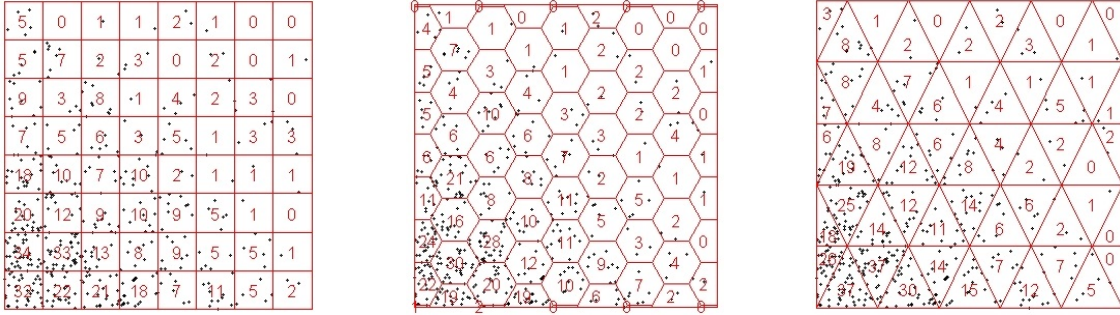


(a) Quadrat counts for square, hexagonal and triangular quadrats

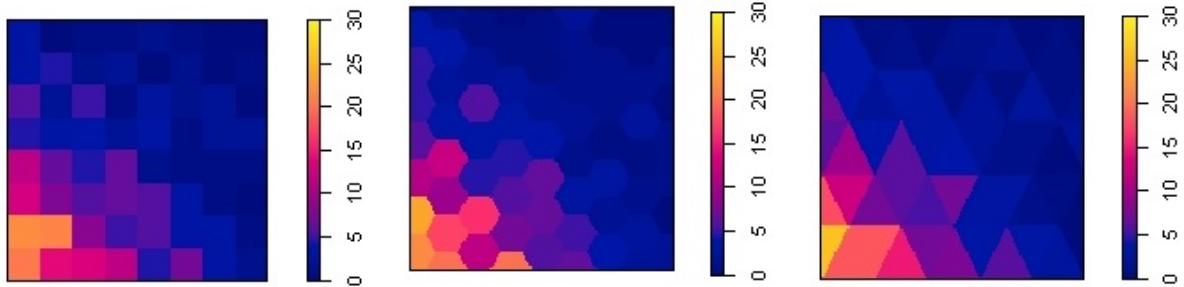


(b) Intensity estimates for square, hexagonal and triangular quadrats

Figure 4.5: Quadrat counts and corresponding intensity plots with quadrats of various shapes with equal area for simulated pattern depicted in Figure 4.1.



(a) Quadrat counts for square, hexagonal and triangular quadrats with equal area



(b) Intensity estimates for square, hexagonal and triangular quadrats with equal area

Figure 4.6: Quadrat counts and corresponding intensity plots with quadrats of various shapes for simulated pattern depicted in Figure 4.2.

4.3 Kernel smoothed intensity estimation

The kernel estimate $\hat{\lambda}(\mathbf{x})$ for the intensity at the point \mathbf{x} , calculated from observed locations $\mathbf{x}_1, \dots, \mathbf{x}_n$ has the standard form given by [4, 16, 23],

$$\hat{\lambda}(\mathbf{x}) = \sum_{i=1}^n \frac{1}{h^2} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (4.3)$$

Here $K(\cdot)$ denotes the kernel function, defined for 2-dimensional \mathbf{x} such that $K(\mathbf{x}) \geq 0$ and

$$\int_{\mathbb{R}^2} K(\mathbf{x}) d\mathbf{x} = 1.$$

In Equation 4.3 the distances between \mathbf{x} and each point \mathbf{x}_i that lies within a region controlled by the bandwidth, h , is measured and the contribution to the intensity determined by how close the point is to \mathbf{x} . An edge corrected kernel estimator [23] for the intensity is given by

$$\hat{\lambda}(\mathbf{x}) = \frac{1}{p_h(\mathbf{x})} \sum_{i=1}^n \frac{1}{h^2} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \quad (4.4)$$

where $p_h(\mathbf{x}) = \int_W \frac{1}{h^2} K\left(\frac{\mathbf{x} - \mathbf{u}}{h}\right) d\mathbf{u}$ is an edge correction term for edge effects (discussed in Chapter 1).

The kernel function $K(\cdot)$ will usually be a radially symmetric unimodal probability density function centered at the origin. Consequently, the kernel is any non-negative function that is integrable over its whole domain with a value equal to one, centered at zero, symmetric about its center, and has first and second moments that exist. Examples include the standard bivariate normal density function

$$K(\mathbf{x}) = (2\pi)^{-1} \exp(-\mathbf{x}^T \mathbf{x}),$$

and the bivariate Epanechnikov kernel

$$K(\mathbf{x}) = \frac{2}{\pi} (1 - \mathbf{x}^T \mathbf{x}) \mathbf{1}(\mathbf{x}^T \mathbf{x} < 1)$$

where $\mathbf{1}(\cdot)$ is an indicator function [97]. Figure 4.7 depicts examples of bivariate kernel functions with the stated properties.

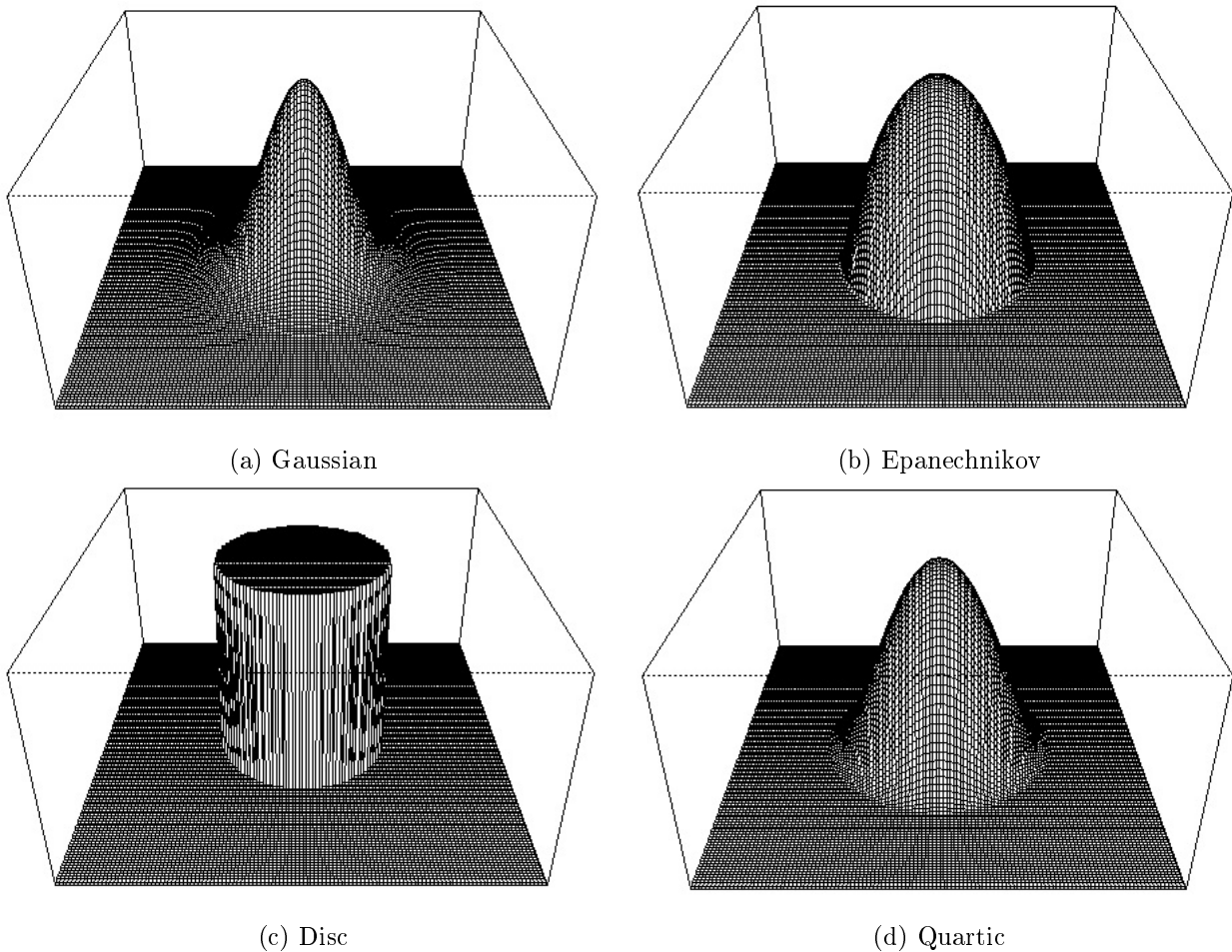


Figure 4.7: Examples of bivariate kernel functions

$K(\cdot)$ could also be of the form of a product kernel $K(\mathbf{u}) = K(u_1) \cdot K(u_2)$. Table 4.1 gives several examples of univariate kernel functions [48, 97] with the properties described above.

Kernel	$K(t)$ for $t \in \mathbb{R}$
Epanechnikov	$\frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2)$ for $ t < \sqrt{5}$, 0 otherwise
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$ for all t
Uniform	$\frac{1}{2a}$ for $-a \leq t \leq a$, 0 otherwise
Triangular	$1 - t $ for $ t < 1$, 0 otherwise
Logistic	$\frac{1}{e^t + 2 + e^{-t}}$
Silverman kernel	$\frac{1}{2}e^{-\frac{ t }{\sqrt{2}}} \sin\left(\frac{4 t + \sqrt{2}\pi}{\sqrt{36}}\right)$

Table 4.1: Examples of univariate kernel functions [48, 97] commonly utilized

The estimation of the intensity of a spatial point process slightly differs from that of a probability density in that the intensity function is not required to integrate to unity, so is not normalised. In the context of a spatial analysis it may be inappropriate to assume a Poisson point process, corresponding to assuming independent and identically distributed random variables in density estimation. This assumption is often used in the derivation of error and standard error results for the density estimate [97].

The bandwidth is denoted by h and controls the level of smoothing. Utility of a single value of the bandwidth implies that the kernel function placed on each observed point is scaled equally in all directions. Some applications, beyond the scope of this mini-dissertation may require the use of a matrix of smoothing parameters allowing for non-constant directional variation [42, 64, 77].

The selected bandwidth is directly responsible for the approximation accuracy [97]. Silverman [97] derived expressions for the approximations of the bias and variance under certain assumptions for the multivariate kernel density estimate,

$$\hat{f}_h(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right),$$

of the density function $f(\mathbf{x})$, where h and $K(\cdot)$ are the bandwidth and kernel function respectively, $K(\cdot)$ is now defined for \mathbb{R}^d , and n is the number of observed data points \mathbf{x}_i contained in \mathbb{R}^d . The derived expression for the approximation of bias is

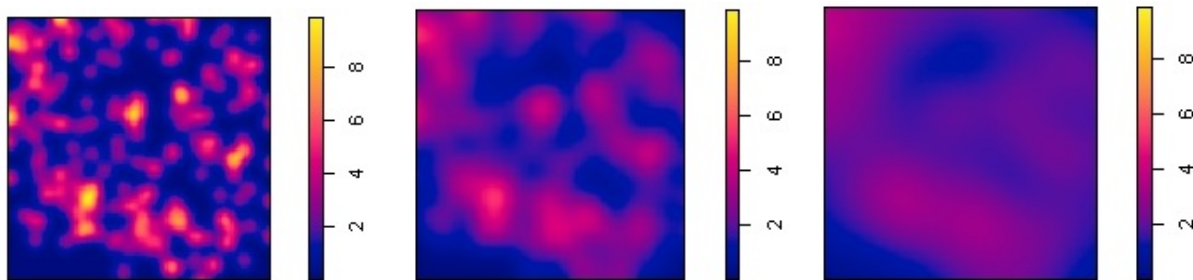
$$bias(\hat{f}_h(\mathbf{x})) \approx \frac{1}{2}h^2 f''(\mathbf{x}) \int \mathbf{t}^T \mathbf{t} K(\mathbf{t}) d\mathbf{t},$$

and for the variance,

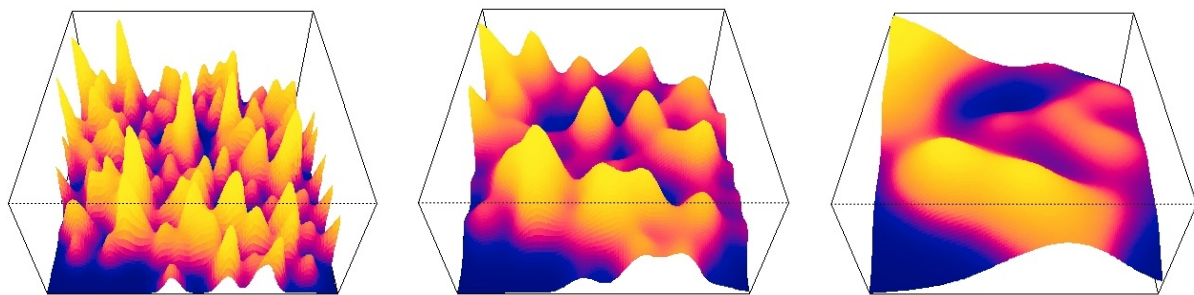
$$\text{var}(\hat{f}_h(\mathbf{x})) \approx \frac{1}{nh} f(\mathbf{x}) \int \{K(\mathbf{t})\}^2 dt.$$

Here $f''(\mathbf{x})$ denotes the second order derivative of $f(\mathbf{x})$. The choice of bandwidth involves a trade-off between bias and variance. As the size of the bandwidth increases the bias increases and the variance decreases.

The choice of h , the bandwidth, is crucial in estimating the kernel smoothed intensity function [17, 23, 27, 97]. A large choice for h will mask the structure of the data and over-smooth the density estimate, reduce precision and decrease variance, whereas, a small h will under-smooth the density estimate, increase precision and increase variance. The effects of the choice of bandwidth are illustrated in Figures 4.8 and 4.9 for the simulated homogeneous point pattern in Figure 4.1, and the simulated inhomogeneous point pattern in Figure 4.2 respectively. As seen in the figures, when the choice of bandwidth is relatively small (i.e. $h = 0.25$), the kernel smoothed intensity estimate will fit the data well, but the intensity estimate function will be jagged and spiked. If we choose a larger bandwidth, say $h = 1$, we get a smooth function for the intensity estimate. A good choice for h would be a value that gives a smooth estimate for the intensity function and that captures the distribution of the data.

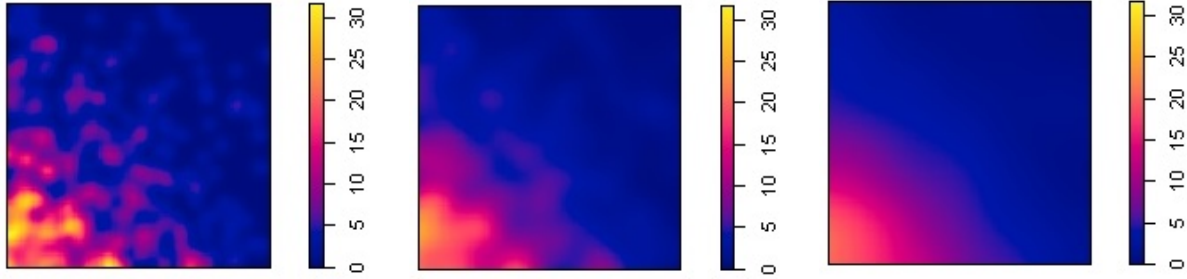


(a) Kernel smoothed intensity estimate for bandwidth size 0.25, 0.5 and 1

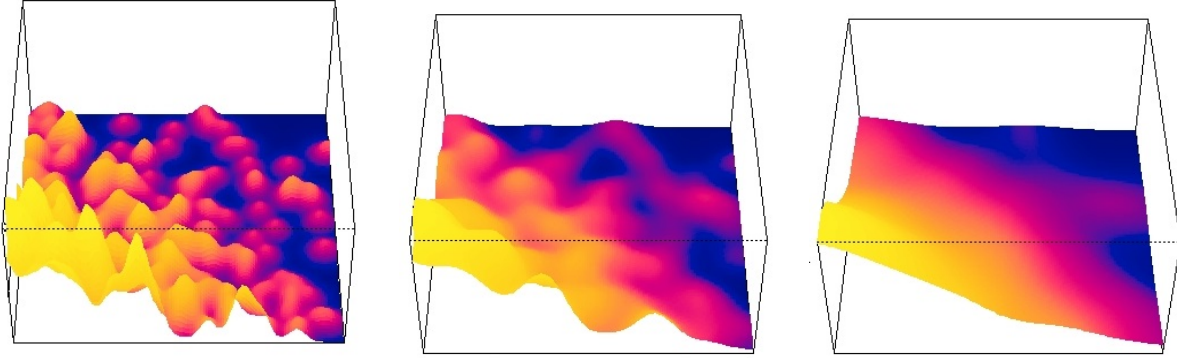


(b) 3D perspective kernel smoothed intensity estimate plots for bandwidth size 0.25, 0.5 and 1

Figure 4.8: Kernel smoothed intensity estimates and corresponding 3D perspective plots are shown in panel (a) and (b) respectively with bandwidths of various sizes for the simulated homogeneous pattern depicted in Figure 4.1.



(a) Kernel smoothed intensity estimate for bandwidth size 0.25, 0.5 and 1



(b) 3D perspective kernel smoothed intensity estimate plots for bandwidth size 0.25, 0.5 and 1

Figure 4.9: Kernel smoothed intensity estimates and corresponding 3D perspective plots are shown in panel (a) and (b) respectively with bandwidths of various sizes for the simulated inhomogeneous pattern depicted in Figure 4.2.

The form of the d -dimensional multivariate kernel density estimator [26, 27] for a random sample $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ is

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = n^{-1} \sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \quad (4.5)$$

where

- $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2} \mathbf{x})$, and $K : \mathbb{R}^d \rightarrow \mathbb{R}$, is a kernel function that takes the argument $\mathbf{x} \in \mathbb{R}^d$.
- n is the number of d -dimensional vectors observed.
- \mathbf{H} is a $d \times d$ bandwidth matrix, which is fixed, symmetric and positive definite¹. As in the univariate case, the level of smoothing of the kernel density function is mostly determined by the bandwidth matrix.

Depending on the choice of bandwidth matrix, the kernel density function can either be too smooth and inaccurately represent the data, under-smoothed and representative of the data, or smooth and a good

¹The determinant of a positive definite matrix is always positive, so a positive definite matrix is always nonsingular and invertible [9]

fit for the data set. There are two main cases of the forms that the bandwidth matrix can assume: The bandwidth matrix is symmetric and positive definite with the form

$$\mathbf{H} = \begin{bmatrix} h_1^2 & h_{12} & \cdots & h_{1d} \\ h_{12} & h_2^2 & \cdots & h_{2d} \\ \cdots & \cdots & \ddots & \cdots \\ h_{1d} & h_{2d} & \cdots & h_d^2 \end{bmatrix}. \quad (4.6)$$

The bandwidth matrix is diagonal, and positive definite with the form

$$\text{diag}(\mathbf{H}) = \begin{bmatrix} h_1^2 & 0 & \cdots & 0 \\ 0 & h_2^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & h_d^2 \end{bmatrix}. \quad (4.7)$$

If we let $h_1 = h_2 = \dots = h_d$, the bandwidth matrix reduces to

$$\mathbf{H} = h^2 \mathbf{I}_d = \begin{bmatrix} h^2 & 0 & \cdots & 0 \\ 0 & h^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & h^2 \end{bmatrix}, \quad (4.8)$$

where \mathbf{I}_d denotes a $d \times d$ identity matrix. For this case of the bandwidth matrix, Equation 4.5 reduces to

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x_1 - x_{i1}}{h}, \frac{x_2 - x_{i2}}{h}, \dots, \frac{x_d - x_{id}}{h}\right).$$

In the case of a symmetric and positive definite bandwidth matrix of the form given in Equation 4.6, the effect of the kernel function placed on each observed point has non-constant directional variation. This indicates that the points in the pattern have directional dependence. In the case of the diagonal matrix of the form given in Equation 4.7, there is no directional dependence, and the kernel function placed on each observed point is scaled equally or unequally (i.e. $h_j = h$ for all $j = 1, \dots, k$ or at least one $h_j \neq h_i$ when $i \neq j$). Figure 4.10 shows bivariate Gaussian kernel functions where the bandwidth matrix has the three forms discussed and are chosen as

$$H_1 = \begin{bmatrix} 0.75 & 0.1875 \\ 0.1875 & 0.25 \end{bmatrix},$$

$$H_2 = \begin{bmatrix} 0.75 & 0 \\ 0 & 0.25 \end{bmatrix},$$

and

$$H_3 = \begin{bmatrix} 0.1875 & 0 \\ 0 & 0.1875 \end{bmatrix}.$$

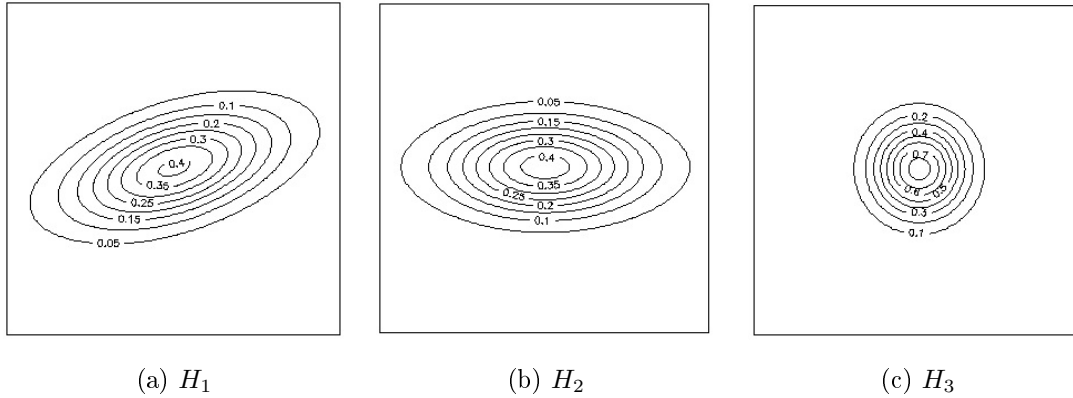


Figure 4.10: Contours of bivariate Gaussian kernel function for varying choices of the bandwidth matrix.

The following example is based on a point pattern created by combining translated realizations of a homogeneous Poisson point process, with $\lambda = 2$, simulated over an elliptical domain rotated by 45 degrees in an anticlockwise direction. The point pattern is given over a rectangular window shown in Figure 4.11(a). The kernel smoothed intensity estimate for the point pattern is fitted using a bivariate Gaussian kernel with the bandwidth matrix H_1 that was used in the preceding example. Figure 4.11(b) shows the result.

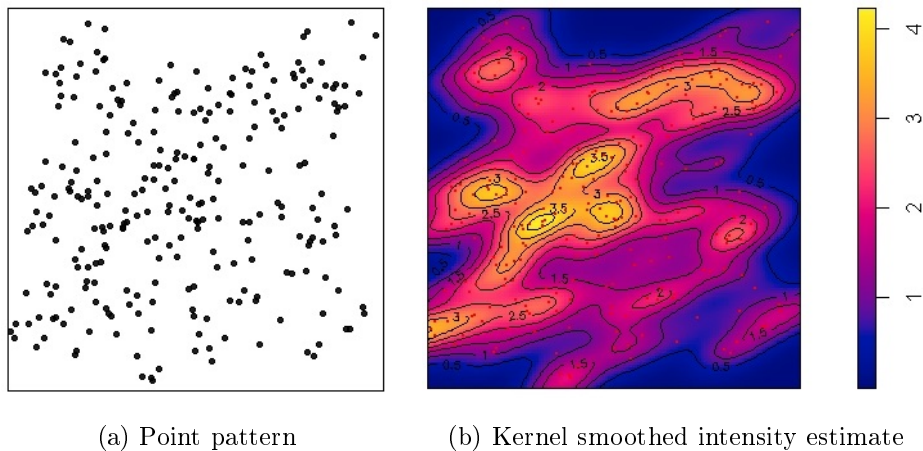


Figure 4.11: Point pattern (points in pattern shown with black dots) and bivariate Gaussian kernel smoothed intensity plot are shown in panel (a) and (b) respectively. Points in pattern are shown with red dots in panel (b). Contour lines are also indicated in the kernel smoothed intensity plot in panel (b).

4.4 Intensity as a function of a covariate

In Chapter 3 a spatial covariate $Z(\mathbf{u})$ was defined as a spatial measurement or continuous regionalised variable defined at every point location \mathbf{u} in the window domain W . Intensity estimation can be extended

to allow for the effect of spatial covariates on the distribution of the point pattern. In this section we give consideration to parametric and nonparametric strategies for modelling the intensity dependent on some underlying covariate. In each case, it is assumed that the intensity is a function of spatial covariates. That is, it is assumed that

$$\lambda(u) = \rho(\mathbf{Z}(u)),$$

where ρ is a function expressing how the intensity of points relate to the covariate values $\mathbf{Z}(u)$. A typical parametric model for the relationship between the covariate $\mathbf{Z}(u)$ and the intensity $\lambda(u)$ is the loglinear model [2]

$$\lambda(u) = \exp(\boldsymbol{\beta}^T \mathbf{Z}(u)),$$

where $\boldsymbol{\beta}$ is a vector of parameters. This model effectively fits a Poisson point process with intensity of a loglinear form. This is a form of the logistic regression model popular in the field of statistics. The expression implies that the intensity varies exponentially as a function of the covariate. Parameter estimates of this model are derived via maximum likelihood.

The probability of occurrence $P(u)$ [2] is defined as the probability that at least one point falls inside a fine cell contained in the study area and is related to the intensity through the expression $\lambda(u) = \frac{P(u)}{1-P(u)}$ [2]. It can be shown via simple algebraic manipulations [2] that the probability of occurrence is expressed by

$$P(u) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}(u))}{1 + \exp(\boldsymbol{\beta}^T \mathbf{Z}(u))}$$

in some applications [2, 3, 4], and may be used to predict and identify areas of high or low point occurrence rates.

Nonparametric models do not assume a functional form and estimation involves finding ρ . Baddeley et al [3, 4] proposes an estimation for ρ that closely resembles probability density estimation from biased sample data and estimation of relative densities. Under regularity conditions ρ is proportional to the ratio of the density of covariate values at the points in the process and the density of covariate values at random locations. They define a spatial cumulative distribution function of a single, real-valued covariate $Z(u)$ on W given by

$$G(z) = \frac{1}{|W|} \int_W \mathbf{1}\{Z(u) \leq z\} du$$

where $|W|$ denotes the d -dimensional volume of the window W and $\mathbf{1}$ is an indicator function that evaluates to 1 if $Z(u) \leq z$ and 0 otherwise. G is assumed to have a derivative g . The non-normalised counterparts $G^*(z) = |W|G(z)$ and $g^*(z) = |W|g(z)$ are used for convenience. In practical applications, $G(z)$ is estimated by getting the value of the covariate evaluated at a fine grid of pixel locations and forming the cumulative distribution function

$$G(z) = \frac{\#\{\text{pixels } u : Z(u) \leq z\}}{\#\text{pixels}}.$$

The form of the kernel estimators of ρ for the ratio, reweighted and transformed case are

$$\hat{\rho}_R(z) = \frac{1}{g^*(z)} \sum_i K(Z(x_i) - z)$$

$$\hat{\rho}_W(z) = \sum_i \frac{1}{g^*(Z(x_i))} K(Z(x_i) - z)$$

and

$$\hat{\rho}_T(z) = \frac{1}{|W|} \sum_i K(G(Z(x_i)) - G(z))$$

respectively, where $K(\cdot)$ is a one-dimensional smoothing kernel, x_1, \dots, x_n are the observed data points, $Z(x_i)$ the value of the covariate at the observed data point, $|W|$ is the area of the observation window W and again G is the spatial cumulative distribution function (cdf) and $g^*(z) = |W|g(z)$ (i.e. $G'(z) = g(z)$). The derivative $G'(z)$ is obtained via an approximation of a differentiated smoothed estimate of G .

The next section is dedicated to a discussion of the effects of intensity estimation over empty areas and the consequence of using the Euclidean metric as a measure of proximity in calculating the kernel smoothed intensity estimate.

4.5 Intensity estimation on nonconvex domains

4.5.1 Estimation in empty space

In Section 3.2, we considered the selection of a nonconvex window domain by incorporating the known effects of a spatial covariate on the distribution of the pattern, particularly, the point occurrence. By filtering out these areas, we effectively assign zero intensity to points in these regions. In the case where a typical rectangular window is naively specified for an analysis, spurious estimation would occur in regions where points are non-observable. This may undermine the integrity of the results derived and inference that can be drawn from them. For example, a point pattern realized from a homogeneous Poisson process bounded in an irregular nonconvex domain and analysed in a rectangular or convex window, may lead to an incorrect conclusion of inhomogeneity when formal statistical testing [4] is done and the kernel smoothed intensity estimate is inspected. This would result from the fact that erroneous estimates of intensity in areas where no point occurrence can be observed would on average differ with intensity estimates in areas where points can be observed.

A useful diagnostic tool used to validate the fit of an intensity estimate is the relative intensity function. The relative intensity function $r(\mathbf{u})$ [4] measures agreement between the true intensity of a point process

X and a fitted intensity. The relative intensity is expressed by

$$r(\mathbf{u}) = \frac{\lambda(\mathbf{u})}{\lambda_0(\mathbf{u})}.$$

where $\lambda(\mathbf{u})$ and $\lambda_0(\mathbf{u})$ denote the true intensity and the fitted intensity respectively. Values of $r(\mathbf{u})$ close to 1 are indicative of the compliance of the fitted intensity to the true intensity value while values greater(less) than 1 indicate that the intensity model underestimates (overestimates) the true intensity. An estimate for the relative intensity in [4] is calculated via kernel smoothing by the expression

$$\hat{r}(\mathbf{u}) = \frac{1}{e(\mathbf{u})} \sum_i \frac{1}{\lambda_0(\mathbf{x}_i)} K(\mathbf{u} - \mathbf{x}_i)$$

where K is the smoothing kernel function and $e(\mathbf{u})$ is an edge correction factor for edge effects.

The following illustration makes use of the relative intensity function. The aim is to illustrate the effect of estimating intensity outside the true window domain. For this task we consider four irregular nonconvex planar domains, with varying areas shown in Figure 4.12, and four Poisson intensity functions. We consider both homogeneous and inhomogeneous cases for the intensity function. The intensity functions are only defined for their respective domains and points outside the domain attain an intensity value of 0. In the case of homogeneous intensity, $\lambda = 0.2$ and $\lambda = 30$ are used to simulate over the domains in Figure 4.12(a) and Figure 4.12(b) respectively. For the case of inhomogeneous intensity,

$$\lambda(\mathbf{x}) = -\frac{2}{5} \exp\left(-\frac{1}{2}((x_1 - 3)^2 + (x_2 - 2)^2)\right) + 30 \quad (4.9)$$

is used to simulate over the domain in Figure 4.12(c) and

$$\lambda(\mathbf{x}) = 90 \exp(-(x_1 - 5)^2 - (x_2 - 5)^2) + 3 \quad (4.10)$$

for the domain in Figure 4.12(d). The kernel smoothed intensity is estimated in each case over a rectangular window domain. Figure 4.13 depicts the point pattern plot of the simulated points in these rectangular windows. The corresponding kernel smoothed intensity estimates are shown in Figure 4.14. A Gaussian kernel function is used for the smoothing. In Figure 4.14 it can be observed that intensity estimates have been calculated for all the points in the rectangular domain including those extending beyond the original domains from which the points were simulated from. Plots of the relative intensity functions for each simulated pattern are shown in Figure 4.15. In Figure 4.15 we see large departures from the true intensity in the kernel estimates in the areas outside the domain. In all instances, the kernel smoothed intensity estimate overestimates the true intensity in these areas (i.e. the relative intensity is less than one and relatively close to zero). For regions inside the domain, the fitted kernel intensity models the true intensity fairly well.

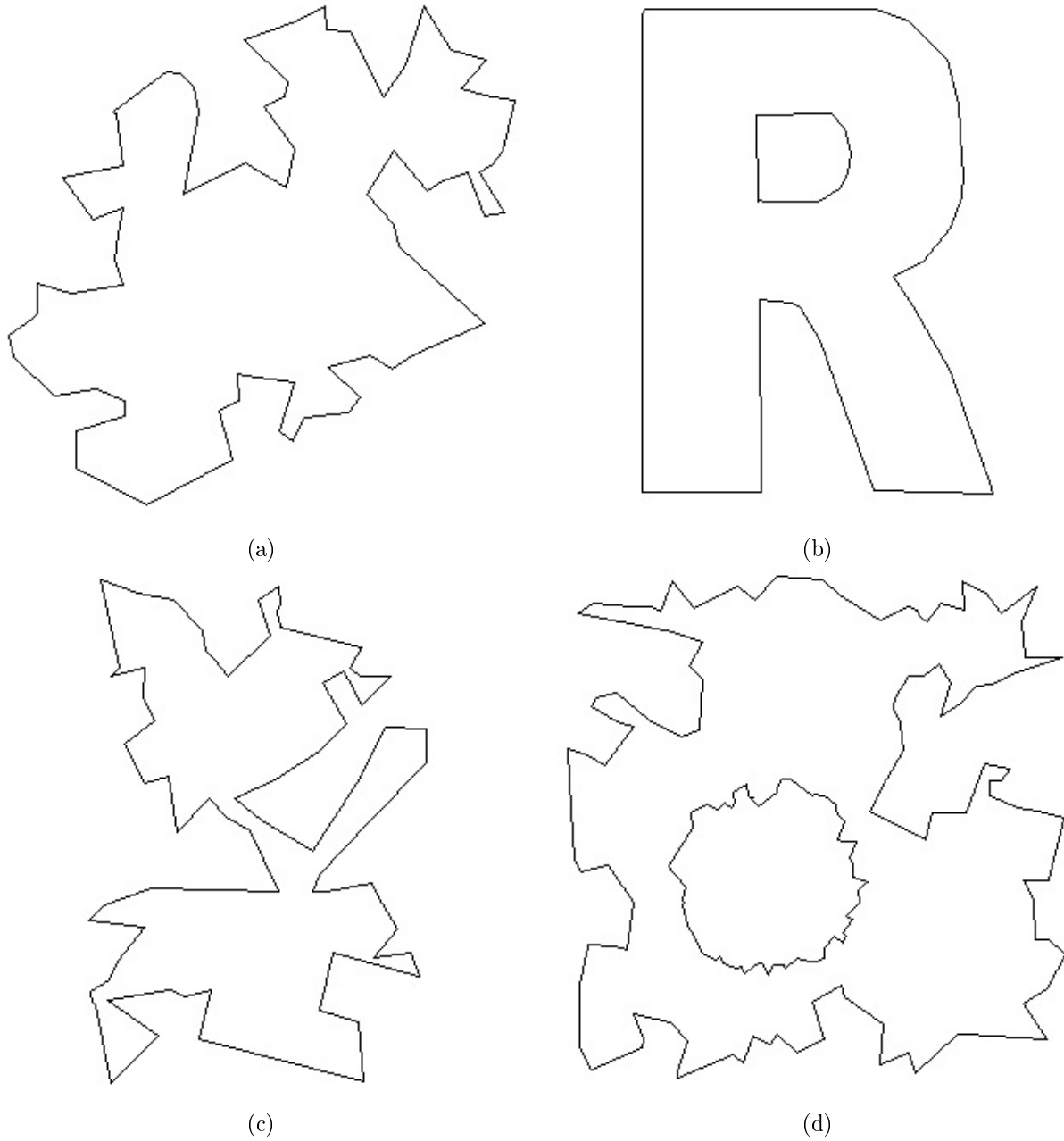


Figure 4.12: Irregular nonconvex window domains with varying areas, where (a), (b), (c), and (d) have areas of 508.6584, 3.697251, 2.15776, and 56.80648 square units respectively.

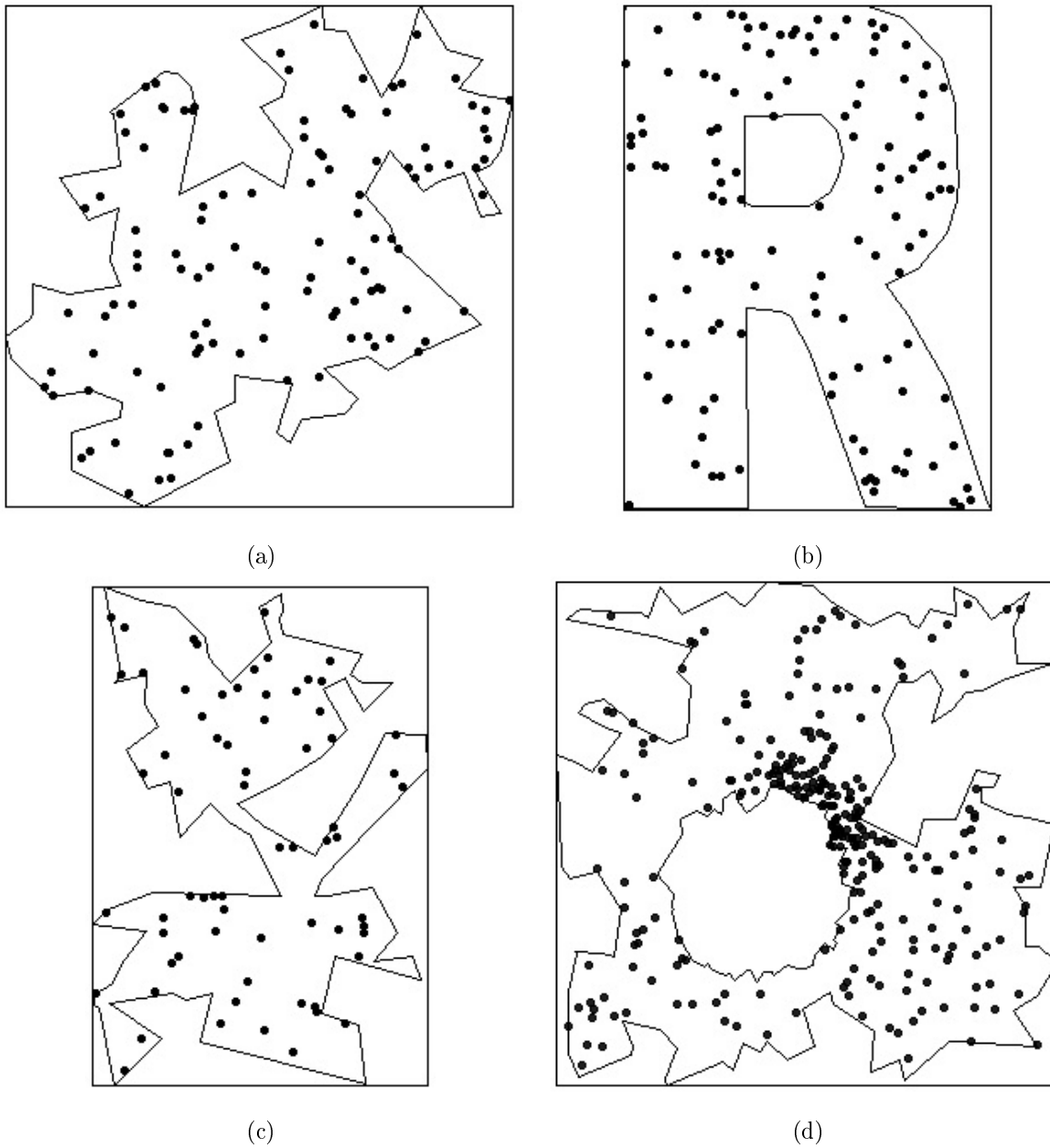


Figure 4.13: Point pattern plots of simulated points in a rectangular window, overlaid with the true window domain.

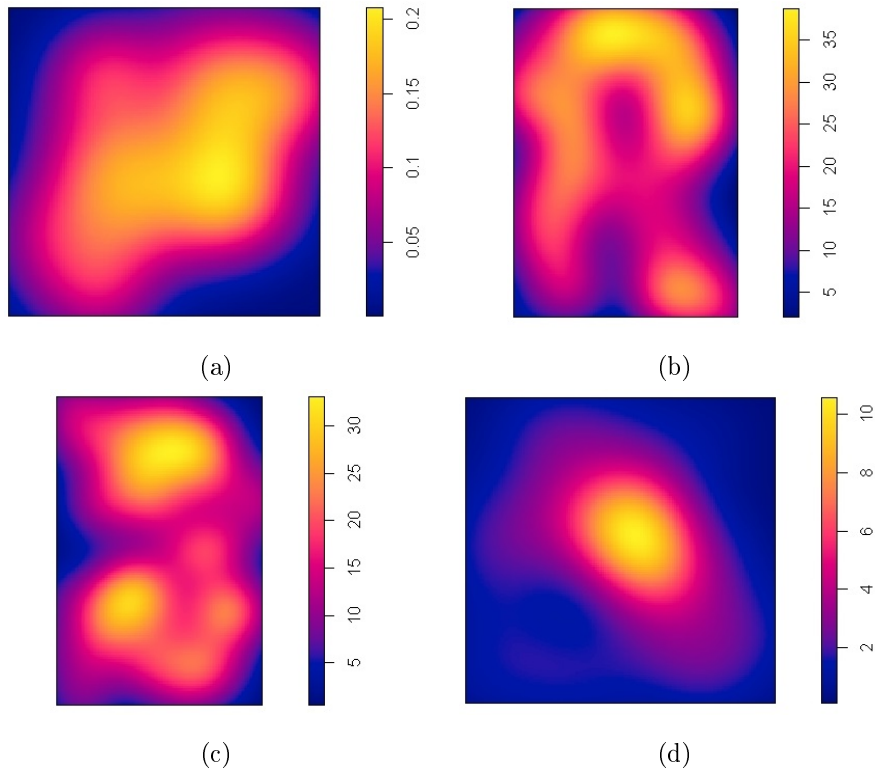


Figure 4.14: Kernel smoothed intensity plots of simulated point patterns.

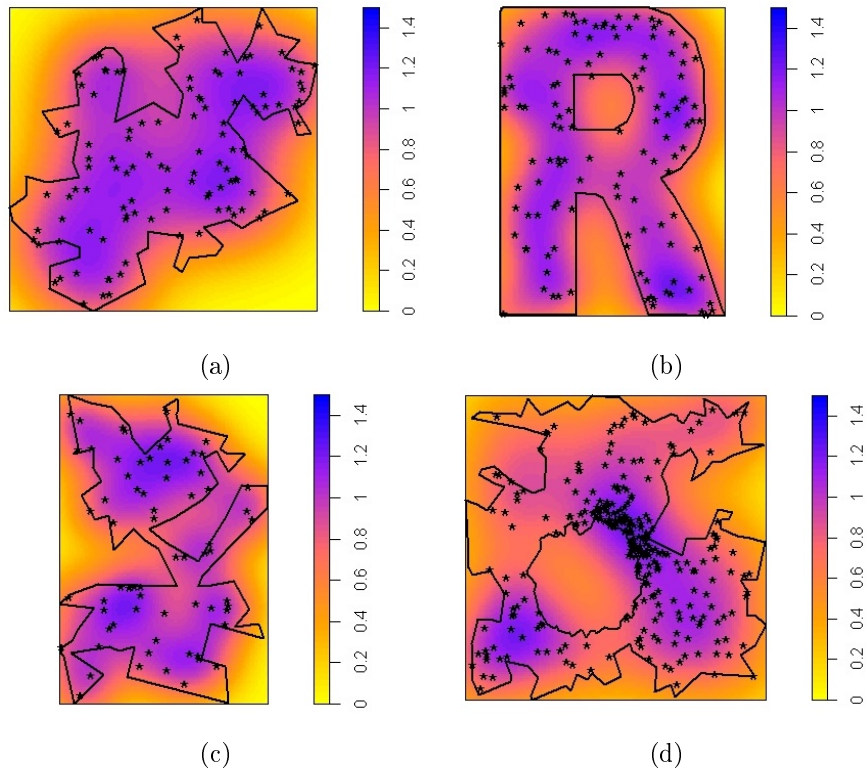


Figure 4.15: Relative intensity plots of simulated point patterns.

4.5.2 Euclidean measure of proximity

The kernel smoothed intensity estimate in Equation 4.3 apportions weights based on the Euclidean distance of a point to the point events observed in the point pattern. Utility of the Euclidean metric assumes that the point pattern is realized from a point process occurring in Euclidean space and that the smallest distance between points is formed along a path formed by the straight line segment representing this distance [55, 62]. The result is a kernel estimate that does not regard the irregular boundaries or holes in areas of the study domain. We illustrate this with the point pattern shown in Figure 4.16. The points are a realization of a homogeneous Poisson process with $\lambda = 25$ in an irregular, nonconvex, window domain shown in Figure 4.12(c). Consider a single point event observed on this window domain shown in Figure 4.17. If the Euclidean distance is used as a metric to characterize the distance between two points, the influence of the point event decays linearly as the distance from it increases. Consequently, the influence of the point event to other points is diffused radially from its center regulated by the linear distance. This means that the contribution to the kernel smoothed intensity estimate is higher even when the path between points are constrained by the boundary of the window domain. This can be observed in Figure 4.17.

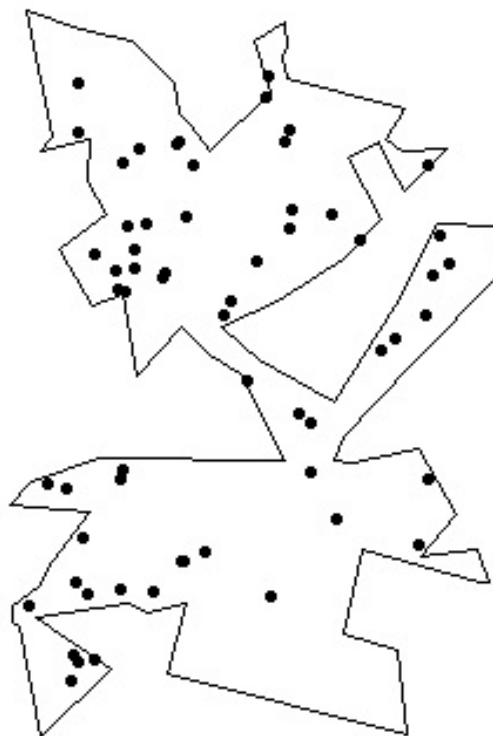


Figure 4.16: Point pattern simulated from a Poisson process with $\lambda = 25$ over an irregular, nonconvex window,

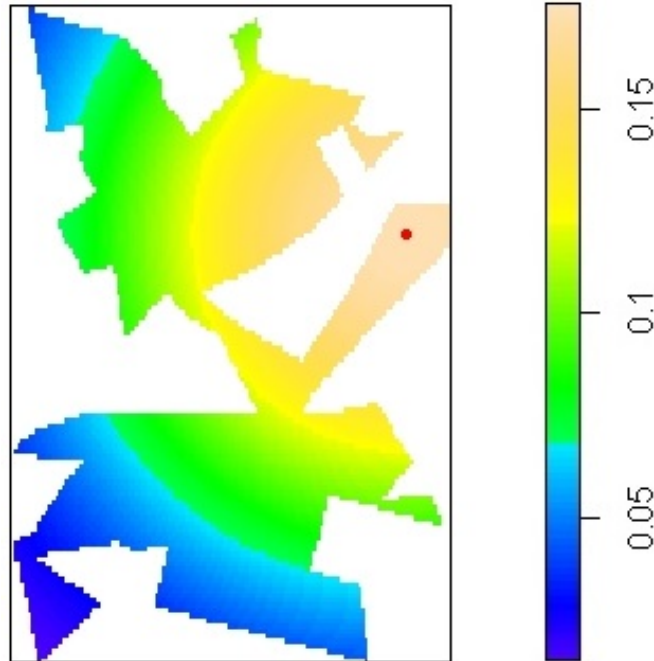


Figure 4.17: Illustration of the effect of using the Euclidean distance, over nonconvex domains, in estimating the kernel smoothed intensity function shown for a single point event from Figure 4.16. A Gaussian kernel is used.

Since traditional kernel density estimation applies kernel weights to Euclidean spaces, a modified kernel smoothed intensity estimator is required that allows for the distances between points to be represented by the true path between them. Several authors such as [10], [70], [114] and [117] have made use of an adapted version of the traditional kernel density estimator to allow for intensity estimation on linear networks, expressed as

$$\hat{\lambda}(\mathbf{x}) = \sum_{i=1}^n \frac{1}{h} K\left(\frac{d_i}{h}\right),$$

where d_i is the shortest path distance from the point \mathbf{x} to the point \mathbf{x}_i on the linear network, h is the bandwidth, and $K(\cdot)$ is a kernel function. In this context, smoothing along the shortest path distances on the network is used as an alternative to the Euclidean distance. This variant of the kernel density estimate will be used for the kernel smoothed intensity estimate on nonconvex domains. Here the Euclidean distance is changed out for the shortest path on the nonconvex domain. Figure 4.18 illustrates the effect of swapping out the Euclidean distance for the shortest path distance in the nonconvex window domain of the point pattern in Figure 4.16. Here we see that the weight of influence of the given point diffuses along a connected physical path.

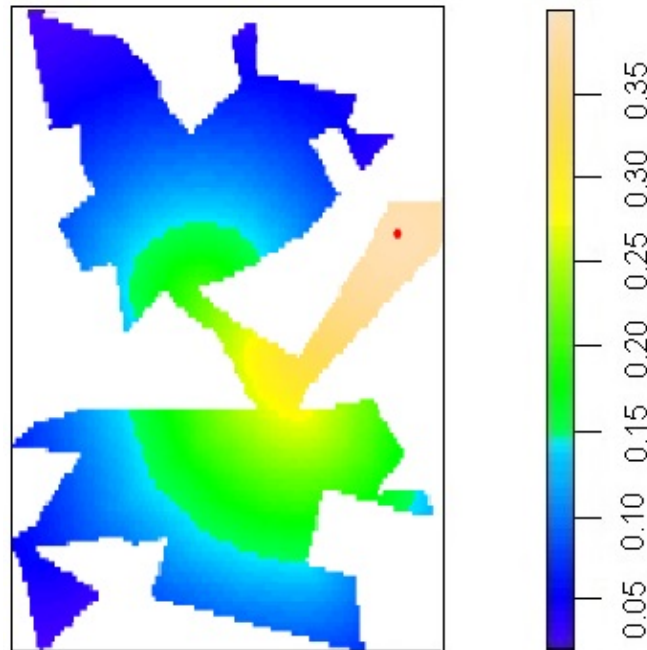


Figure 4.18: Illustration of the effect of using the shortest path distance, over nonconvex domains, in estimating the kernel smoothed intensity function shown for a single point event from Figure 4.16. A Gaussian kernel is used.

4.6 Concluding remarks

In this chapter we considered nonparametric methods for intensity estimation of point pattern data. Quadrat and kernel smoothing methods were reviewed in Sections 4.1 and 4.2 respectively. Particular focus was given to the kernel smoothed technique for intensity estimation. The kernel method for intensity estimation can also be extended to allow for covariate effects, a topic covered in Section 4.3.

For kernel intensity estimates, the bandwidth needs to be chosen carefully since it is directly responsible for the approximation accuracy. Large bandwidths mask the structure of the data and over-smooth the kernel estimate whilst a small bandwidth will under-smooth the data and produce estimates with large spikes.

It is important to select a representative window domain because this will directly affect the intensity estimate. If a window is chosen too large we have estimation occurring over areas for which data has not been observed and it has not been confirmed that a point can occur there. The result is spurious estimation of intensity in void areas where the point occurrence of an object or event can not happen.

The kernel smoothed intensity function apportions weights based on the Euclidean metric of distance.

This assumes that the point pattern is realized from a point process occurring in Euclidean space and that the smallest distance between points is formed along a path represented by this straight line segment. In this instance the kernel smoothed intensity function does not respect the boundaries or void areas in the domain. The solution here is to switch out the Euclidean metric for the distance in the intensity estimation, with that of the shortest path distance on the nonconvex domain, a concept which has been used for density estimation on linear networks.

Chapter 5

Application

In this chapter we apply the algorithm proposed in Section 3.2 to a point pattern where the points denote the locations of households in a rural setting. The chapter is organised as follows. Section 5.1 describes the data that will be utilised in applying the algorithm, namely the point pattern and spatial covariate data. The spatial covariate data is extracted from a Digital Elevation Model, the subject of which is detailed in Section 5.2. Section 5.3 is devoted to a discussion on mathematical morphology and morphological operators that will be used in the processing of the Digital Elevation Model. Section 5.4 presents the results of applying the proposed algorithm discussed in Section 3.2. Analysis in this chapter is done using R [76].

5.1 Data description

The data that will be used was collected in a census in the Serengeti District, Mara province, situated in Northern Tanzania^{1,2}. The census comprises of georeferenced data for 35947 households spread across 88 villages. The locations of the households are given as latitude and longitude in decimal degree coordinates. For the purposes of this mini-dissertation 5 villages, namely Iseresere, Nyamakobiti, Magatini, Majimoto and Hekwe with households that number 295, 412, 100, 336 and 235 respectively, will be utilized. The set of locations of each household in the villages will form the point pattern. The left pane of Figures 5.2 and 5.1 depict point pattern plots of the villages over a rectangular window.

As a spatial covariate, elevation data of Tanzania from a Digital Elevation Model (DEM) will be used (See Section 5.2 for definition of a DEM). The data was collected in the Shuttle Radar Topographic Mission

¹<https://www.gla.ac.uk/researchinstitutes/bahcm/staff/katiehampson/>,<http://www.katiehampson.com/>

²Ethics Reference number: NAS339/2019

(SRTM)^{2,3}, an international effort to generate a digital elevation model database of the Earth's topology. The SRTM data was sampled over a grid of 1 arc-second by 1 arc-second (approximately 30m by 30m), with linear vertical absolute height error of less than 16m.

Elevation is a property of the terrain of an area that influences the distribution of environmental phenomena and the nature of environmental processes [80]. Owing to this, it is useful as a covariate in describing terrain viable for (new) house locations. Terrain properties such as elevation at any point, slope and aspect are attributes of surface form that can be characterized using a DEM. Plots of the terrain elevation for the selected villages are depicted in the right pane of Figures 5.1 and 5.2.

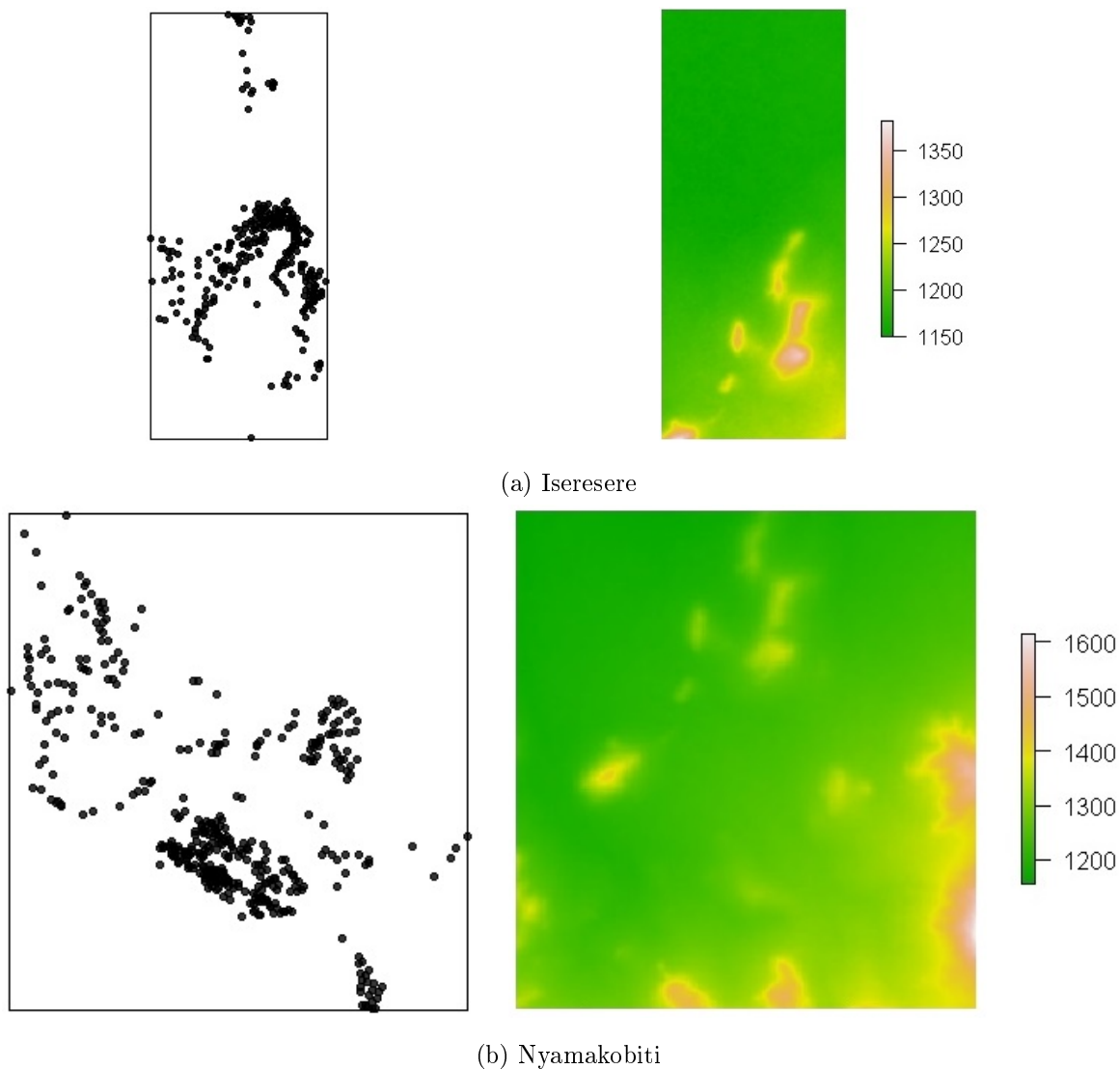


Figure 5.1: Point pattern plots (left) and terrain elevation (right) for villages Iseresere and Nyamakobiti in Tanzania's Mara province on a rectangular window.

³<https://www2.jpl.nasa.gov/srtm/>

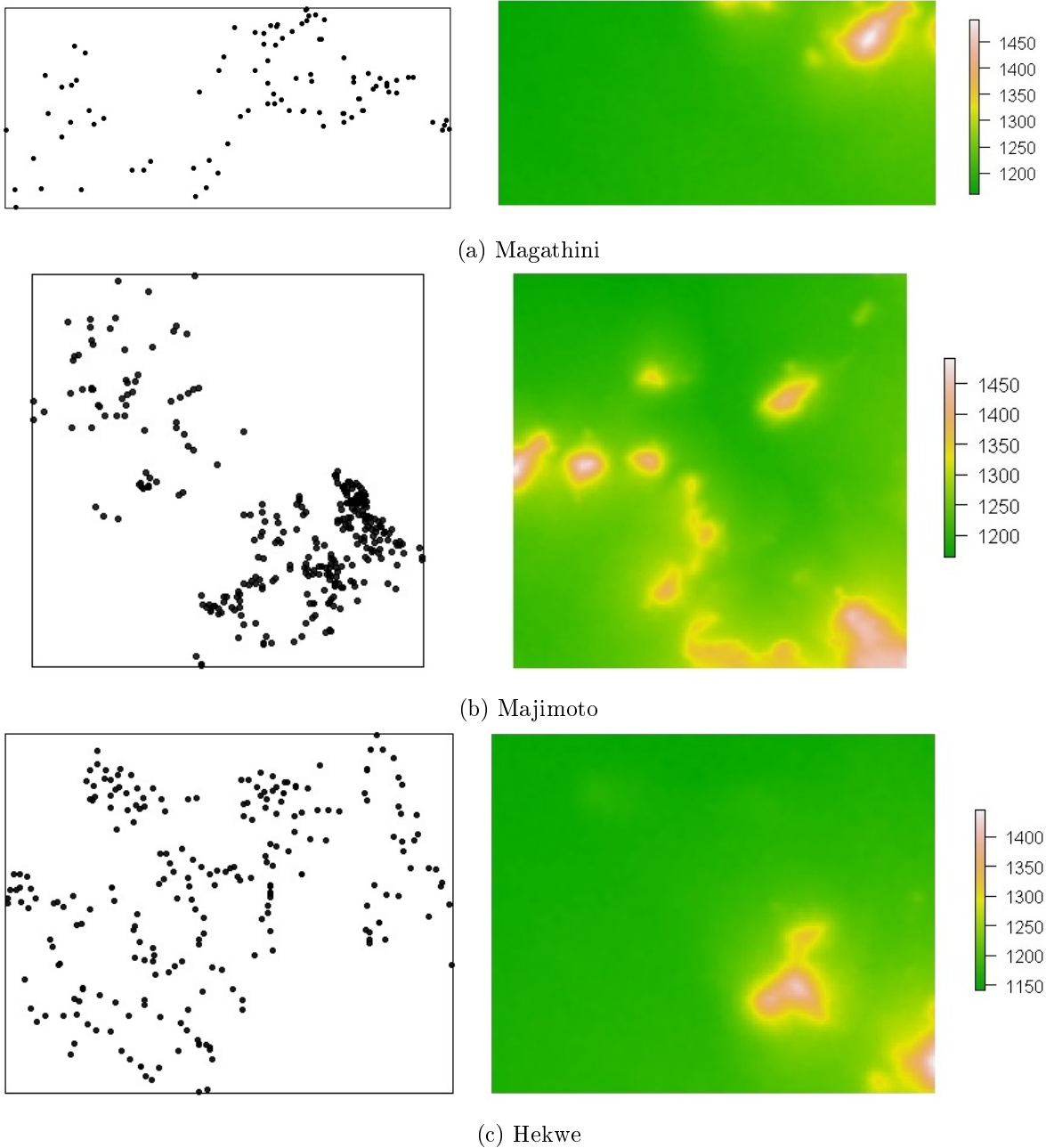


Figure 5.2: Point pattern plots (left) and terrain elevation (right) for villages Magathini, Majimoto and Hekwe in Tanzania's Mara province on a rectangular window.

5.2 Digital Elevation Model

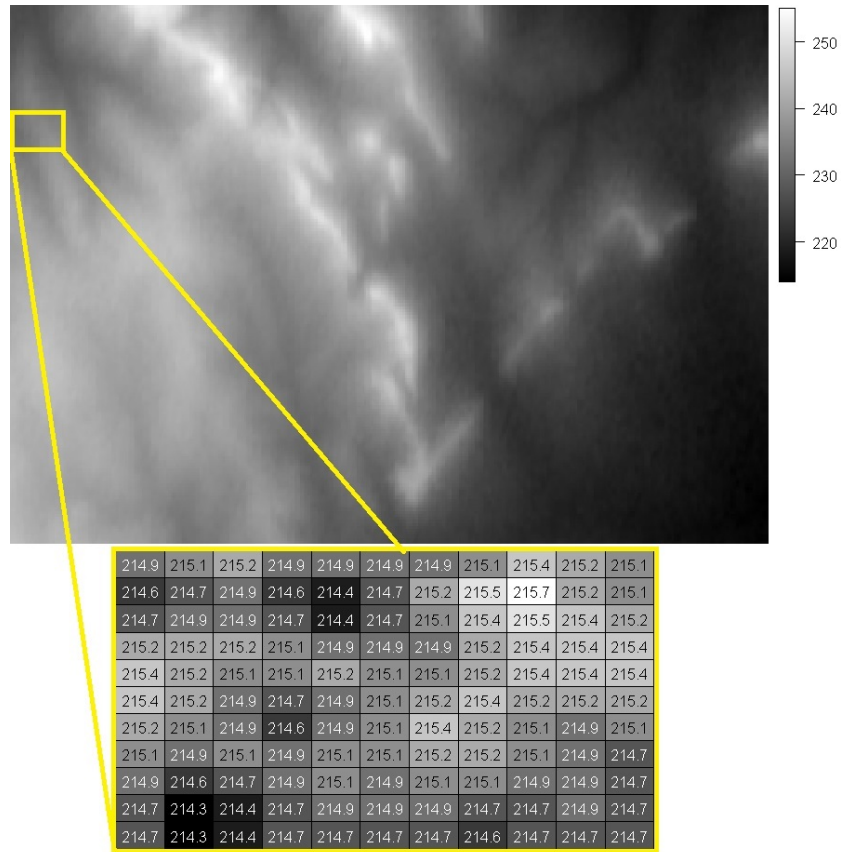
DEMs are utilized in a wide field of applications [1, 46, 47, 94, 101] that include hydrologic and geologic analyses, hazard monitoring, natural resources exploration and agricultural management. In hydrologic applications [40, 49, 57, 72, 102], the DEM gives insights about the relief of ground surface that aids the modelling of water, glaciers and ice movement, determining landslide probability, flood prone area mapping and estimating the volume of proposed reservoirs and wells. Applications in the field of geoscience [33, 63] involve the classification of the earth's surface at a global scale for large-scale modelling of geo-processes such as soil erosion, migration, and desertification.

A DEM represented as a raster grid, is a matrix of cells, with each cell containing a numeric value representing the elevation of the earth's surface above sea level in meters. The grid cells represent a square unit of area and each hold a measurement at sampled points, or estimates at unsampled points, of the elevation value referenced horizontally to a geographic coordinate system. A common method for generating a DEM is through the use of elevation data collected via remote sensing techniques. Methods for obtaining elevation data for DEM surfaces include,

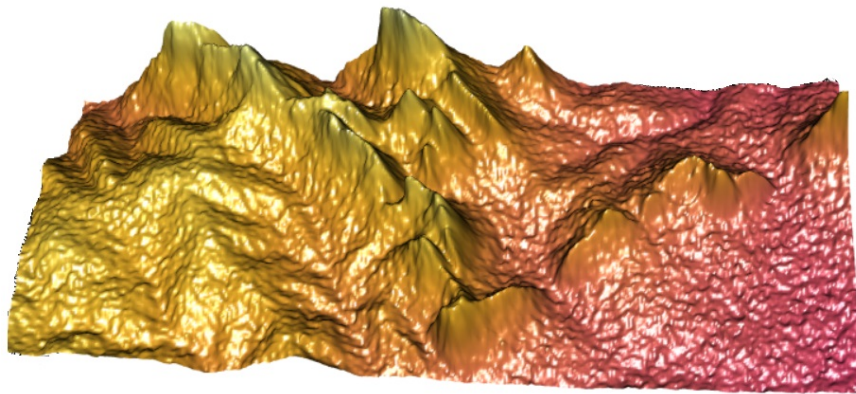
- satellite interferometry [8], a method that involves using two or more synthetic aperture radar (SAR) images to create maps of digital elevation,
- photogrammetry [113], making use of aerial photographs from different vantage points,
- LiDAR (Light Detection and Ranging) [30], measuring light reflected from the ground to obtain elevation of the Earth's surface, and
- land survey methods [60].

When a DEM is viewed on a map it appears as a surface layer symbolized by a colour ramp as shown in Figure 5.3(b). Elevation data may also be digitally represented on a triangular irregular network, and contours [11, 50, 53, 116]. Figure 5.3 depicts the raster grid representation and surface plot of elevation data from a DEM for Bokore village in Tanzania's Mara Province.

When considering DEMs as models for a surface form, a few things should be noted. A DEM is a set of discrete elevation measurements, used to model a usually undifferentiable continuous surface form. The accuracy with which the DEM models the true surface will depend on several factors [35, 52]. These are the surface roughness, the DEM resolution (i.e. the size of the raster cells), the elevation data collection method, and the interpolation algorithm used to obtain elevation values for unsampled points.



(a) Raster grid representation



(b) Surface plot

Figure 5.3: DEM elevation data plots for Bokore village in Tanzania’s Mara Province.

A global measure commonly used to assess the DEM accuracy is the Root Mean Squared Error (RMSE) statistic [35, 52, 108]. The RMSE is an indicator of vertical accuracy in a DEM. The RMSE is calculated by comparing the DEM with the most probable elevation points obtained at locations from field surveys, aerotriangulated test points, or contours from existing source maps. It quantifies the average deviation between more precise field based observations and the DEM value at the set of control locations. The

RMSE is given by the expression

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - y_j)^2}{N - 1}}$$

where y_i is an elevation from the DEM, y_j is the ‘true’ measured elevation of a test point and N is the number of sampled points. The treatment of DEM error is beyond the scope of this mini-dissertation and models used are assumed to give a fairly accurate representation of landform.

The DEM is a topographic surface with the grey level of a pixel being the elevation. Consequently, mathematical morphology is well suited for the processing of elevation data. Mathematical morphology is a branch of image analysis that deals with the extraction of image components [92, 96]. In the next section, we give a brief discussion on mathematical morphology and the mathematical morphological operators that are used to analyse a DEM in our application. The section will also detail an algorithm from [93], based on mathematical morphology, used to characterize physiographic features, the physical geography (i.e. mountains, basins, etc), of the earths’s terrain represented by the DEM.

5.3 Morphological segmentation of physiographic features from a DEM

5.3.1 Mathematical morphology

Mathematical morphology deals with the theory for analysing and processing geometric structures based on set theory, lattice theory, topology and random functions. It involves the mathematical theory of describing shapes using sets [95, 96, 98]. In mathematical morphology an image is defined by a set of coordinate vectors in Euclidean space. A greyscale image f is a mapping

$$f : \mathcal{D}_f \rightarrow T,$$

where \mathcal{D}_f , the definition domain, is the domain of f contained in a subspace \mathbb{Z}^2 of 2-dimensional Euclidean space, and T is the set of possible pixel values [95, 98]. In a greyscale image pixels with high values appear more white or bright and pixels with low values are black or dark. In the next section we define morphological operators that will be used in the algorithm [93] discussed in Section 5.3.3.

5.3.2 Mathematical morphological operators

In image analysis, morphological operators are used to extract desired structures from an image [95, 98]. This is done by probing the image with a set of known shapes termed structuring elements. Structuring elements have a defined origin, the centre pixel of the structuring element, that identifies the pixel in the

image being processed and defines the neighbourhood used in the processing of each pixel [98]. The shape of the structuring element is typically chosen based on prior knowledge about the geometry of the desired structure of the object to be extracted from the image [98]. Figure 5.4 shows an example of a 3×3 square structuring element around a target pixel and the neighbourhood defined by the structuring element.

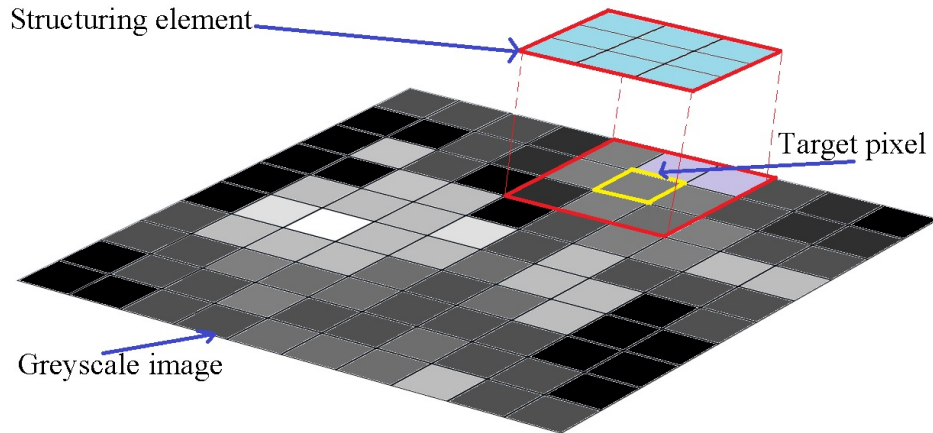


Figure 5.4: Illustration of an example of a 3×3 square structuring element around a target pixel and the neighbourhood defined by the structuring element

Erosion

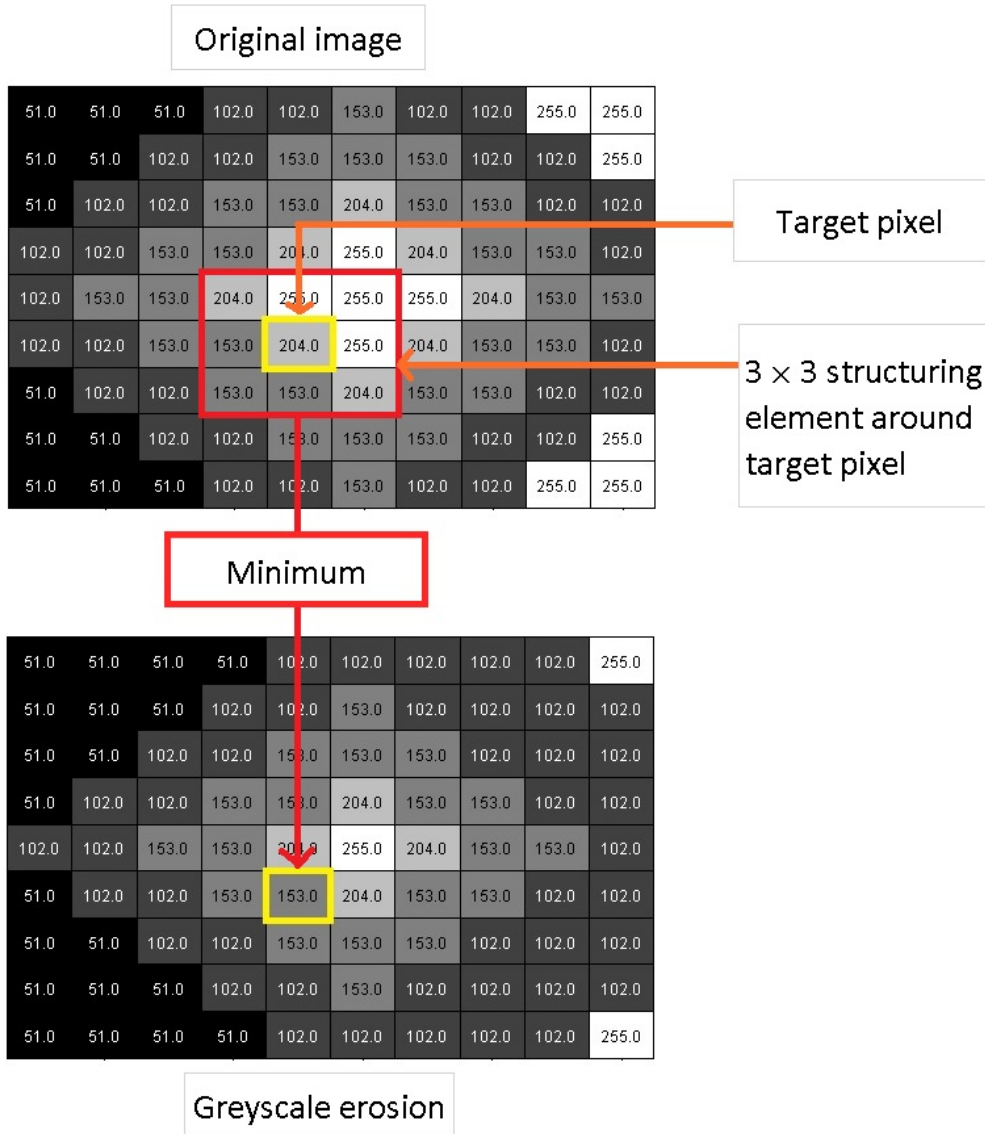
The erosion of a set X by a structuring element B is given by

$$X \ominus B = \{\mathbf{x} | B_{\mathbf{x}} \subseteq X\},$$

where \ominus is the erosion operator and $B_{\mathbf{x}}$ is the structuring element when its origin coincides with \mathbf{x} [98]. The eroded value of a greyscale image $f(\mathbf{x})$ at a given pixel \mathbf{x} is the minimum value of the image in the window defined by the structuring element B when its origin is at \mathbf{x} [98],

$$\min_{\mathbf{b} \in B} \{f(\mathbf{x} + \mathbf{b})\}.$$

Erosion sets the pixel values within the structuring element to the minimum value of the image. The result of applying erosion to a greyscale image is illustrated in Figure 5.5. The erosion is performed using a diamond structuring element. The figure shows the processing of a particular pixel in the input image. The structuring element defines the neighbourhood of pixels to be used. The pixel being processed attains the lowest value in this neighbourhood.

Figure 5.5: Illustration of an erosion of a greyscale image by a 3×3 square structuring element

Dilation

The dilation of a set X by a structuring element B is given by

$$X \oplus B = \{\mathbf{x} | B_{\mathbf{x}} \cap X \neq \emptyset\},$$

where \oplus is the dilation operator and $B_{\mathbf{x}}$ is the structuring element when its origin is at \mathbf{x} [98]. The dilated value of a greyscale image at a pixel \mathbf{x} is the maximum value of the image in the window defined by the structuring element B when its origin is at \mathbf{x} [98],

$$\max_{\mathbf{b} \in B} \{f(\mathbf{x} + \mathbf{b})\}.$$

Erosion and dilation are dual operators⁴. Erosion shrinks objects in an image whilst dilation expands them. Figure 5.6 shows the resultant image when a dilation is performed with a diamond structuring element. The pixel being processed attains the highest value in the neighbourhood defined by the structuring element.

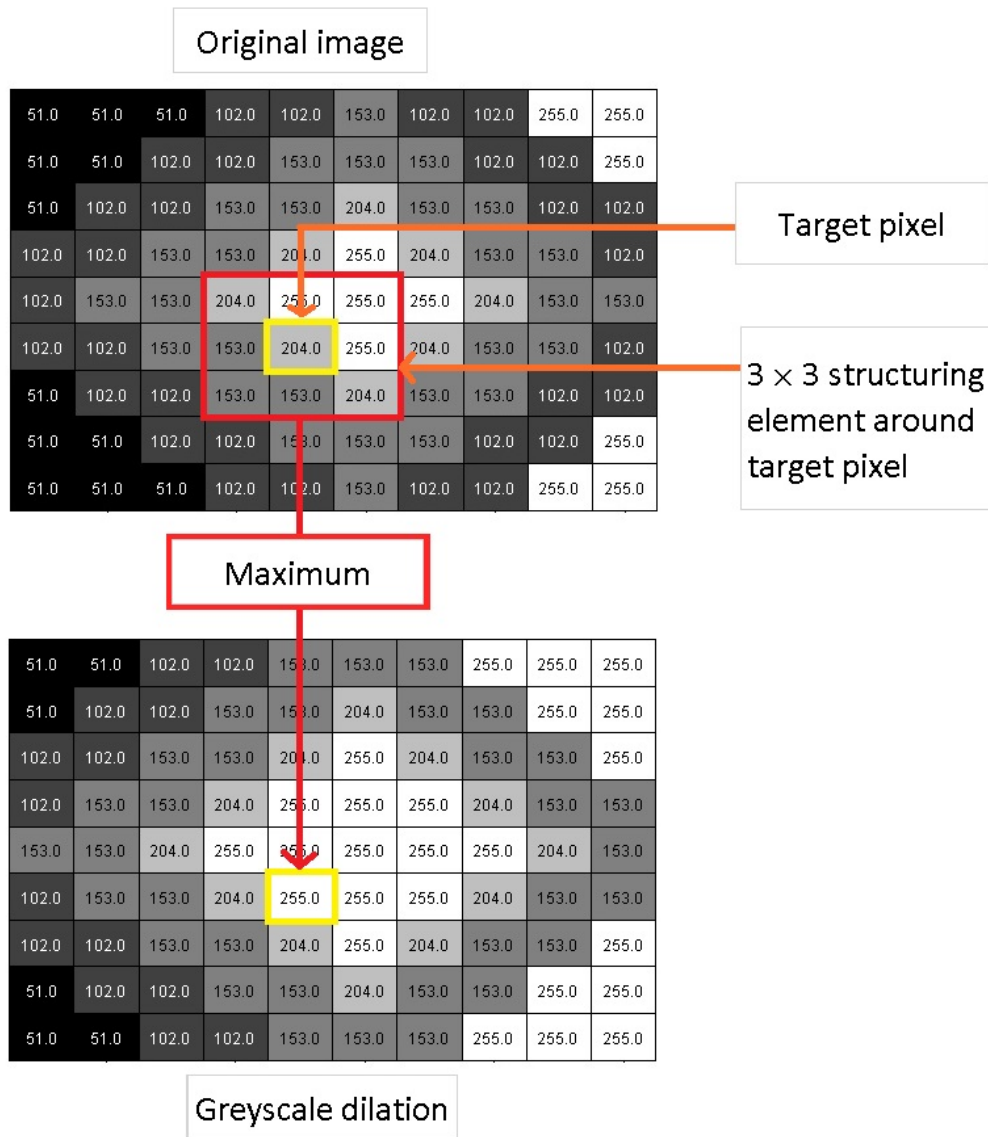


Figure 5.6: Illustration of a dilation of a greyscale image by a 3 × 3 square structuring element

⁴Erosion of an image is equivalent to the complement of the dilation of the complemented image with the same structuring element and vice versa. i.e. $X \ominus B = (X^C \oplus B)^C$ [98]. This characterizes the operators duality property.

Geodesic dilation

Geodesic dilation [98] involves the use of two images. One image is termed a marker and the other is termed as mask. Let f and g denote the marker and mask image respectively. Both images must have the same definition domain (i.e. $\mathcal{D}_f = \mathcal{D}_g \in \mathbb{Z}^2$) and it is also required that $f \leq g$, that is, the pixel values in the marker image should be less than or equal to the mask image pixel counterpart. The marker image f is first dilated by a basic structuring element B . The resultant image is then forced to remain below the mask image. The geodesic dilation, of size 1, of a marker f by a mask g , denoted by $\delta_g^{(1)}(f)$, is expressed as

$$\delta_g^{(1)}(f) = (f \oplus B) \wedge g$$

where the symbol \wedge denotes a pointwise minimum [98]. A geodesic dilation of size n , $\delta_g^{(n)}(f)$, of a marker image f by a mask image g is obtained by successively applying a geodesic dilation of size 1 n times [98]. Geodesic dilation on a greyscale image is illustrated in Figure 5.7. A 3×3 square structuring element is used in the dilation step.

5.3.3 Peak extraction in DEM using mathematical morphological operators

In the context of terrain features, a peak is the pointed top of a mountain or ridge and relates to the highest points of a mountain. In a DEM, the peaks are connected components that are enclosed in pixels of lower elevation. These areas are visually discernible as bright regions in the greyscale image. We now detail an algorithm presented in [93] for peak extraction from a DEM. This is based on mathematical morphological operators namely, ultimate erosion and geodesic dilation. In the algorithm, ultimate erosion involves the application of the erosion operator successively on an input image until components within the image disappear and reconstructing each eroded image using geodesic dilation on an erosion of smaller size. The reconstructed images are then subtracted from the corresponding eroded image to form the eroded sets; the components that appear in the eroded image, but not in the reconstructed image. When a component in the image disappears after an erosion, it is not reconstructed and will thus appear in the eroded set. The union of all eroded sets is termed the ultimate eroded set and contains all the peak pixels. The process of ultimate erosion in the algorithm is illustrated in Figure 5.8. In the figure, the images in the left panel are the result of applying successive erosions. The middle panel contains images reconstructed via the process of geodesic dilation where the eroded image is used as the mask, and an erosion of smaller size as the marker. The right-hand panel contains the eroded sets formed by subtracting the reconstructed image and corresponding eroded images and taking the unions of the eroded sets with the preceding iterations.

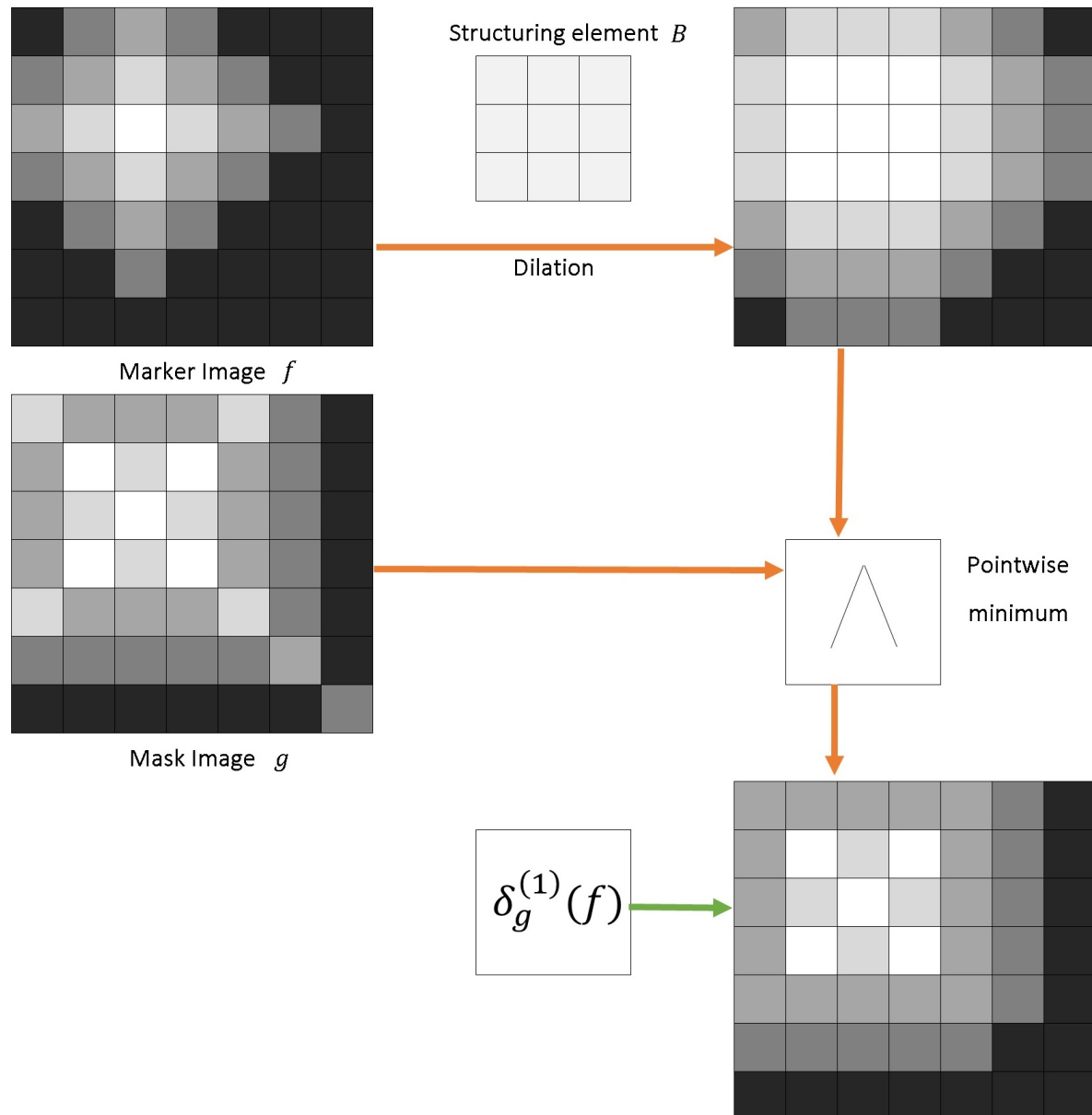


Figure 5.7: Illustration of a geodesic dilation of size 1 using a marker f and mask g , and a 3×3 square structuring element in the dilation step.

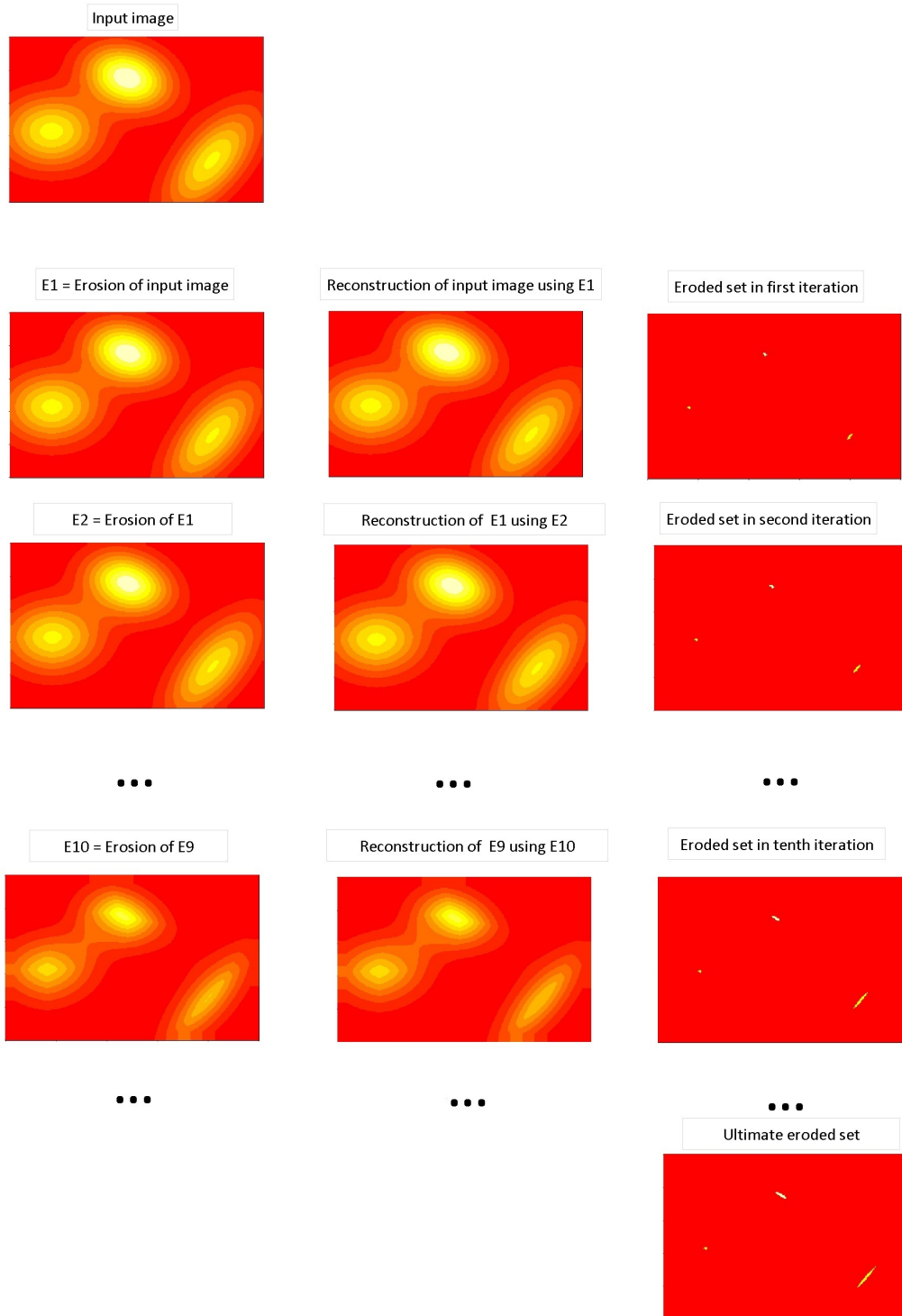


Figure 5.8: Illustration of ultimate erosion operation performed by the successive erosion (left) on the image until objects vanish and reconstructing each eroded image (middle) using an erosion of smaller size. The eroded sets (right) are formed by subtracting the reconstructed images with the corresponding eroded images.

5.3.4 Gaussian blurring

Gaussian blurring [32] is a smoothing operator that uses an isotropic Gaussian function as a kernel (or filter) to remove the extremities of an image. The Gaussian kernel [32] $G(x, y)$ used to smooth (i.e. blur) the image is of the form

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right),$$

where σ is the standard deviation of the Gaussian function and (x, y) is the pixel coordinate. Figure 5.9 depicts an example of the Gaussian kernel centered at zero and a standard deviation of one. The Gaussian blurring process involves moving over each pixel of the image and recalculating the pixel value based on a weighted average of the pixels that surround it. The pixel neighbourhood and values are regulated by the Gaussian kernel. Each new pixel value is determined by calculating a weighted average of the surrounding pixel values using the values from the Gaussian function as weights. Pixels close to the center pixel have higher weights than those further away. A discrete approximation of the Gaussian function is required before smoothing can be performed, since the image is stored as a collection of discrete pixels and not as a continuous function. The theoretical Gaussian distribution is non-zero everywhere which would require an infinitely large filter. When used practically however, it is effectively zero more than three standard deviations from the mean and thus can be truncated at this point. Figure 5.10 shows the result of applying Gaussian blurring, with $\sigma = 6$, to an image of two parrots⁵. The input image and blurred image are shown in Figures 5.10(a) and 5.10(b) respectively.

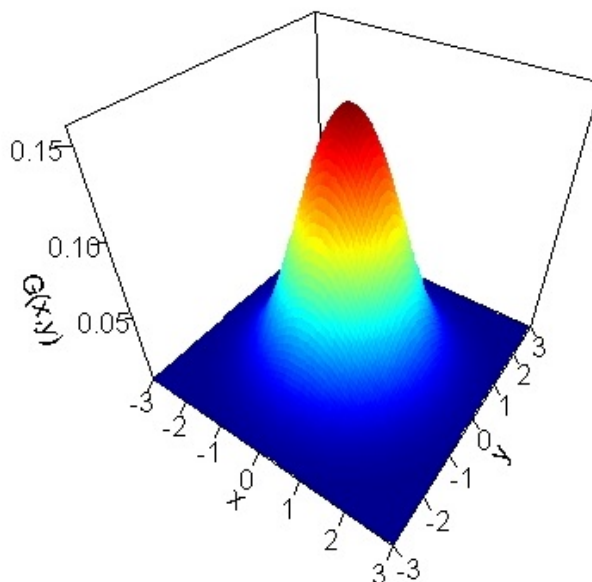


Figure 5.9: Gaussian kernel with mean $(0, 0)$ and $\sigma = 1$

⁵<https://www.bioconductor.org/packages/devel/bioc/vignettes/EImage/inst/doc/EImage-introduction.html>. Date accessed: 20 January 2020

2. A square moving window with side length equal to a multiple of the minimum Euclidean distance between the observed points is defined.
3. The moving window scans the neighbourhood of each point in the pattern and calculates the average slope.
4. A range for these values is used as the selection criterion to determine which quadrats should be included for the window W : a quadrat should be deleted from the window if the average slope is larger than the average slope of values considered viable as given by the range derived from the observed points.

The result for the five villages and the corresponding terrain plots are depicted in Figures 5.11 and 5.12.

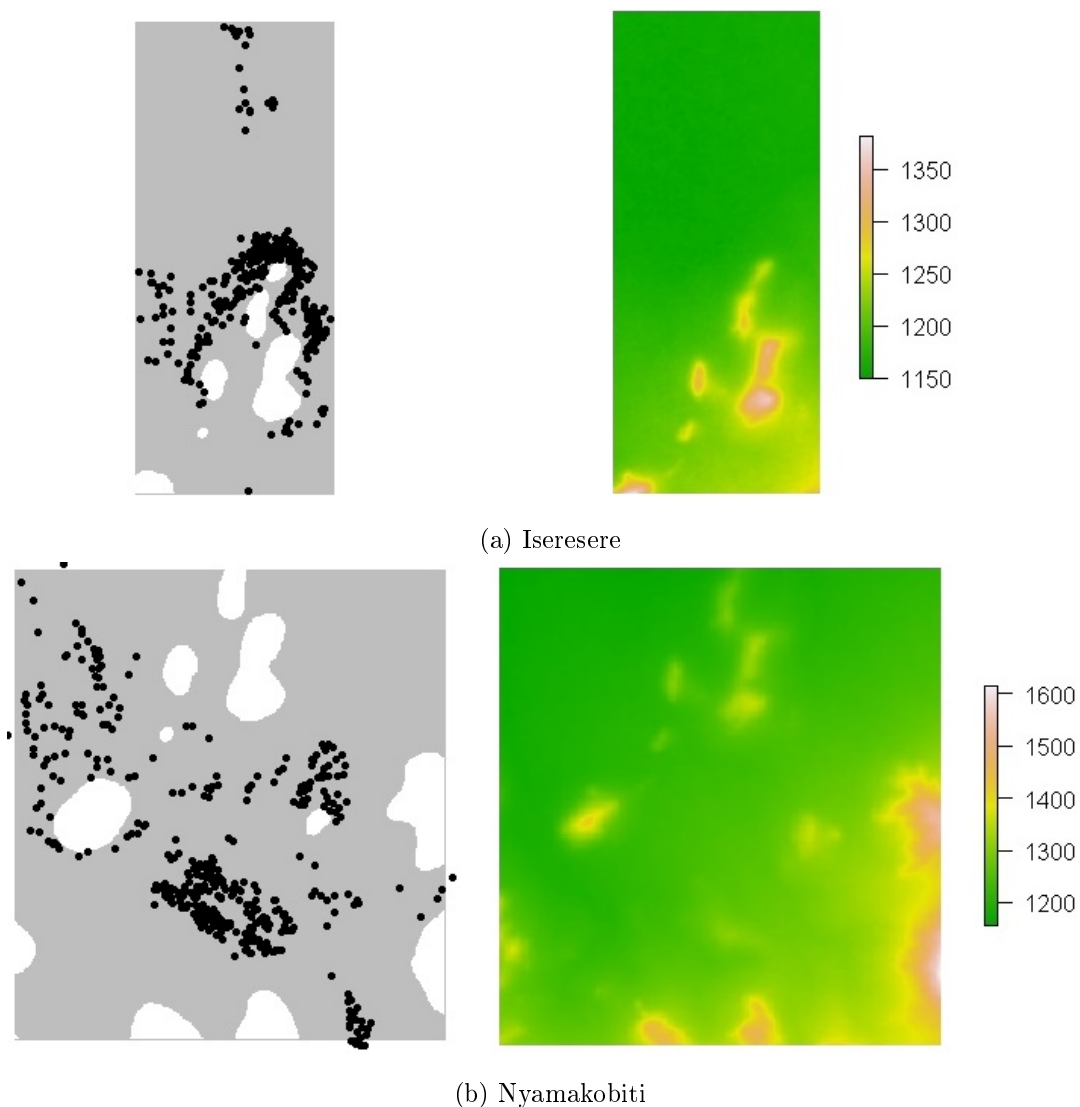
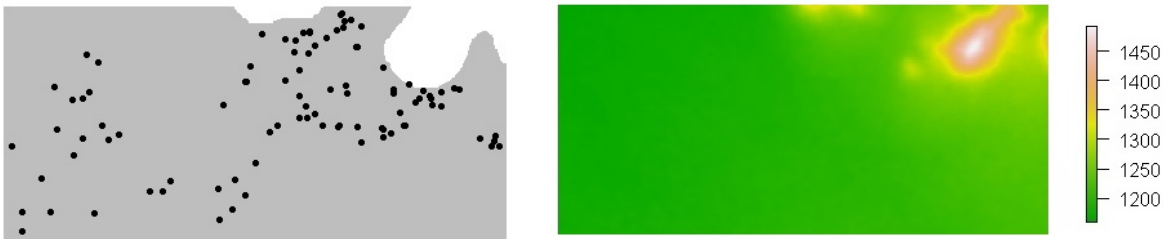
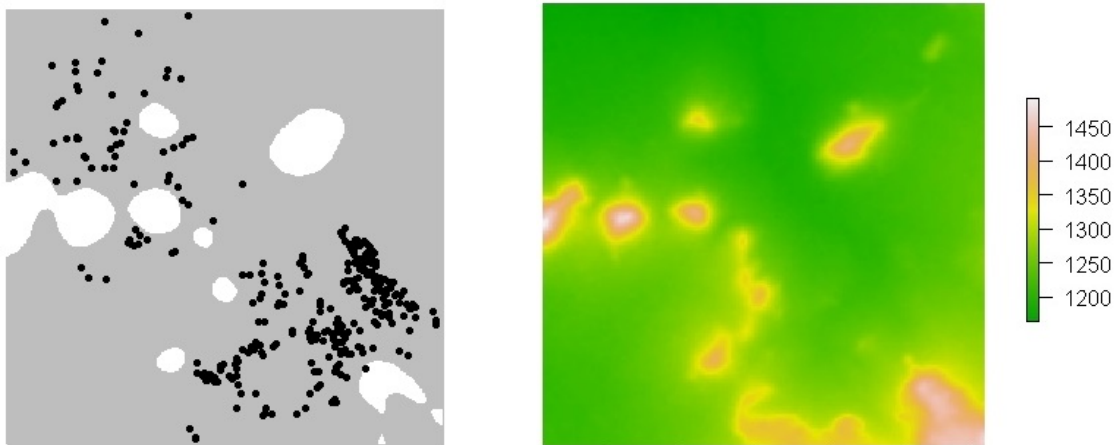


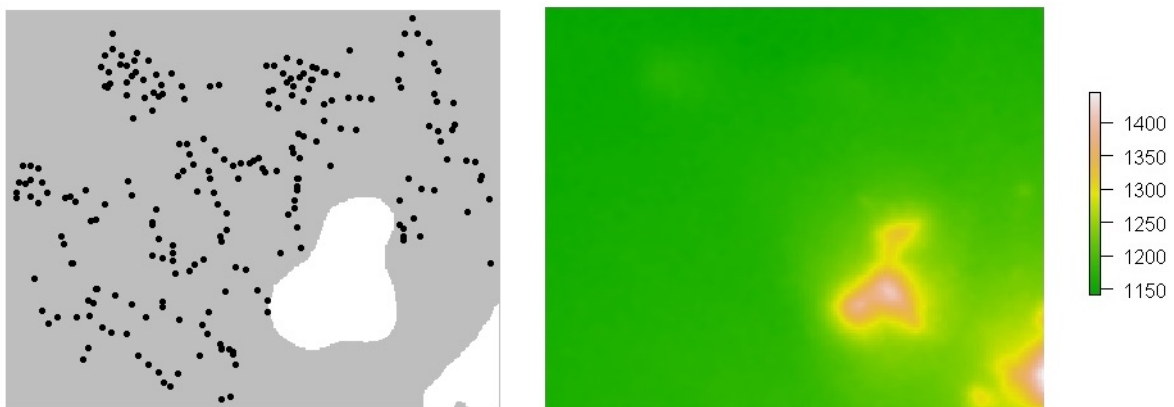
Figure 5.11: Point pattern plots (left) and corresponding terrain slope (right) for villages Iseresere and Nyamakobiti in Tanzania's Mara province, on nonconvex window constructed using covariate data



(a) Magathini



(b) Majimoto



(c) Hekwe

Figure 5.12: Point pattern plot (left) and corresponding terrain slope (right) for villages Magathini, Majimoto and Hekwe in Tanzania's Mara province, on nonconvex window constructed using covariate data

5.4.1 Landscape and terrain features

For the selection of sites for human settlement, in a rural setting, the location for shelter is typically guided by some attractive and restrictive trait of the natural landscape. These may include features such as the proximity to water and food, population density and usable land for agriculture and building. Figure 5.13⁷ depicts a Google Earth satellite image of Iseresere village in Mara province, Tanzania. The geographical locations, latitude and longitude, of the households are indicated on the image with red circles.

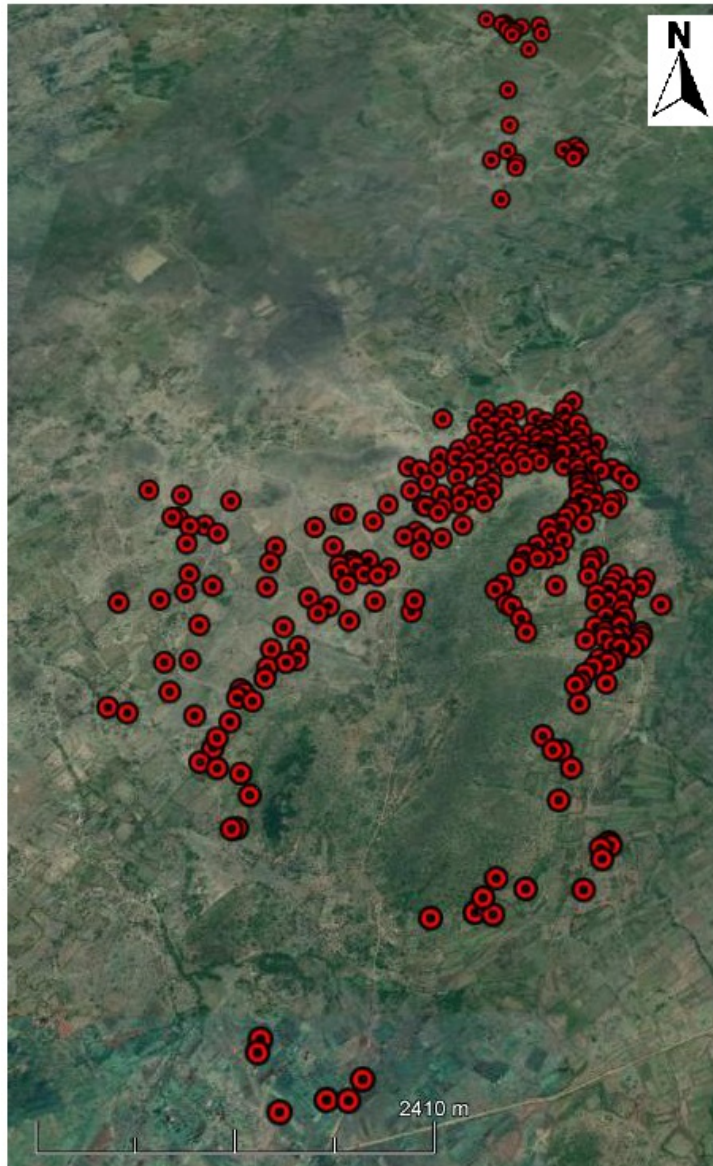


Figure 5.13: Google Earth satellite image of Iseresere village in Tanzania’s Mara province, with household locations indicated with red circles.

⁷Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°34'53.62”S, 34°23'58.54”E. Eye alt 12.01km. Maxar Technologies 2020, CNES/Airbus 2020. <http://www.earth.google.com>[March 20, 2020]

Figures 5.14⁸, 5.15⁹ and 5.16¹⁰ show enlarged images of different sites in the village. In the figures, terrain features such as hills (areas of high ground), ridges (sloping line of high ground) and flat plains (even landmass of relatively uniform elevation) are identifiable. Households are distributed at the edge of terrains with high relief. The occurrence of a household is only seen on flat plain areas and at the base of the mountainous regions. The households cluster on plains adjacent to scarps (i.e. steep slopes). On further inspection of the figures, we observe that the patterns of settlements are arranged based on the configuration of the site, the shape of agricultural fields and hills, streets and roads. The plan of the village is mostly adjusted to the relief features of the region, some along the edges of the hill slopes. It is also observed that homesteads are surrounded with patchy agricultural land.

5.5 Discussion

Notwithstanding the discussion on the appropriateness of the Euclidean metric as a measure of distance on nonconvex domains, the scope of this mini-dissertation is to illustrate the implementation of the proposed nonconvex window construction algorithm and thus the Euclidean distance is used as a measure of proximity for ease of implementation. The Euclidean metric is used in the algorithm in the definition of the moving window. If the shortest path distance was used to specify the size of the moving window, a window larger in size would be defined that may violate the requirement of having a single point contained in the area spanned by the window centered over observed point locations. Future research will implement non-Euclidean metrics in the proposed algorithm.

The relative intensity estimate, introduced in Section 4.5.1, is used as a diagnostic tool to evaluate the spatial domain constructed through the application of the proposed algorithm. The estimation is done on rectangular windows that extend beyond the constructed nonconvex spatial domains shown in Figures

⁸Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°31'49.52"S, 34°21'41.08"E. Eye alt 1.83km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°32'36.14"S, 34°22'12.67"E. Eye alt 2.98km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°36'48.25"S, 34°21'42.04"E. Eye alt 2.00km. CNES/Airbus 2020. <http://www.earth.google.com>[January 20, 2020]

⁹Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°35'57.22"S, 34°22'24.69"E. Eye alt 3.39km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°35'10.67"S, 34°22'39.62"E. Eye alt 2.94km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°34'31.19"S, 34°22'12.20"E. Eye alt 4.43km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020]

¹⁰Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°35'24.53"S, 34°21'40.57"E. Eye alt 5.49km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°33'30.73"S, 34°22'03.92"E. Eye alt 2.93km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020], Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. 1°35'35.08"S, 34°21'59.23"E. Eye alt 3.46km. Maxar Technonogies 2020. <http://www.earth.google.com>[January 20, 2020]

5.11 and 5.12. A Gaussian kernel function is used for the smoothing. The result is overlaid with the nonconvex spatial domains. The relative intensity plots for the villages are depicted in Figures 5.17 and 5.18. In the figures we observe large departures from the true intensity in the kernel estimates in the areas empty of data. In these areas, the relative intensity is less than one and relatively close to zero. Consequently, there is a tendency for the true intensity to be overestimated in these regions. This is observed for regions inside and regions outside the constructed nonconvex spatial domain and for all the villages. The fitted kernel intensity models the true intensity fairly well in regions where the observed occurrence of households are recorded as seen by the values close to one.

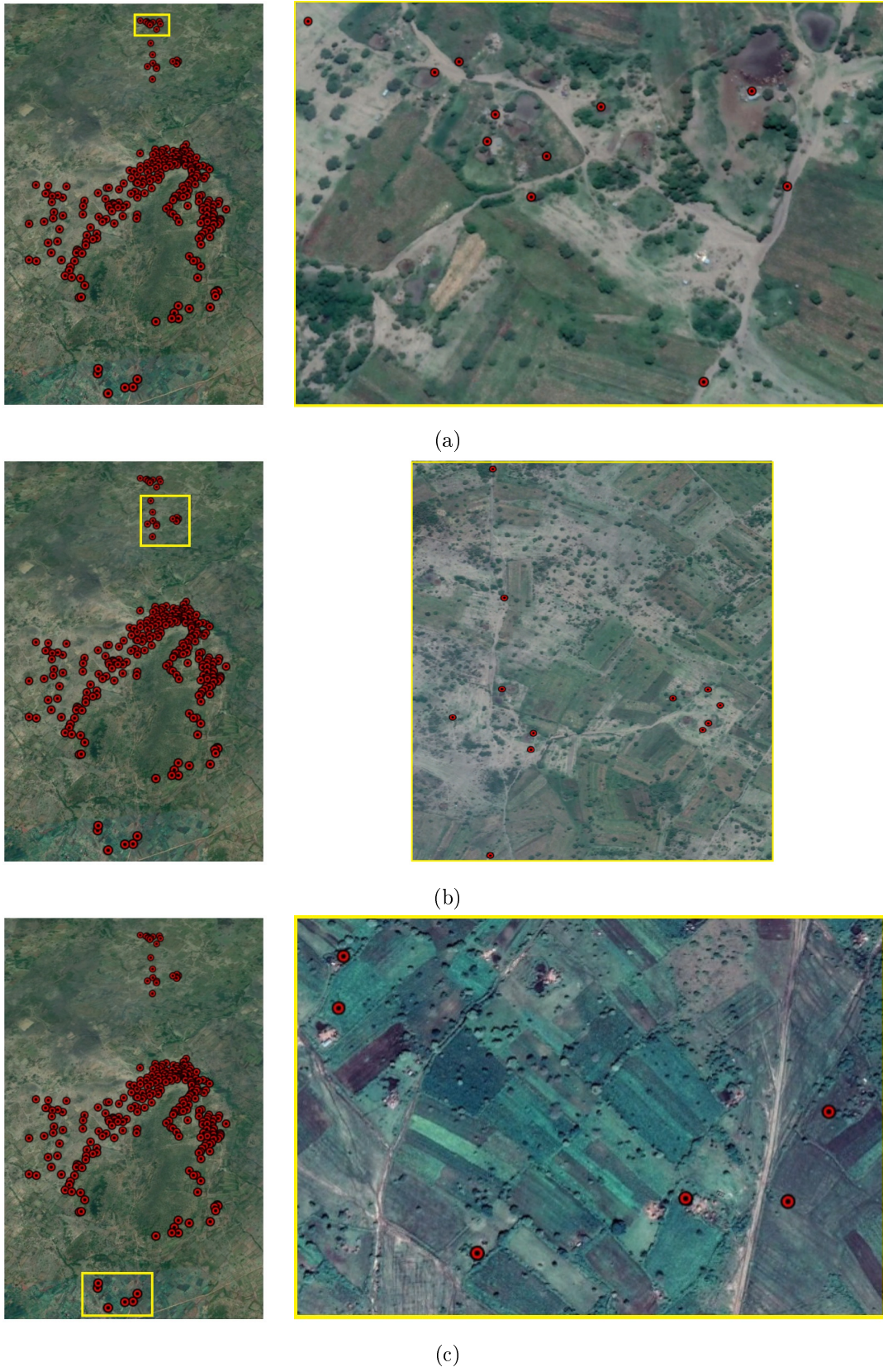


Figure 5.14: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province.

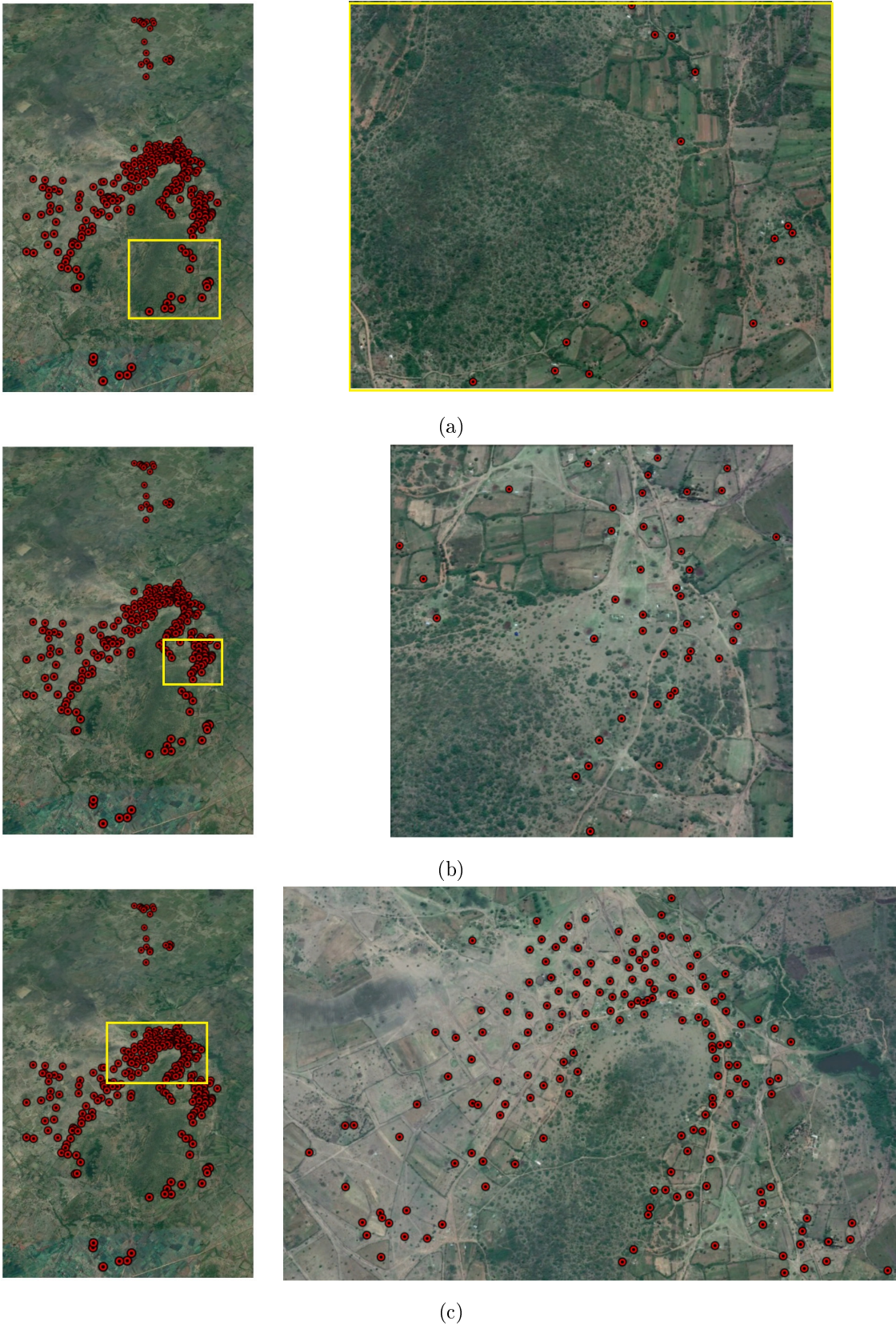
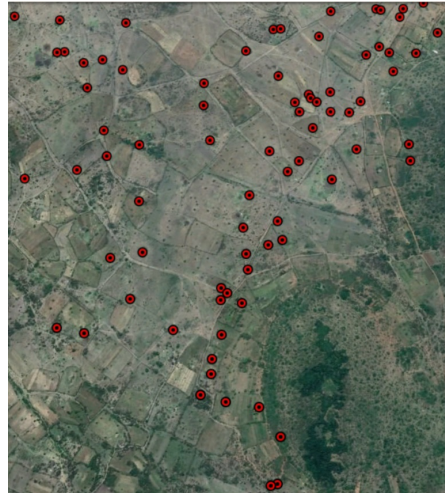
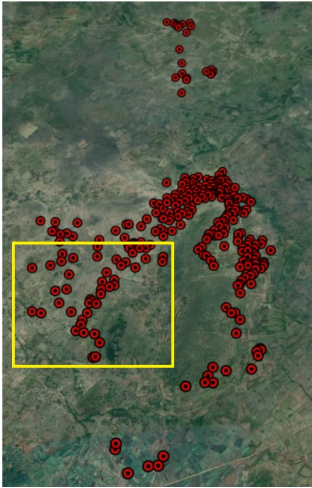
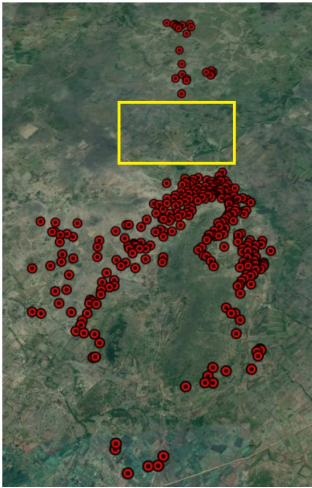


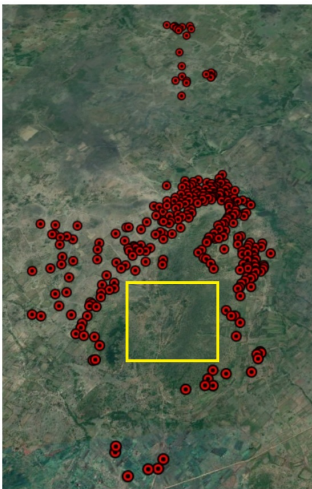
Figure 5.15: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).



(a)

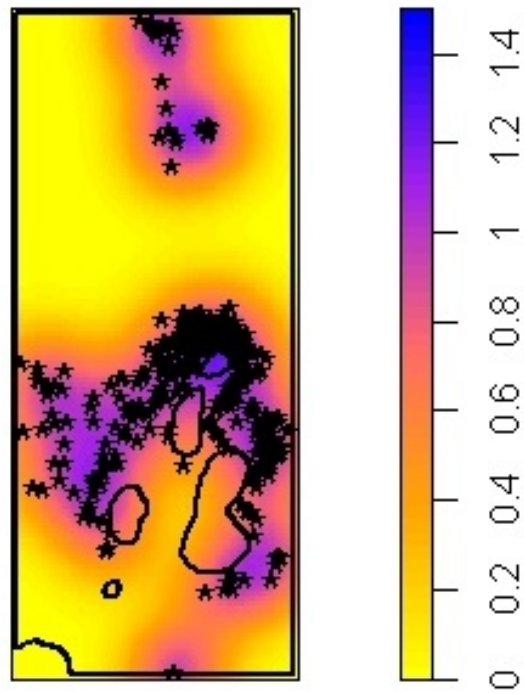


(b)

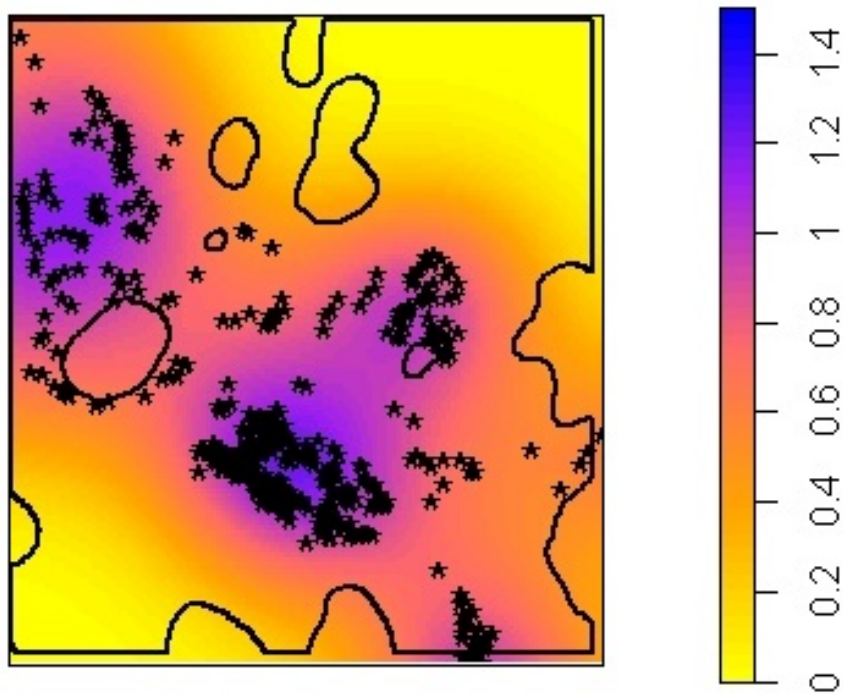


(c)

Figure 5.16: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).

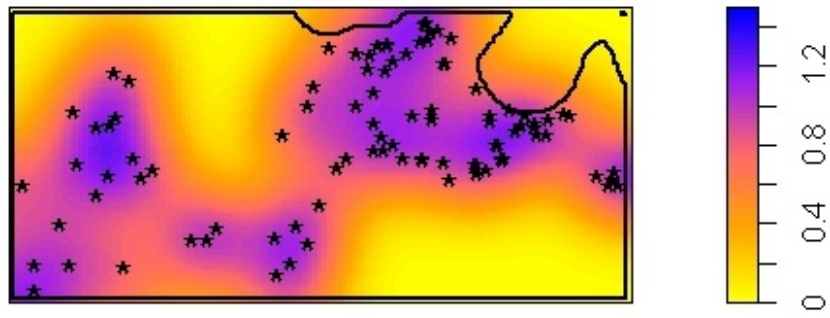


(a) Iseresere

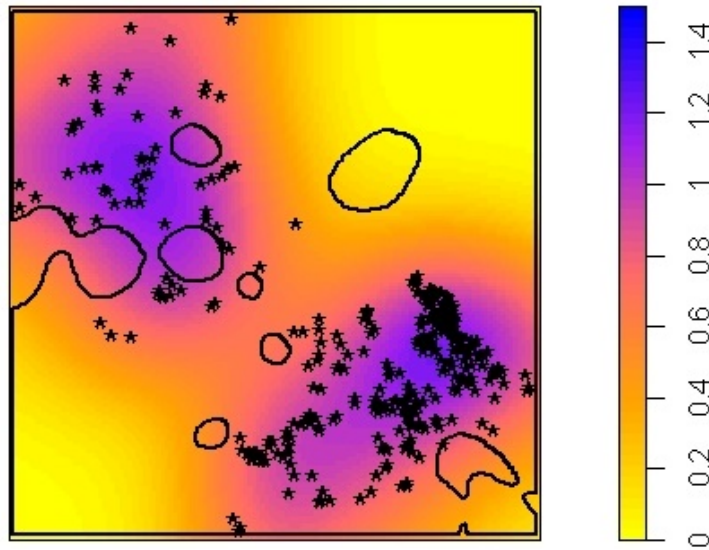


(b) Nyamakobiti

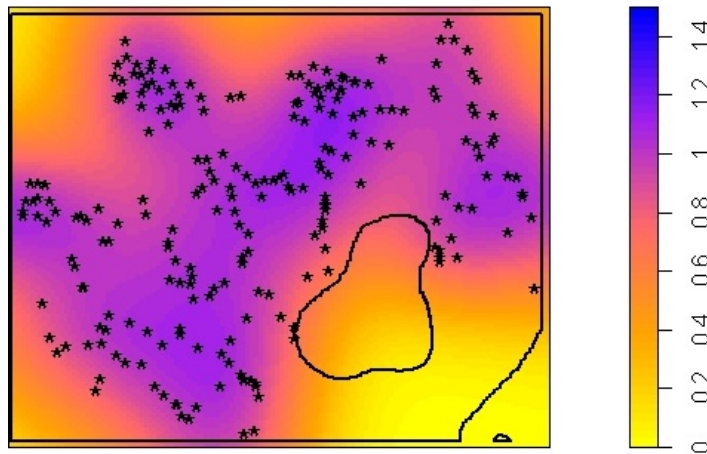
Figure 5.17: Relative intensity plots for villages Iseresere and Nyamakobiti estimated on a rectangular window and overlaid with the nonconvex constructed window.



(a) Magatini



(b) Majimoto



(c) Hekwe

Figure 5.18: Relative intensity plots for villages Magatini, Majimoto and Hekwe estimated on a rectangular window and overlaid with the nonconvex constructed window.

Chapter 6

Conclusion

In this mini-dissertation we considered spatial window selection techniques for point pattern data. The selection of a spatial window precedes analysis of a point pattern data set and is important since it is where estimation and prediction are done. A review of methods that involved the estimation of convex and nonconvex sets when points realized from the unknown set, and points outside the set, are observed was given. Some methods typically rely on the assumption of a homogeneous Poisson process and the assumption of a convex domain, which may not necessarily be true in practice. In real world applications, the distribution of points may be influenced by some underlying process, expressed as a covariate, resulting in more complex spatial windows.

We presented a new algorithm for selecting the spatial point pattern domain without the restriction of convexity. The algorithm works by using a moving window to search over a larger domain than that of the true window and, using a function or feature of the spatial covariate in regions at observed points in the pattern, construct a nonconvex window. Test for the dependence of the point pattern on the spatial covariate were discussed and should precede the application of this algorithm. Covariates that impact the distribution and abundance of the points of interest are desired and should be confirmed using these methods. Testing for global and local dependence of point patterns on covariates in parametric models can also be done using the method proposed in [68]. We applied the algorithm in the setting of rural villages in Tanzania's Mara province and used remote sensed data from a DEM as a covariate. The extraction of peaks through the use of mathematical morphological operators was included as a preprocessing step. The algorithm did well in detecting and filtering out areas of high relief and steep slopes, observed characteristics that were seen to make the occurrence of a household improbable.

We discussed the importance of selecting representative spatial window domains and the effect that this has on the kernel smoothed intensity estimate, a spatial measure used to investigate the first-order properties

of a point pattern. The selected spatial window directly affects the intensity estimate. If a window is chosen too large we have estimation occurring over areas for which data has not been observed and it has not been confirmed that a point can occur there. The result is spurious estimation of intensity in void areas where the point occurrence of an object or event can not happen. The kernel smoothed intensity estimate can also be extended to allow for covariate effects.

The kernel smoothed intensity function weights the influence of a point based on Euclidean distance. This assumes that the point pattern is realized from a point process occurring in Euclidean space and that the smallest distance between points is formed along a path represented by this straight line segment. Consequently this would mean that the kernel smoothed intensity function does not respect the boundaries or void areas in the domain. The solution here is to switch out the Euclidean metric for the distance in the intensity estimation, with that of the shortest path distance on the nonconvex domain, a concept which has been used for density estimation on linear networks.

In future work the algorithm could be extended to allow for an ensemble of spatial covariate effects. The possibility of weighting covariates based on their influence on points in the pattern could also be explored. One could also investigate ways to refine the selection criterion (function of the covariate), used as a filter to remove regions, and make the identification of this feature more automatic. The definition of the moving window using a non-Euclidean metric also warrants further implementation.

In this research we have contributed the following:

- reviewed methods for convex and nonconvex spatial window estimation,
- presented a new algorithm for nonconvex window construction of a point pattern using spatial covariates,
- illustrated the effects of the choice of window on the kernel smoothed intensity estimate, and
- illustrated the effects of using the Euclidean metric as a measure of distance when computing the kernel smoothed intensity estimate.

Bibliography

- [1] Sameh K Abd-Elmabod, Antonio Jordán, Luuk Fleskens, Jonathan D Phillips, Miriam Muñoz-Rojas, Martine van der Ploeg, María Anaya-Romero, Soad El-Ashry, and Diego de la Rosa. Modeling agricultural suitability along soil transects under current conditions and improved scenario of soil factors. In *Soil Mapping and Process Modeling for Sustainable Land Use Management*, pages 193–219. Elsevier, 2017.
- [2] Adrian Baddeley, Mark Berman, Nicholas I Fisher, Andrew Hardegen, Robin K Milne, Dominic Schuhmacher, Rohan Shah, and Rolf Turner. Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, 4:1151–1201, 2010.
- [3] Adrian Baddeley, Ya-Mei Chang, Yong Song, and Rolf Turner. Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and its Interface*, 5(2):221–236, 2012.
- [4] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, 2015.
- [5] Trevor C Bailey and Anthony C Gatrell. *Interactive Spatial Data Analysis*, volume 413. Longman Scientific & Technical Essex, 1995.
- [6] Amparo Baillo, Antonio Cuevas, and Ana Justel. Set estimation and nonparametric detection. *Canadian Journal of Statistics*, 28(4):765–782, 2000.
- [7] Lee J Bain and Max Engelhardt. *Introduction to Probability and Mathematical Statistics*. Brooks/Cole, 1987.
- [8] Richard Bamler and Philipp Hartl. Synthetic aperture radar interferometry. *Inverse Problems*, 14(4):R1, 1998.
- [9] Rajendra Bhatia. *Positive Definite Matrices*, volume 24. Princeton University Press, 2009.
- [10] Giuseppe Borruso. Network density and the delimitation of urban areas. *Transactions in GIS*, 7(2):177–191, 2003.

- [11] James R Carter. Digital representations of topographic surfaces. *Photogrammetry Engineering and Remote Sensing*, 54(11):1577–1580, 1988.
- [12] Achmad Choiruddin, Jean-François Coeurjolly, and Frédérique Letué. Spatial point processes intensity estimation with a diverging number of covariates. *arXiv preprint arXiv:1712.09562*, 2017.
- [13] Philip J Clark and Francis C Evans. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35(4):445–453, 1954.
- [14] Jean-François Coeurjolly and Jesper Møller. Variational approach for spatial point process intensity estimation. *Bernoulli*, 20(3):1097–1125, 2014.
- [15] Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.
- [16] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.
- [17] Ottmar Cronie and Maria Nicolette Margaretha Van Lieshout. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105(2):455–462, 2018.
- [18] Antonio Cuevas. Set estimation: Another bridge between statistics and geometry. *Boletín de Estadística e Investigación Operativa*, 25(2):71–85, 2009.
- [19] Antonio Cuevas, Ricardo Fraiman, and Alberto Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *The Annals of Statistics*, 35(3):1031–1051, 2007.
- [20] PJ Dare and JT Barry. Population size, density and regularity in nest spacing of Buzzards Buteo Buteo in two upland regions of North Wales. *Bird Study*, 37(1):23–29, 1990.
- [21] Jon Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Lulu, 2010.
- [22] Luc Devroye and Gary L Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.
- [23] Peter Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.
- [24] Peter J Diggle. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics*, pages 87–101, 1979.
- [25] Peter J Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, 2013.
- [26] Tarn Duong and Martin Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, 15(1):17–30, 2003.

- [27] Tarn Duong and Martin L Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32(3):485–506, 2005.
- [28] Rex A Dwyer. On the convex hull of random points in a polytope. *Journal of Applied Probability*, 25(4):688–699, 1988.
- [29] Bradley Efron. The convex hull of a random set of points. *Biometrika*, 52(3-4):331–343, 1965.
- [30] Juan Carlos Fernandez Diaz, William E Carter, Ramesh L Shrestha, and Craig L Glennie. LiDAR remote sensing. *Handbook of Satellite Applications*, pages 757–808, 2013.
- [31] Manfred M Fischer. *Spatial Analysis in Geography*. Springer, 2006.
- [32] Robert Fisher, Simon Perkins, Ashley Walker, and Erik Wolfart. Hypermedia image processing reference. *Department of Artificial Intelligence, University of Edinburg*, 1996.
- [33] Igor Florinsky. *Digital Terrain Analysis in Soil Science and Geology*. Academic Press, 2016.
- [34] Zhongliang Fu and Yuefeng Lu. An efficient algorithm for the convex hull of planner scattered point set. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 39:B2, 2012.
- [35] Jay Gao. Resolution and accuracy of terrain representation by grid dems at a micro-scale. *International Journal of Geographical Information Science*, 11(2):199–212, 1997.
- [36] Anthony C Gatrell, Trevor C Bailey, Peter J Diggle, and Barry S Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, pages 256–274, 1996.
- [37] Jacques Gignoux, Camille Duby, and Sébastien Barot. Comparing the performances of Diggle’s tests of spatial randomness for small samples with and without edge-effect correction: Application to ecological data. *Biometrics*, 55(1):156–164, 1999.
- [38] Samuel F Ginrich. Measuring and evaluating stocking and stand density in upland hardwood forests in the central states. *Forest Science*, 13(1):38–53, 1967.
- [39] Victor Goldsmith, Philip G McGuire, John B Mollenkopf, and Timothy A Ross. *Analyzing Crime Patterns: Frontiers of Practice*. Sage Publications, 1999.
- [40] Peter V Gorsevski, Randy B Foltz, Paul E Gessler, and Terrance W Cundy. Statistical modeling of landslide hazard using GIS. In *In: Proceedings of the Seventh Federal Interagency Sedimentation Conference, March 25 to 29, 2001, Reno, Nevada. Washington, DC: US Inter-agency Committee on Water Resources, Subcommittee on Sedimentation: XI-103-XI-109*, 2001.

- [41] Ulf Grenander. Statistical geometry: a tool for pattern analysis. *Bulletin of the American Mathematical Society*, 79(5):829–856, 1973.
- [42] Yongtao Guan, Michael Sherman, and James A Calvin. Assessing isotropy for spatial point processes. *Biometrics*, 62(1):119–125, 2006.
- [43] Peter Haase. Spatial pattern analysis in ecology based on Ripley’s K-function: Introduction and methods of edge correction. *Journal of Vegetation Science*, 6(4):575–582, 1995.
- [44] Jacques Hadamard. *Lessons in Geometry: Plane Geometry*. American Mathematical Society, 2008.
- [45] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons, 2008.
- [46] Manoj K Jain and Vijay P Singh. DEM-based modelling of surface runoff using diffusion wave equation. *Journal of Hydrology*, 302(1-4):107–126, 2005.
- [47] Xu Jingwen, Zhang Wanchang, Sun Chengwu, and Fu Congbin. An efficient method on deriving topographic index from DEM for land surface hydrological model simulations. *Journal of Meteorological Research*, 23(5):609–616, 2009.
- [48] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous Univariate Distributions*, volume 1. Houghton Mifflin Boston, 1970.
- [49] Daisaku Kawabata and Joel Bandibas. Landslide susceptibility mapping using geological data, a DEM from ASTER images and an artificial neural network (ANN). *Geomorphology*, 113(1-2):97–109, 2009.
- [50] Mark Kumler. An intensive comparison of triangulated irregular networks (TINs) and digital elevation models (DEMs). *Cartographica*, 31(2):1, 1994.
- [51] Yu A Kutoyants. *Statistical Inference for Spatial Poisson Processes*, volume 134. Springer Science & Business Media, 2012.
- [52] Phaedon C Kyriakidis, Ashton M Shortridge, and Michael F Goodchild. Geostatistics for conflation and accuracy assessment of digital elevation models. *International Journal of Geographical Information Science*, 13(7):677–707, 1999.
- [53] Jay Lee. Comparison of existing methods for building triangular irregular network, models of terrain from grid digital elevation models. *International Journal of Geographical Information System*, 5(3):267–285, 1991.
- [54] Fasheng Li and Lianjun Zhang. Comparison of point pattern analysis methods for classifying the spatial distributions of spruce-fir stands in the north-east usa. *Forestry*, 80(3):337–349, 2007.

- [55] Paul Longley and Michael Batty. *Advanced Spatial Analysis: the CASA book of GIS*. ESRI, Inc., 2003.
- [56] Sotirios E Louridas and Michael T Rassias. Problem-solving and selected topics in Euclidean geometry. *AMC*, 10:12, 2013.
- [57] Salvatore Manfreda, Margherita Di Leo, and Aurelia Sole. Detection of flood-prone areas using digital elevation models. *Journal of Hydrologic Engineering*, 16(10):781–790, 2011.
- [58] Jerrold E Marsden and Michael J Hoffman. *Elementary Classical Analysis*. W. H. Freeman, 1993.
- [59] Eric S McCord and Jerry H Ratcliffe. Intensity value analysis and the criminogenic effects of land use features on local crime patterns. *Crime Patterns and Analysis*, 2(1):17–30, 2009.
- [60] Ronald C McDonald, RF Isbell, James G Speight, Joe Walker, and MS Hopkins. *Australian Soil and Land Survey: Field Handbook*. Number Ed. 2. CSIRO publishing, 1998.
- [61] Bruce E Meserve. *Fundamental Concepts of Geometry*. Courier Corporation, 2014.
- [62] Harvey J Miller and Elizabeth A Wentz. Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers*, 93(3):574–594, 2003.
- [63] Helena Mitasova, Jaroslav Hofierka, Maros Zlocha, and Louis R Iverson. Modelling topographic potential for erosion and deposition using GIS. *International Journal of Geographical Information Systems*, 10(5):629–641, 1996.
- [64] Jesper Møller and Håkon Toftaker. Geometric anisotropic spatial point pattern analysis and Cox processes. *Scandinavian Journal of Statistics*, 41(2):414–435, 2014.
- [65] Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.
- [66] Jesper Moller and Rasmus Plenge Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, 2003.
- [67] Marc Moore. On the estimation of a convex set. *The Annals of Statistics*, pages 1090–1099, 1984.
- [68] Mari Myllymäki, Mikko Kuronen, and Tomáš Mrkvička. Testing global and local dependence of point patterns on covariates in parametric models. *Spatial Statistics*, 2020. doi: 10.1016/j.spasta.2020.100436.
- [69] I Newton, Mick Marquiss, DN Weir, and Dorian Moss. Spacing of Sparrowhawk nesting territories. *The Journal of Animal Ecology*, pages 425–441, 1977.

- [70] Atsuyuki Okabe and Kokichi Sugihara. *Spatial Analysis Along Networks: Statistical and Computational Methods*. John Wiley & Sons, 2012.
- [71] Guy Ouillon and Didier Sornette. Segmentation of fault networks determined from spatial clustering of earthquakes. *Journal of Geophysical Research: Solid Earth*, 116(B2), 2011.
- [72] Yashon Ouma. Evaluation of multiresolution digital elevation model (DEM) from real-time kinematic GPS and ancillary data for reservoir storage capacity estimation. *Hydrology*, 3(2):16, 2016.
- [73] RR Phelps. Convex sets and nearest points. *Proceedings of the American Mathematical Society*, 8(4):790–797, 1957.
- [74] Arne Pommerening and Dietrich Stoyan. Edge-correction needs in estimating indices of spatial forest structure. *Canadian Journal of Forest Research*, 36(7):1723–1739, 2006.
- [75] Charles Chapman Pugh. *Real Mathematical Analysis*, volume 2011. Springer, 2002.
- [76] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. url:<http://www.R-project.org/>.
- [77] Tuomas Rajala, Claudia Redenbach, Aila Särkkä, and Martina Sormani. A review on anisotropy analysis of spatial point patterns. *Spatial Statistics*, 28:141–168, 2018.
- [78] Jean-Paul Rasson, Marcel Rémon, Tite Kubushishi, and Florence Henry. Finding the edge of a Poisson forest with inside and outside observations: a theoretical point of view. In *Internal Report 94/22*. Department of Mathematics, FUNDP Namur, 1994.
- [79] JP Rasson, M Rémon, and Fl Henry. Finding the edge of a Poisson forest with inside and outside observations: The discriminant analysis point of view. In *From Data to Knowledge*, pages 94–101. Springer, 1996.
- [80] Mandla V Ravibabu and Kamal Jain. Digital elevation model accuracy aspects. *Journal of Applied Sciences*, 8(1):134–139, 2008.
- [81] M Rémon. A discriminant analysis algorithm for the inside/outside problem. *Computational Statistics & Data Analysis*, 23(1):125–133, 1996.
- [82] Marcel Rémon. The estimation of a convex domain when inside and outside observations are available. *Supplemento ai Rendiconti del Circolo Matematico di Palermo*, 35:227–235, 1994.
- [83] Marcel Rémon. Discriminant analysis tools for non convex pattern recognition. In *Data Analysis, Classification, and Related Methods*, pages 241–246. Springer, 2000.
- [84] BD Ripley and J-P Rasson. Finding the edge of a Poisson forest. *Journal of Applied Probability*, 14(3):483–491, 1977.

- [85] Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977.
- [86] Brian D Ripley. *Spatial Statistics*, volume 575. John Wiley & Sons, 2005.
- [87] Timothy P Robinson. Spatial statistics and geographical information systems in epidemiology and public health. *Advances in Parasitology*, 47:81–128, 2000.
- [88] R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.
- [89] Marcos S Rodrigues, José E Corá, and Carolina Fernandes. Soil sampling intensity and spatial distribution pattern of soils attributes and corn yield in no-tillage system. *Engenharia Agrícola*, 32(5):852–865, 2012.
- [90] Alberto Rodríguez Casal. Set estimation under convexity type assumptions. In *Annales de l'Institut Henri Poincaré Probabilités et statistiques*, volume 43, pages 763–774, 2007.
- [91] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real Analysis*, volume 32. Macmillan New York, 1988.
- [92] Behara Seshadri Daya Sagar. *Mathematical Morphology in Geomorphology and GISci*. Chapman and Hall/CRC, 2013.
- [93] D Sathymoorthy, R Palanikumar, and BSD Sagar. Morphological segmentation of physiographic features from DEM. *International Journal of Remote Sensing*, 28(15):3379–3394, 2007.
- [94] Gabriel B Senay, Andrew D Ward, John G Lyon, Norman R Fausey, and Sue E Nokes. Manipulation of high spatial resolution aircraft remote sensing data for use in site-specific farming. *Transactions of the ASAE*, 41(2):489, 1998.
- [95] Jean Serra. *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, Inc., 1983.
- [96] Frank Y Shih. *Image Processing and Mathematical Morphology: Fundamentals and Applications*. CRC press, 2017.
- [97] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press, 1986.
- [98] Pierre Sollié. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag Berlin Heidelberg, 2 edition, 2004.
- [99] Robert W Sterner, Christine A Ribic, and George E Schatz. Testing for life historical changes in spatial patterns of four tropical tree species. *The Journal of Ecology*, pages 621–633, 1986.

- [100] A Stewart Fotheringham and Peter A Rogerson. GIS and spatial analytical problems. *International Journal of Geographical Information Science*, 7(1):3–19, 1993.
- [101] Hua Sun, Peter S Cornish, and TM Daniell. Spatial variability in hydrologic modeling using rainfall-runoff model and digital elevation model. *Journal of Hydrologic Engineering*, 7(6):404–412, 2002.
- [102] Eric C Tate, David R Maidment, Francisco Olivera, and David J Anderson. Creating a terrain model for floodplain mapping. *Journal of Hydrologic Engineering*, 7(2):100–108, 2002.
- [103] Joao Manuel Tavares and Jorge RM Natal. *Computational Modelling of Objects Represented in Images. Fundamentals, Methods and Applications: Proceedings of the International Symposium CompIMAGE in Engineering, Water and Earth Sciences*. Taylor & Francis, Inc., 2007.
- [104] Kenji Ueno, Kōji Shiga, Toshikazu Sunada, and Shigeyuki Morita. *A Mathematical Gift, III: The Interplay Between Topology, Functions, Geometry, and Algebra*, volume 3. American Mathematical Society, 2003.
- [105] Eduardo Velázquez, Isabel Martínez, Stephan Getzin, Kirk A Moloney, and Thorsten Wiegand. An evaluation of the state of spatial point pattern analysis in ecology. *Ecography*, 39(11):1042–1055, 2016.
- [106] Lance A Waller and Carol A Gotway. *Applied Spatial Statistics for Public Health Data*, volume 368. John Wiley & Sons, 2004.
- [107] Qingchun Wen, Xin Chen, Yi Shi, Jian Ma, and Qian Zhao. Analysis on composition and pattern of agricultural nonpoint source pollution in Liaohe River basin, China. *Procedia Environmental Sciences*, 8:26–33, 2011.
- [108] Qihao Weng. Quantifying uncertainty of digital elevation models derived from topographic maps. In *Advances in Spatial Data Handling*, pages 403–418. Springer, 2002.
- [109] Thorsten Wiegand and Kirk A. Moloney. Rings, circles, and null-models for point pattern analysis in ecology. *Oikos*, 104(2):209–229, 2004.
- [110] Thorsten Wiegand, Savithri Gunatilleke, and Nimal Gunatilleke. Species associations in a heterogeneous Sri Lankan Dipterocarp forest. *The American Naturalist*, 170(4):E77–E95, 2007.
- [111] Thorsten Wiegand and Kirk A Moloney. *Handbook of Spatial Point Pattern Analysis in Ecology*. Chapman and Hall/CRC, 2013.
- [112] Stefan Wiemer. Earthquake statistics and earthquake prediction research. *Institute of Geophysics; Zürich, Switzerland*, 2000.

- [113] Paul R Wolf and Bon A Dewitt. *Elements of photogrammetry: with applications in GIS*, volume 3. McGraw-Hill New York, 2000.
- [114] Zhixiao Xie and Jun Yan. Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5):396–406, 2008.
- [115] Ikuho Yamada and Peter A Rogerson. An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis*, 35(2):97–109, 2003.
- [116] Bisheng Yang, Qingquan Li, and Wenzhong Shi. Constructing multi-resolution triangulated irregular network model for visualization. *Computers & Geosciences*, 31(1):77–86, 2005.
- [117] Wenhao Yu, Tinghua Ai, and Shiwei Shao. The analysis and delimitation of central business district using network kernel density estimation. *Journal of Transport Geography*, 45:32–47, 2015.