

Enhanced point pattern analysis on nonconvex spatial domains

by

Kabelo Mahloromela

14194237

Supervisor: Prof. Inger Nicolette Fabris-Rotelli



**UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA**

Submitted in fulfillment of the requirements for the degree
Philosophiae Doctor (Mathematical Statistics)
in the Faculty of Natural and Agricultural Sciences, Department of Statistics
University of Pretoria, Pretoria

November 2024

Enhanced point pattern analysis on nonconvex spatial domains

by

Kabelo Mahloromela

Abstract

Point pattern analysis is the study of the spatial arrangement of points in space, usually two-dimensional space. The points arise from a stochastic mechanism, termed a point process, whose characteristics are of scientific interest. The properties of point patterns are characterised using statistical measures that are a function of the study area and distance. Consequently, the domain in which points are observed and the distance metric used to quantify proximity between points plays an important role. Convex domains with the Euclidean distance are often used. This choice of domain and distance measure, however, makes an implicit assumption that all points are connected in a space without obstacles. In real-world applications, points may be constrained by their environments, thus a convex window and the Euclidean distance may not correctly capture spatial proximity relationships and restrictions imposed by the domain's geometry. In this thesis, a presentation of methodology that accounts for the nonconvex structure of the spatial domain in point pattern analysis is provided. Firstly, consideration is given to the selection of nonconvex windows (when unknown) for point patterns realised from a process that is governed by a covariate. The proposed algorithm uses a weighted distance-based outlier scoring scheme that considers the distribution of covariates at observed data point locations. The robustness of the algorithm is demonstrated through a simulation study. Subsequently, a framework is developed to quantify proximity relationships using a graph theoretic approach based on visibility graphs. This characterisation of distance is used to extend first- and second-order point pattern measures for appropriate use on nonconvex domains. Finally, we provide an implementation strategy to efficiently compute summary measures based on a query to the visibility graph.

Keywords: point pattern, spatial domain, nonconvex, visibility graph, Euclidean distance

Supervisor : Prof. I. N. Fabris-Rotelli
Department : Department of Statistics, University of Pretoria
Degree : Philosophiae Doctor (Mathematical Statistics)

Declaration

I, *Kabelo Mahloromela*, declare that the thesis, which I hereby submit for the degree PhD in Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution. This research was approved by the Faculty of Natural and Agricultural Science Research Ethics committee at the University of Pretoria under the reference NAS339/2019.

Signature:  _____

Date: 29 November 2024

Acknowledgements

I would like to express my sincerest gratitude to the many people who have provided support and encouragement throughout this challenging but rewarding journey. I am wholeheartedly grateful to my supervisor, Prof. Inger Fabris-Rotelli. I appreciate the time and effort you have invested, which not only made this research possible but also helped me grow as an academic and individual. You are a beacon of inspiration and your insight, guidance, and patience have been invaluable. To my friends, too many to list, thank you for your empathy and camaraderie which have kept me balanced and sane throughout this work. To my family - my parents, Ntombifuthi Mahloromela and Masilo Mahloromela, my sister, Keneilwe Mahloromela, and other distant relatives, I am grateful for your endless support. Thank you for your love, encouragement, and unwavering belief in me.

This research also received support from

- STATOMET, the Bureau for Statistical and Survey Methodology, in the Department of Statistics at the University of Pretoria,
- ESRI, the Environmental Systems Research Institute in South Africa,
- NRF, the National Research Foundation of South Africa (Grant Number 137785), and
- DST-NRF-SAMRC SARChI, the South Africa National Research Foundation and South Africa Medical Research Council (South Africa DST-NRF-SAMRC SARChI Research Chair in Biostatistics, Grant number 114613).

The content and opinions expressed herein are the sole responsibility of the authors and do not necessarily represent the official views of STATOMET, ESRI, and the NRF.

Contents

List of Figures	iv
List of Tables	viii
1 Introduction	1
1.1 Spatial point processes	1
1.2 First- and second-order properties	1
1.2.1 First-order properties	2
1.2.2 Second-order properties	2
1.3 Analysis on nonconvex spatial domains	3
1.4 On the use of visibility graphs to quantify proximity relationships	5
1.4.1 Weighted visibility graph	5
1.5 Contributions	5
1.5.1 Published work	6
1.5.2 Paper Competition Awards	6
1.5.3 Presented work	7
1.6 Thesis Outline	7
2 Point pattern domain estimation using distance-based outlier scores	9
2.1 Abstract	9
2.2 Introduction	10
2.3 Methodology	12
2.3.1 Point patterns	12
2.3.2 Nearest-neighbour distance outlier score	12

2.3.3	Otsu thresholding	13
2.3.4	Window selection algorithm	14
2.4	Simulation study	16
2.4.1	Simulation outline	17
2.4.2	Simulating the covariate	17
2.4.3	Using the covariate to define the window	19
2.4.4	Simulating the point pattern	19
2.4.5	Evaluation metrics	19
2.4.6	Results	21
2.5	Application	23
2.5.1	Landscape and terrain features	27
2.6	Discussion	33
2.7	Conclusion	38
3	Kernel-smoothing of point patterns constrained on nonconvex domains	39
3.1	Abstract	39
3.2	Introduction	40
3.3	Methodology	41
3.3.1	Spatial point pattern data	42
3.3.2	First-order intensity	42
3.3.3	Weighted visibility graph	43
3.3.4	Modified kernel-smoothed intensity estimate	45
3.4	Application	50
3.5	Discussion	52
3.6	Conclusion	55
4	Assessing point pattern interactions constrained on nonconvex domains using visibility graphs	56
4.1	Abstract	56
4.2	Introduction	56
4.3	Methodology	58
4.3.1	Spatial point pattern data	58

4.3.2	Second-order quantities	59
4.3.3	Weighted visibility graph	62
4.3.4	Shortest path distance	63
4.4	Visibility graph-based adaptations	63
4.4.1	Visibility-based pairwise, nearest-neighbour, and empty-space distance	63
4.5	Results	65
4.6	Discussion	66
4.7	Conclusion	67
5	A framework for analysing point patterns on nonconvex domains using visibility graphs and multidimensional scaling	71
5.1	Abstract	71
5.2	Introduction	72
5.3	Background Theory	73
5.3.1	Spatial point pattern data	74
5.3.2	First- and second-order properties	74
5.3.3	Weighted visibility graph	76
5.3.4	Shortest path distance	76
5.3.5	Multidimensional scaling (MDS)	77
5.4	Methodology	79
5.5	Results	84
5.6	Discussion	87
5.7	Conclusion	90
6	Conclusion	91
	Bibliography	94

List of Figures

1.1	The straight-line (Euclidean distance) between two points confined on irregular nonconvex domains. Points are represented in a polygon of the Great Bear lake in Canada in (a), and the American state of Massachusetts in (b).	4
2.1	Spatial covariates simulated on a window $W^* = [0, 10] \times [0, 10]$ from a Gaussian random field with a Whittle-Matern covariance model and ranges 0.5 ((a) Covariate I and (b) Covariate II) and 1.5 ((c) Covariate III and (d) Covariate IV).	18
2.2	Simulation extent of points based on selecting thresholds for (a) Covariate I, (b) Covariate II, (c) Covariate III, and (d) Covariate IV.	20
2.3	Violin plots of the accuracy of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	24
2.4	Violin plots of the precision of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	25
2.5	Violin plots of the recall of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	26
2.6	Violin plots of the specificity of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	27
2.7	Violin plots of the F1-score of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	28
2.8	Violin plots of the intersection over union of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.	29

2.9	Household locations for 3 villages in Tanzania's Mara province.	30
2.10	Constructed window (right) and terrain slope (left) for villages in Tanzania's Mara province.	31
2.11	Google Earth satellite image of Iseresere village in Tanzania's Mara province, with household locations indicated with red dots.	32
2.12	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province.	34
2.13	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).	35
2.14	Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).	36
3.1	Illustration of the effect of using the Euclidean distance when estimating the kernel smoothed intensity function on a nonconvex window, shown for a single point event.	44
3.2	Shortest path (red line) between two points (green dots) and visibility graph (blue lines) on a nonconvex polygons.	45
3.3	Point pattern plots and intensity estimates on convex (left) and nonconvex (right) windows fitted using Equations 3.1 and 3.5, respectively.	47
3.4	Intensity estimates of simulated point pattern. (a) Point pattern simulated from a homogeneous Poisson process with intensity 0.5. (b) Intensity estimate using Equation 3.3. (c) Intensity estimate using Equation 3.7. (d) Pointwise difference between intensity estimate using Equations 3.3 and 3.7.	48
3.5	Heterogeneous intensity function of a point process on two different nonconvex windows (left) and corresponding raincloud plot (right) of ISE values in the case of the Euclidean shortest path (ESP) intensity estimator and diffusion (DF) smoothing.	50
3.6	Google Earth plot of village household locations for three villages in Tanzania's Mara province with terrain map	53

3.7	Point pattern plot (left) and modified kernel smoothed intensity estimate plot (right) for villages in Tanzania’s Mara province, on nonconvex windows constructed using terrain slope covariate data	54
4.1	An illustration showing the effect of using the Euclidean distance (green arrows) and shortest path distance (orange arrows) to quantify proximity on nonconvex windows when computing several types of distances used in second-order quantities	58
4.2	Modified K , F , G , and J -function estimates for a simulated clustered point pattern.	68
4.3	Modified K , F , G , and J -function estimates for a simulated random point pattern.	69
4.4	Modified K , F , G , and J -function estimates for a simulated regular point pattern.	70
5.1	Shortest path (red line) between two points (green dots) and visibility graph (blue lines) on a nonconvex polygons.	77
5.2	An illustration of the result of embedding (a) the shortest path distances on a weighted visibility graph in a Euclidean space. (a) shows the path of the shortest path in the original nonconvex domain and (b) shows the result after applying MDS.	79
5.3	A discretisation of a nonconvex polygon using a triangular mesh in (a) and its corresponding representation in Euclidean space after the application of MDS in (b).	80
5.4	The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), a U-shaped window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.	81

5.5	The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), an S-shaped window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.	82
5.6	The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), an irregular window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.	83
5.7	A point pattern realised from a point process with inhomogeneous intensity.	85
5.8	A discretisation of a nonconvex polygon using a triangular mesh in (a) and its corresponding representation in Euclidean space after the application of MDS in (b).	86
5.9	A kernel smoothed intensity estimate produced using (a) Equation 5.1 and (b) Equation 5.7.	87
5.10	Modified empirical K -function estimates for simulated (a) clustered, (b) random, and (c) regular point patterns.	88

List of Tables

3.1	Summary statistics of the ISE values when using the intensity estimators based on the Euclidean shortest path (ESP) distance and diffusion (DF) smoothing on the L-shaped nonconvex window.	51
3.2	Summary statistics of the ISE values when using the intensity estimators based on the Euclidean shortest path (ESP) distance and diffusion (DF) smoothing on an irregularly shaped nonconvex window.	51

Chapter 1

Introduction

Point pattern methodology is applied widely across various application areas to model the location of random events in space. Some application fields include hotspot detection in crime analysis [4, 3, 10], species distribution characterisation in ecology [74, 50, 28], disease spread modelling in epidemiology [24, 23], urban analysis in urban planning [51], and characterising the distribution of celestial bodies in astronomy [68].

1.1 Spatial point processes

A spatial point processes $X = \{X_1, \dots, X_n\}$ [39, 13, 30, 9] is a stochastic mechanism whose outcome is a finite (or countably infinite) set of points $x = \{x_1, \dots, x_n\}$ in a domain $S \subseteq \mathbb{R}^d$, typically $d = 2$. In practice, points are observed in a bounded region $W \subseteq S$, termed a window, which may assume any geometric shape. Point pattern data is assumed to be realised from a point process. The aim of a scientific inquiry then becomes estimating the properties of X using x .

1.2 First- and second-order properties

The properties of point processes are defined based on statistical measures that describe the density and interaction of points within a specified domain. These characteristics are so-called first- and second-order properties.

1.2.1 First-order properties

First-order properties are concerned with the individual characteristics of locations and their variation across space. The first-order property of a point process is described using the intensity measure

$$\mu(B) = E[N(B)],$$

where $E[\cdot]$ denotes the expected value and $N(B)$ is a random count variable of the number of points in a region $B \subseteq \mathbb{R}^d$. The function $\lambda(u)$, is called the intensity function of X if the intensity measure μ can be written as

$$\mu(B) = \int_B \lambda(u) du.$$

The intensity function is a local measure that gives the expected number of points per unit area. Stated more formally,

$$\lambda(u) = \lim_{|du| \rightarrow 0} \frac{E[N(du)]}{|du|},$$

where du is a small region around u , and has an area $|du|$. When $\lambda(u) = \lambda$ is constant for any given location in the spatial domain, the point process is said to be homogeneous or first-order stationary [39, 13, 30, 9]. The most common method for estimating the intensity function is kernel smoothing [18, 39, 9]. This is especially useful if there is no prior knowledge of the parametric form of the underlying process.

1.2.2 Second-order properties

Second-order properties describe the interactions or dependencies between points. These properties explain how the presence of a point at one location influences the probability of finding another point at some distance away, making them useful as tools to identify inter-point relationships such as clustering, regularity, or randomness. The second-order intensity [61, 39, 9] of a point process is defined as

$$\lambda_2(u, v) = \lim_{\substack{|du| \rightarrow 0 \\ |dv| \rightarrow 0}} \frac{E[N(du) \times N(dv)]}{|du| \times |dv|}.$$

where du and dv are small regions around u and v , and have areas $|du|$ and $|dv|$ respectively. A point pattern is considered second-order stationary when the statistical properties of the point pattern, namely the interpoint relationships, remain unchanged regardless of the location within the pattern [61, 39, 9]. This means that there is no overall trend or directionality in the point pattern, and the only factor influencing the relationship between points is the distance between them, not their absolute location, i.e., $\lambda_2(u, v)$ only depends on the distance $\|u - v\|$.

1.3 Analysis on nonconvex spatial domains

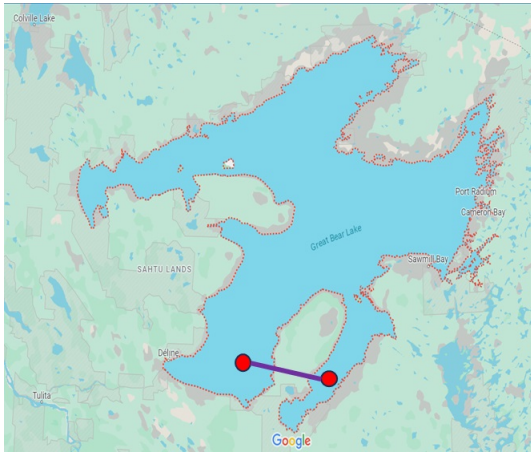
It is typical in point pattern methodology to capture the spatial structure of a set of locations in space by quantifying proximity relationships within a Cartesian coordinate system. This, however, overlooks spatial relationships that are governed by the topological structure of physically connected spaces, inherent in many spatial phenomena [27, 45]. When proximity relationships use straight-line (Euclidean) distances an implicit assumption is being made about the geometric structure of the observation domain, namely that it is convex [17].

Definition 1 A set $A \subseteq \mathbb{R}^d$ is convex if $\forall x, y \in A$ and $\alpha \in [0, 1]$, $\alpha x + (1 - \alpha)y \in A$.

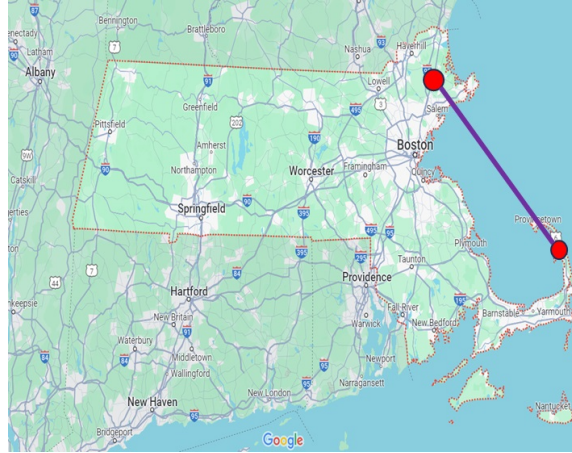
A convex set is thus a connected set, meaning that there exists a straight line path contained in the set between any two points. This is however not true for nonconvex domains for which there exists some points x and y in the domain for which $\alpha x + (1 - \alpha)y$ is not a fully contained element.

The effect of quantifying proximity relationships on nonconvex polygons is illustrated¹ in Figure 1.1. Figure 1.1(a) shows the straight-line distance between two points (i.e., fish) contained in a polygon of the Great Bear lake in Canada. Fish are physically restricted to the lake water, so using the Euclidean distance does not appropriately represent interactions in this space. A similar problem is shown in Figure 1.1(b) where points are confined to and interact on a connected landmass represented by the American state of Massachusetts. These issues, if not accounted for, appear as “tunnelling” artefacts in

¹Google Maps. Retrieved July 29, 2024, from <https://www.google.com/maps>



(a) Great Bear Lake



(b) The State of Massachusetts

Figure 1.1: The straight-line (Euclidean distance) between two points confined on irregular nonconvex domains. Points are represented in a polygon of the Great Bear lake in Canada in (a), and the American state of Massachusetts in (b).

point pattern analysis. For example, when estimating a varying function of intensity, the contribution of two points separated by disconnected spaces is overestimated [7].

The challenge of incorporating the topological structure of a domain in point pattern analysis has been considered for non-planar cases, such as points on a network of lines [8, 9, 77]. Okabe et al. [44], for example, develop techniques to address methodological pitfalls that arise when points are constrained in linear network spaces. Some adaptations to point pattern statistical summary measures have been proposed in these settings. These include the network kernel smoothed intensity estimate and network K -function [45, 8, 44].

Although there is an acknowledgement of the need to use appropriate distance metrics to measure spatial relationships [42, 44, 7], to our knowledge, methodology for planar point processes confined to nonconvex spatial domains is not well-developed. Owing to this, we propose using a representative distance measure on a path connected space, through the application of a weighted visibility graph.

1.4 On the use of visibility graphs to quantify proximity relationships

In this Section, the definition of a visibility graph is provided, and its usefulness in determining proximity relationships in point patterns as an alternative to the Euclidean distance is discussed.

1.4.1 Weighted visibility graph

A weighted visibility graph [25] of a polygon P , denoted by $G(P)$, is an ordered triple (V, E, w) comprising

- $V = \{v_1, \dots, v_m\}$, a set of the vertices² of P ,
- $E = \{(v_i, v_j) | v_i, v_j \in V \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, m\}$, a set of edges between vertices that are intervisible (in a geometric context, two vertices v_i and v_j , $v_i \neq v_j$, are visible to each other if the line segment connecting them does not intersect the outer boundary of P or hole boundaries, that is $\alpha v_i + (1 - \alpha)v_j \in P$, $\forall \alpha \in (0, 1)$), and
- $w : E \mapsto \mathbb{R}^+$, a function mapping edges to the value of the Euclidean distance between the vertices it connects.

A representative distance measure between two points in a point pattern, one that quantifies the length of a path that does not cross the window's boundary and holes in its domain, can be found by determining the shortest path on the weighted visibility graph of the window.

1.5 Contributions

The overarching aim of this thesis is to develop a novel framework to appropriately characterise spatial relationships in point patterns that are confined to nonconvex planar

²In the case of a polygon with holes, the vertices of the hole boundaries are included in the visibility graph vertex set.

regions. With this broader goal in mind, we have contributed the following:

- A novel window estimation technique to select nonconvex spatial domains (when unknown) based on covariates via distance-based outlier scoring and Otsu thresholding.
- A modified kernel smoothed intensity estimator that accounts for the constraints imposed by obstacles in the spatial domain and the domain boundaries through the use of a visibility graph.
- A visibility graph-based technique to quantify proximity relationships and inter-point interactions of points constrained in nonconvex domains.
- A framework to analyse point patterns constrained on nonconvex domains using visibility graphs and an embedding of the shortest path-distances on this graph in abstract Euclidean space using multidimensional scaling.

1.5.1 Published work

- K. Mahloromela, I.N. Fabris-Rotelli, C. Kraamwinkel. Covariate construction of nonconvex windows for spatial point patterns. South African Statistical Journal, 57(2):65-87, 2023

1.5.2 Paper Competition Awards

- South African Statistical Association 2023 postgraduate paper competition 1st prize winner (published works) - Covariate construction of nonconvex windows for spatial point patterns.
- South African Statistical Association 2024 postgraduate paper competition 3rd prize winner (unpublished works) - Enhanced point pattern analysis on nonconvex spatial domains.

1.5.3 Presented work

- Nonconvex window selection for spatial point pattern data (Using moving window statistics) - 63rd Annual Conference of the South African Statistical Association, SASA 2022, George, South Africa (28 November 2022 - 2 December 2022).
- Covariate construction of nonconvex windows for spatial point pattern data - 64th Annual Conference of the South African Statistical Association, SASA 2023, Durban, South Africa (27 November 2023 - 1 December 2023).
- Window selection for point patterns - 2023 International Symposium on Modern Biostatistics and Statistics hosted by the University of Pretoria, Pretoria, South-Africa (3 - 4 July 2023).
- Window selection for point patterns - Spatial Statistics Conference 2023, Boulder, United States of America (18 - 21 July 2023).
- Enhance point pattern analysis in nonconvex spatial domains - 2024 International Symposium on Modern Biostatistics and Statistics hosted by the University of Pretoria, Pretoria, South-Africa (26 - 28 June 2024).
- A window selection algorithm for enhanced spatial point pattern analysis (Outlier-based scoring method) - International Society of Photogrammetry and Remote Sensing (ISPRS) Technical Commission IV Symposium 2024, Perth, Australia (22 - 25 October 2024).
- Enhanced point pattern analysis on nonconvex spatial domains - 65th Annual Conference of the South African Statistical Association, SASA 2024, Stellenbosch, South Africa (18 - 22 November 2024).

1.6 Thesis Outline

This thesis comprises a series of academic articles that have been either published or are in preparation for submission to peer-reviewed journals. Consequently, each chapter

is designed to be self-contained. Owing to this, some definitions and discussions are restated therein. The organisation of this thesis is as follows.

- **Chapter 2** provides a presentation of an estimation technique that uses covariate-based distance outlier scoring to select nonconvex domains. It improves on the work published in the paper entitled *Covariate construction of nonconvex windows for spatial point patterns* [36] and its robustness is demonstrated against several point process models.
- **Chapter 3** is dedicated to a presentation of a kernel-smoothed intensity estimation technique for point patterns constrained on nonconvex domains. The estimator developed is based on a weighted visibility graph. This work forms part of research that has been published in the article entitled *Covariate construction of nonconvex windows for spatial point patterns* [36] where it is applied as a validation technique. The paper is extended in this chapter to include a comparison between this new proposed estimator and its only other known competitor, the diffusion smoother of [7].
- **Chapter 4** gives an extension of several second-order quantities used to assess spatial interaction in planar point patterns via the use of a visibility graph.
- **Chapter 5** provides an implementation strategy to efficiently compute the measures developed in Chapters 3 and 4 by embedding visibility graph-based distances in abstract higher-dimensional Euclidean space using multidimensional scaling.
- **Chapter 6** finishes off with a concluding discussion and final remarks.

Chapter 2

Point pattern domain estimation using distance-based outlier scores

For many analyses of point pattern data, it is important to have knowledge of the window in which points are observed. This is a methodological problem that is inherent to spatial point patterns since estimation is done on the window or is a function of the window's area. Owing to this, it is important to choose the window appropriately so that the spatial structure in the data is captured correctly. This chapter is dedicated to a presentation of an estimation technique to choose the spatial domain of a point pattern in a setting where points are constrained by physical or other phenomenon. It differs from the work published in [36] in that it is based on an outlier-scoring technique and Ostu thresholding. Whereas in [36] moving window statistics and the minimum and maximum of these statistics are used.

2.1 Abstract

To conduct an appropriate analysis of point pattern data, knowledge of the location of events and the spatial domain, termed a window, in which these locations are observed is required. The window defines the extent for analysis and is directly involved in estimating and inferring on the first-and-second order properties of the point process that generated the data. Often, the window is known and recorded during data collection.

Otherwise, it must be chosen objectively by the researcher. When the window is unknown, the typical approach is to use the minimum bounding box or the convex hull of the observed locations. Choosing too large a window, however, may lead to spurious estimation and inference in regions where points cannot occur. In a setting where points are restricted by physical or other phenomena, a nonconvex window that accounts for these constraints provides a more representative domain for analysis. Herein, we propose a domain estimation technique for point pattern data using covariate-based distance outlier scores and Otsu thresholding. The robustness of the algorithm is evaluated on various point pattern types namely regular, clustered, inhomogeneous, and completely spatially random point patterns. The algorithm is then applied to household locations of rural villages in Tanzania.

Keywords: point pattern, window domain, covariate, nonconvex

2.2 Introduction

In point pattern analysis, the window is often given a priori. In this setting, the window boundaries are known and can be recorded during data collection. For example, in epidemiological applications, disease incidence locations may form points in a point pattern, with municipal boundaries delineating the window's perimeter [23, 55]. In other cases, the window boundaries may not be clearly defined and must be chosen objectively by the researcher, e.g., the boundary of nesting territories in an ecological study [43, 16]. Herein, we consider the latter.

Selecting the window must be done carefully as it gives information about where observations were and were not made and where they could be predicted [9]. A larger window choice may result in spurious estimation and incorrect conclusions about the properties of the underlying point process: since the description of these properties rely implicitly or explicitly on the specification of the window, e.g., when computing density measures based on the area of window, when correcting for edge effects [9, 18, 7], and when using distance to quantify spatial dependence and correlation [19, 12].

Efforts to infer the window based on observed data locations have been made. Ripley and Rasson [60] proposed a technique for window selection by reconstructing an unknown

compact convex set from a realization of a homogeneous planar Poisson process observed from this unknown set. Their solution is the dilation of the convex hull about its centroid. Moore [40] presents a method for addressing a similar problem; i.e., estimate a compact convex set, given independent observations sampled uniformly from the unknown set. Rasson et al [54, 53, 58] extend these methods to estimate convex sets with observations that are inside and outside the convex set. Other literature that address determining a convex set from a random set of points include and are not limited to [20] and [17].

While standard methods for window selection typically consider rectangular or convex domains, real-world applications may involve nonconvex windows that emerge due to phenomenon that constrain the occurrence of points: for example, features such as mountains, valleys, rivers, or deserts may be obstacles that create discontinuities or boundaries in space that limit where points can move to or be placed. In [36], a method was developed to construct nonconvex windows by using covariate information in an algorithm that makes use of moving window statistics. The algorithm, however, was only tested on real cases, and its suitability for different point patterns was not investigated.

In this chapter, a new algorithm for nonconvex window selection is provided that makes use of a weighted repeated sample nearest neighbour distance outlier score and Otsu's method for automatic threshold selection. We work in a setting where points in a point pattern are constrained by physical or other phenomenon, represented as spatial covariates. The algorithm's robustness is evaluated on point patterns with various first- and second-order characteristics to assess its behaviour, reliability, and generalisability across different point pattern types. The algorithm is then applied to select the appropriate window of village household locations in rural Tanzania.

The remainder of this chapter is organised as follows. Section 2.3 provides some definitions and a presentation of the proposed nonconvex window selection algorithm. The performance of the algorithm is evaluated via a simulation study in Section 2.4. An application to rural village household locations in Tanzania is given in Section 2.5, a discussion in Section 2.6, and concluding remarks in Section 2.7.

2.3 Methodology

Herein, a formal definition of a point pattern data set is provided in Section 2.3.1. The proposed nonconvex window selection algorithm makes use of repeatedly sampled nearest-neighbour distance outlier scores and Otsu thresholding, thus a discussion of these topics is provided in Sections 2.3.2 and 2.3.3 respectively. The nonconvex window selection algorithm based on covariate data is outlined in Section 2.3.4.

2.3.1 Point patterns

A point pattern [9, 13, 30] data set typically comprises a set of locations, $x = \{x_1, \dots, x_n\}$, observed in a spatial domain $W \subset \mathbb{R}^2$, where $n \geq 0$ is not predetermined. Additional covariate information, as a spatial measurement $z(u)$ with $u \in W$, may also be available. In practice, the points u usually form a regular lattice extracted from continuous data over W and do not necessarily coincide with the data points. When covariate information is available, the dependence of a point pattern on the covariate should be investigated and quantified. A more thorough discussion of these techniques are provided in [9, 6, 41].

2.3.2 Nearest-neighbour distance outlier score

Covariates at observed locations represent a random sample¹ of possible data values of a point pattern. When points are constrained by covariates, the distribution of covariate values at observed locations will differ significantly from those at random locations generated by a mechanism independent of the covariate. A value that deviates significantly from the set of covariates at observed locations may thus provide an indication that it arose from a mechanism that is different to that of the possible data values. Owing to this, the use of a distance-based outlier detection framework is proposed to classify covariate values into two sets: inliers, which coincide with possible point locations, and outliers, where points are unlikely to occur. In particular, the outlier-score of [48] is used. The score is defined as the distance between a point, u , to its nearest neighbour in repeated independent random samples of the data, i.e.,

¹Albeit in some cases a biased sample, e.g., for clustered point patterns

$$q(u) := \frac{1}{r} \sum_{j=1}^r \min_{s \in S_j(x)} \|u - s\|,$$

where $S_j(x)$ is the j -th random sample of the data set $x = \{x_i\}_{i=1}^n$, $\|u - s\|$ is the Euclidean distance between u and s , and r is the number of random subsets. This technique combines an ensemble of multiple nearest-neighbour outlier scores from data re-samples. Data values that obtain relatively large scores are considered outliers. In order to classify a data point as an inlier or outlier based on its outlier score, a threshold value must be chosen. For this task, we use Otsu thresholding.

2.3.3 Otsu thresholding

Otsu thresholding [46] is a technique that is widely used for automatic greyscale image segmentation. The image pixels are divided into two groups, namely foreground and background, by a threshold τ , i.e., a set of pixel values less than or equal to τ and a set of pixel values greater than τ (relabelled as 0 and 1 respectively).

An image is represented as a $M \times K$ array of pixels with greyscale values $\{y_i\}_{i=1}^L$, where L is the total number of pixels. Let $I(\cdot)$ be an indicator function that is 1 when a condition is true and 0 otherwise. Otsu's thresholding method exhaustively searches for a threshold that maximises the inter-class variance,

$$\sigma^2(\tau) = \omega_0(\tau)\omega_1(\tau)(\mu_0(\tau) - \mu_1(\tau))^2,$$

where

$$\omega_0(\tau) = \frac{1}{L} \sum_{i=1}^L I(y_i \leq \tau),$$

is the fraction of background pixels,

$$\omega_1(\tau) = 1 - \omega_0(\tau) = \frac{1}{L} \sum_{i=1}^L I(y_i > \tau),$$

is the fraction of foreground pixels,

$$\mu_0(\tau) = \sum_{i=1}^L \frac{y_i I(y_i \leq \tau)}{L\omega_0(\tau)},$$

is the average of the background pixel values, and

$$\mu_1(\tau) = \sum_{i=1}^L \frac{y_i I(y_i > \tau)}{L\omega_1(\tau)},$$

is the average of the foreground pixel values.

Otsu thresholding aims to find a value that best separates an image into foreground and background pixels. The inter-class variance is the criterion used to measure separability between these group of pixels, with larger values indicating better separation. The process tests each pixel value as a threshold, grouping pixels into foreground or background based on whether they are below or above this value. The proportion of background and foreground pixels is determined along with the average of the pixel values for each group. These are then used to compute the inter-class variance. The pixel value that yields the highest inter-class variance is chosen as the optimal threshold. We apply this technique to automatically select a threshold for the outlier-scores. In our case the outlier-scores determined based on covariates denote the pixel values for which we was to select a threshold for.

2.3.4 Window selection algorithm

In this section we present a new algorithm for nonconvex window selection. The algorithm constructs nonconvex spatial domains by making use of covariate data. This algorithm differs in approach from that provide in [36], i.e., in [36] the window is selected using moving window statistics and thresholds are determined based on the minimum and maximum values of the statistical summary measures.

It is important to investigate the dependence of the point pattern on the covariate before implementation of this algorithm. Ideally, covariates believed to impact the distribution and abundance of points, or that are correlated to them should be used, see [9, 6, 5].

Let $x = \{x_1, \dots, x_n\}$ denote the set of observed locations in the point pattern. The points are assumed to be observed from an unknown domain $W \subset \mathbb{R}^2$. Let $z(u)$ be the value of a spatial covariate at location $u \in W^*$ where $W^* \supset W$. The covariate is represented as an image with m pixel cells, denoted by P_j , $j = 1, \dots, m$, each holding the

value of the covariate, i.e., $z(u_j)$ is the value of the covariate associated with the j -th pixel, where u_j is the location of the pixel center. Without loss of generality we suppose that W^* is the smallest bounding rectangular window that contains all the observed data points.

The algorithm makes use of the notion that the covariate values at the observed points are a random subset of potential data values within the point pattern, although in some instances, the sample may be biased. When the placement of points is influenced by covariates, the covariate distribution at observed points will deviate substantially from that at randomly chosen points generated by a process not tied to the covariate. Thus, to determine whether a given value of a covariate is “anomalous” relative to the covariates at observed locations, the repeated sampled nearest neighbour distance outlier score is used, given by

$$q(u) = \frac{1}{r} \sum_{k=1}^r \min_{s \in S_k(x)} \|z(u) - z(s)\|,$$

where $S_k(x)$ is the k -th weighted random sample of the data set $x = \{x_i\}_{i=1}^n$, $\|z(u) - z(s)\|$ is the absolute difference between the covariate values $z(u)$ and $z(s)$, and r is the number of random subsets. To account for potential selection bias in the values of covariates induced by point patterns with high clustering and inhomogeneous intensity, $S_k(x)$ is chosen as a weighted random sample (with replacement) where the weight w_i at a location x_i is the inverse kernel density estimate of the locations $\{x_i\}_{i=1}^n$,

$$w_i = \left(n^{-1} \sum_{v=1}^n K_h(x_i - x_v) \right)^{-1},$$

where $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ is scaled kernel weighting function, $K(\cdot)$ is a probability density on $\mathbb{R}^2$, and $h > 0$ is a smoothing bandwidth.

When the value of the covariates at a given location differs substantially from those at observed locations, the outlier score will be significantly large. To increase the contrast between inlier and outlier scores, we apply a log transformation to the values of the scores and then use Otsu’s method to classify the pixels associated with the values of the transformed scores into an inlier and outlier group. The final window is constructed by taking the union of all pixel cells associated with covariates what are part of the inlier group that is produced via Otsu thresholding. The threshold, τ , chosen maximises

$$\sigma_q^2(\tau) = \alpha_0(\tau)\alpha_1(\tau)(\bar{q}_0(\tau) - \bar{q}_1(\tau))^2,$$

where

$$\alpha_0(\tau) = \frac{1}{m} \sum_{j=1}^m I(-\ln(q(u_j)) \leq \tau),$$

is the fraction of transformed outlier scores less than or equal to τ ,

$$\alpha_1(\tau) = 1 - \alpha_0(\tau) = \frac{1}{m} \sum_{j=1}^m I(-\ln(q(u_j)) > \tau),$$

is the fraction of transformed outlier scores greater than τ ,

$$\bar{q}_0(\tau) = \sum_{j=1}^m \frac{-\ln(q(u_j))I(-\ln(q(u_j)) \leq \tau)}{m\alpha_0(\tau)},$$

is the average of transformed outlier scores less than or equal to τ , and

$$\bar{q}_1(\tau) = \sum_{j=1}^m \frac{-\ln(q(u_j))I(-\ln(q(u_j)) > \tau)}{m\alpha_1(\tau)},$$

is the average of transformed outlier scores greater than τ . The resultant window is given by

$$\widehat{W} = \bigcup_{\forall j \ni \ln(q(u_j))^{-1} \leq \tau} P_j.$$

Opening and closing (mathematical morphological operations) can be applied to the constructed window as a post-processing step to remove noise and improve the window smoothness. This new algorithm better accounts for the bias in the sampled covariates that is induced by heterogeneity that arises in certain point pattern types, which was not considered in [36]. Next, we perform a simulation study to evaluate the robustness of the algorithm to different point patterns.

2.4 Simulation study

In this section, we perform a simulation study to test the robustness of the window selection algorithm against point patterns with different first- and second-order properties.

All computations are performed using the R programming language [52]. Simulation of point patterns and spatial covariates are done using functions in the `spatstat` [9] and `SpatialExtremes` [59] packages respectively.

2.4.1 Simulation outline

The performance of the proposed methodology is evaluated with simulated data. A basic outline of the simulation is provided below:

- Select (or simulate) a spatial covariate. In the case of simulation, a Gaussian random field [67] is used.
- Choose a threshold for the covariate values that will delineate the simulation domain for point pattern locations. The window is defined as the set of locations that coincide with covariate values based on the following cases:
 - **Case 1:** Points may only occur at locations with a covariate value that is less than the chosen threshold.
 - **Case 2:** Points may only occur at locations with a covariate value that is greater than the chosen threshold.
 - **Case 3:** Points may only occur at locations with covariate values that lie between two chosen threshold limits.
- Simulate different point patterns on the created window with different first- and second-order characteristics, namely point patterns that are completely spatially random, regular, and clustered.
- To evaluate the quality of the constructed window, statistics based on the area of the set difference between the simulation extent and the constructed window are used.

2.4.2 Simulating the covariate

A Gaussian random field with a Whittle-Matern covariance model [67, 75] is used to simulate the spatial covariate in a window $W^* = [0, 10] \times [0, 10]$. Two different values,

0.5 and 1.5, are used as range parameters in the Whittle-Matern covariance model to simulate spatial covariates with short and moderate autocorrelation structures respectively. Figure 2.1 shows a plot of the resultant spatial covariates. For ease of reference, we label the covariates and refer to these labels in the rest of this document.

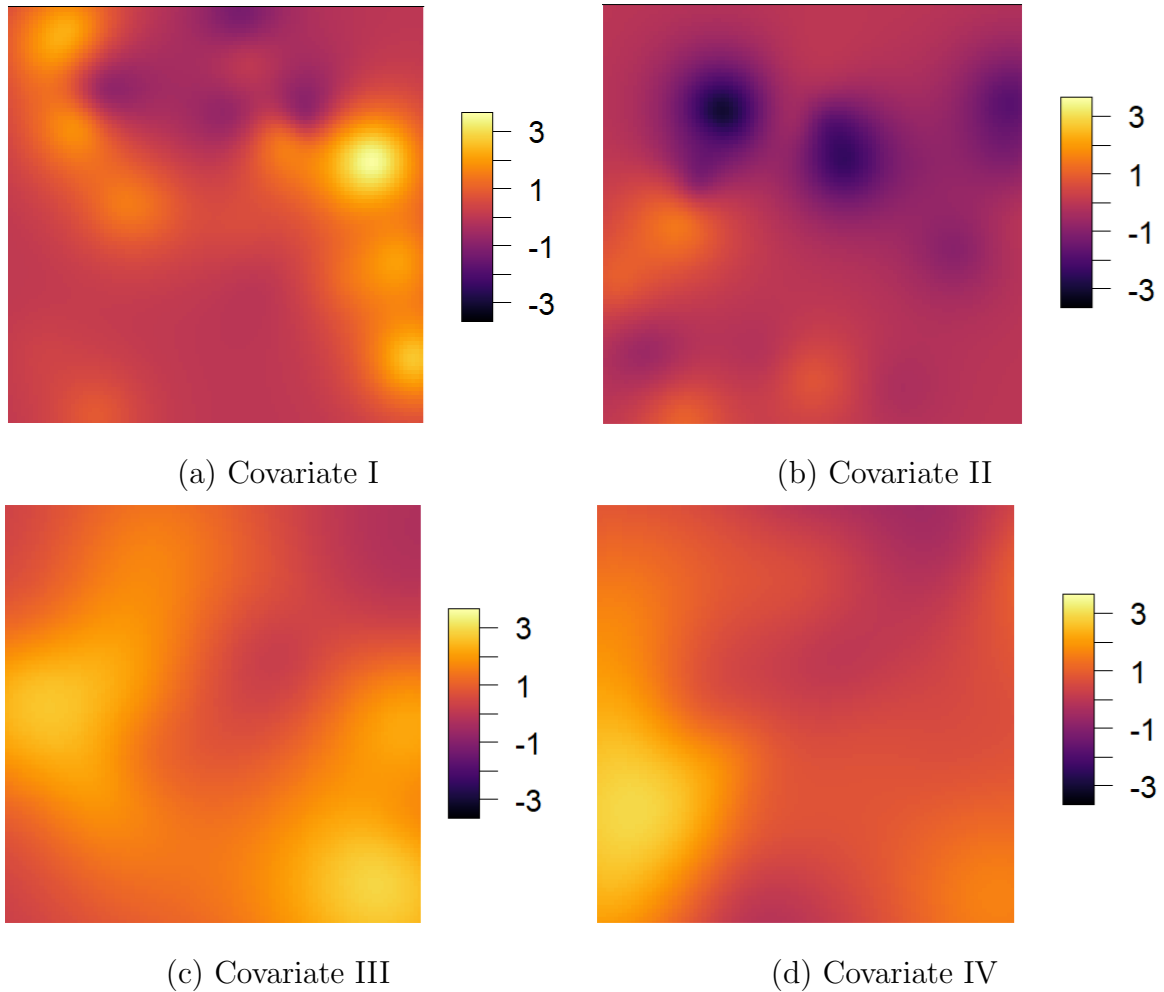


Figure 2.1: Spatial covariates simulated on a window $W^* = [0, 10] \times [0, 10]$ from a Gaussian random field with a Whittle-Matern covariance model and ranges 0.5 ((a) Covariate I and (b) Covariate II) and 1.5 ((c) Covariate III and (d) Covariate IV).

2.4.3 Using the covariate to define the window

In order to define the simulation window, W , on which the points will be generated, we use thresholds based on the covariate values to specify the locations at which points should be simulated. The following cases are considered.

- **Case 1:** Points may only occur at locations that coincide with a covariate value that is less than a chosen threshold. Figure 2.2(a) illustrates the result of applying this case to Covariate I.
- **Case 2:** Points may only occur at locations that coincide with a covariate value that is greater than a chosen threshold. Figures 2.2(b) and (c) illustrate the result of applying this case to Covariate II and III, respectively.
- **Case 3:** Points may only occur at locations that coincide with covariate values that lie between two chosen limits. Figure 2.2(d) illustrates the result of applying this case to Covariate IV.

2.4.4 Simulating the point pattern

For the point pattern simulation, we consider four different spatial point process models, namely a homogeneous Poisson point process, heterogeneous Poisson point process, a simple sequential inhibition process, and a Matérn cluster process. The simulation of point patterns from these models are performed using the functions `rpoispp`, `rSSI`, and `rMatClust` in the `spatstat` package in R. For each window, we simulate 1000 point patterns from the different models and consider small (S), medium (M), and large (L) sized point patterns with 100, 500, and 1000 points, respectively.

2.4.5 Evaluation metrics

To assess the performance of the algorithm, statistics based on the area of the set difference between the simulation window, W , and the constructed window, $\widehat{W}$, is computed for each simulated point pattern, i.e.,

$$\text{True Positive (TP)} = |W \cap \widehat{W}|,$$

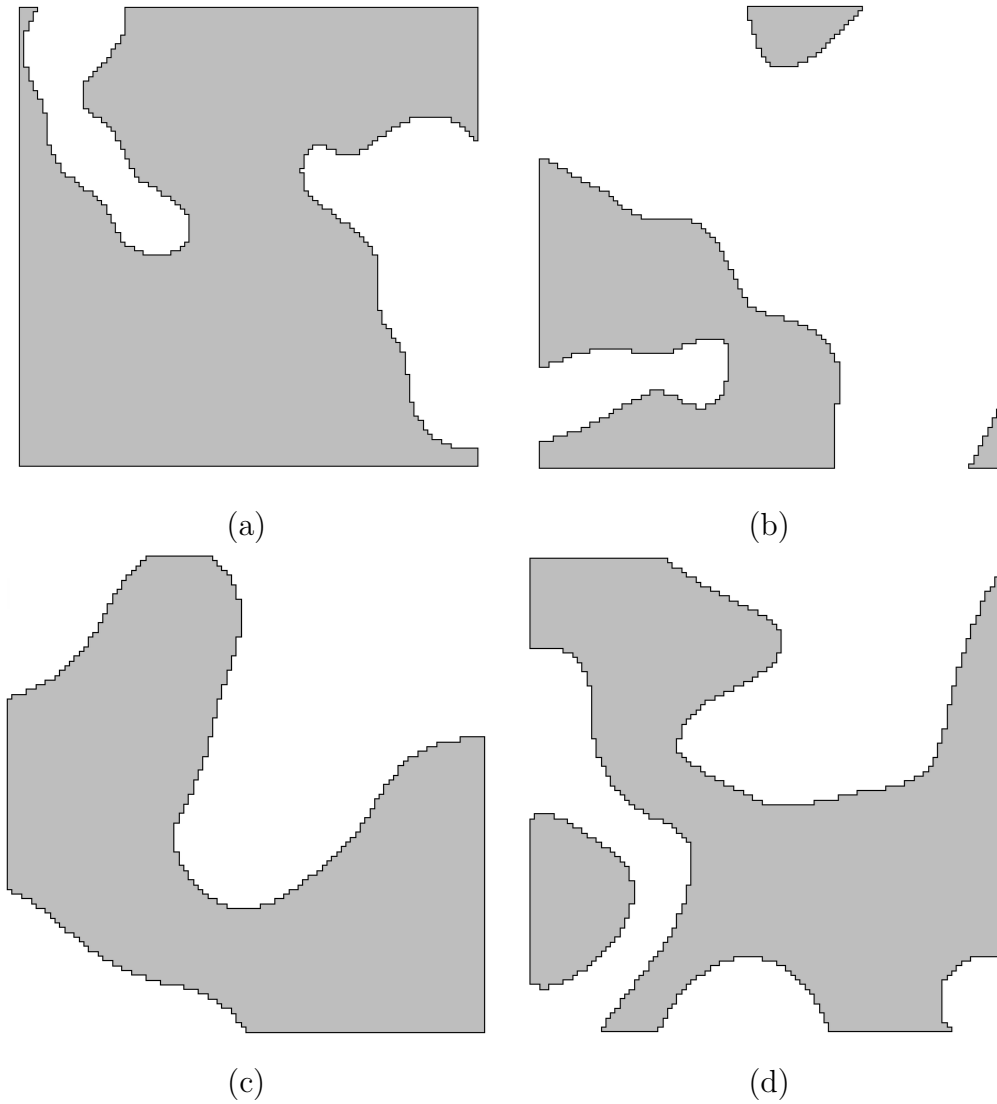


Figure 2.2: Simulation extent of points based on selecting thresholds for (a) Covariate I, (b) Covariate II, (c) Covariate III, and (d) Covariate IV.

$$\text{True Negative (TN)} = |W' \cap \widehat{W}'|,$$

$$\text{False Positive (FP)} = |W' \cap \widehat{W}|,$$

$$\text{False Negative (FN)} = |W \cap \widehat{W}'|,$$

where the symbol ' denotes the complement. These values are then used to compute the accuracy,

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN},$$

precision,

$$\text{Precision} = \frac{TP}{TP + FP},$$

recall,

$$\text{Recall} = \frac{TP}{TP + FN},$$

secificity,

$$\text{Specificity} = \frac{TN}{TN + FP},$$

F₁-score,

$$\text{F}_1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

and the intersection over union score,

$$\text{Intersection over Union (IoU)} = \frac{|W \cap \widehat{W}|}{|W \cup \widehat{W}|} = \frac{TP}{TP + FP + FN}.$$

The results are presented in Section 2.4.6.

2.4.6 Results

The results of the simulation are presented in this document in Figures 2.3 - 2.8. The corresponding Tables referred to in this section are supplied as additional supplementary material and can be accessed using the following link [here](#).

Accuracy

In Table 2.1 and Figure 2.3 a high mean accuracy for the window selection algorithm can be observed, with values ranging from 91% to 99% across various point patterns, sizes, and covariates. Accuracy increases as the size of the point patterns increases from small to large across all point pattern types and covariates. Small clustered point patterns show the poorest performance overall. The algorithm is relatively stable demonstrated through low standard deviation values between 0.01% and 4.41% for the different point patterns and covariates.

Precision

Table 2.2 and Figure 2.4 show relatively high mean precision values demonstrating the algorithms ability correctly identify regions to retain in the constructed window. A marked improvement in the algorithm's detection ability is observed here as well with increasing point pattern size. The poorest performance, here as well, is in the case of small clustered point patterns.

Recall

The algorithm has on average high recall, illustrated in Table 2.3 and Figure 2.5. Thus, in most cases, it is able to identify the majority of relevant regions to retain in the window. The variation in standard deviation of the recall values indicates that the consistency of recall may vary between different covariates and point pattern types but is more stable for point patterns that are larger in size.

Specificity

The algorithm, on average, has relatively high specificity values. This is observable in Table 2.4 and Figure 2.6. Thus, in most cases, it demonstrates a strong ability to identify the regions that must be exclude in the window domain. Visible in the figure as well is the marked improvement in the ability of the algorithm as the size increases from small to large. As, in the case of the previously discussed metrics the poorest performance is observed for small clustered point patterns across all covariates.

F₁-score

The results in Table 2.5 and Figure 2.7 show that, on average, the algorithm has high F₁-score values that are fairly close 1. This suggests the algorithm performs in identifying regions to retain in the window while also being accurate in doing so. A pattern of the improvement in performance of the algorithm is also notable here again for point patterns as they increase in size.

Intersection over Union

The summaries of the IoU score presented in Table 2.6 and Figure 2.8 indicate a high degree of overlap between the predicted and actual regions, implying that the algorithm typically performs well in localizing the extent of the window domain. Poorer performance in the case of smaller point patterns is observed especially in the case of clustered point patterns.

The window selection algorithm shows good performance across all evaluation metrics. The algorithm has strong capabilities in retaining relevant regions, with improvements observed as point pattern sizes increase. The Intersection over Union (IoU) shows large overlap between the estimated and true simulation extent, demonstrating the algorithm's ability to appropriately defining window domains. Although performance is lower for small clustered point patterns, the algorithm performs relatively well in most scenarios.

2.5 Application

We now consider an application to rural household locations in Tanzania. This data was collected in a census in the Serengeti District, Mara province, Northern Tanzania². The data comprises georeferenced locations for 35 947 households spread across 88 villages. The locations are in latitude and longitude decimal degree coordinates. Three villages are considered for this chapter, namely Iseresere, Kono and Nyamakobiti with households that number 295, 320 and 412 respectively. Figure 2.9 depicts the household locations in the three villages.

Spatial covariate data of terrain elevation for Tanzania is extracted from a Digital Elevation Model (DEM): a raster grid with each cell containing a value of the elevation (in meters) of the earth's surface above sea level. The data was collected in the Shuttle Radar Topographic Mission (SRTM). The SRTM data were sampled over a grid of 1 arc-second by 1 arc-second (approximately 30m by 30m). The left pane of Figure 2.10

²Provided by Katie Hampson, <http://www.gla.ac.uk/researchinstitutes/bahcm/staff/katiehampson/>, <http://www.katiehampson.com/#intro>, and approved for use by the Faculty of Natural and Agricultural Science Research Ethics committee under the reference NAS33/2019.

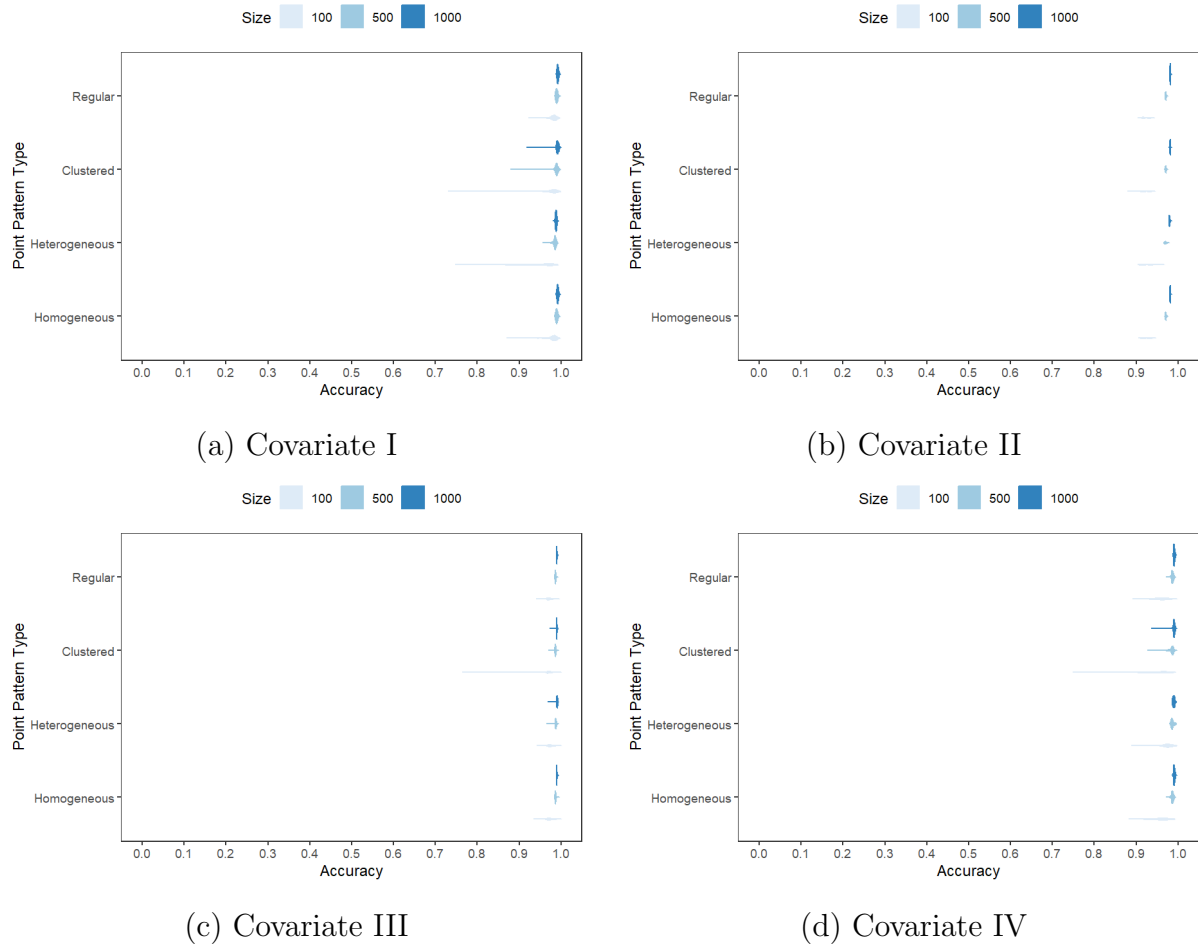


Figure 2.3: Violin plots of the accuracy of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

shows the terrain slope of the elevation data from the DEM.

Human settlements in a rural setting often rely on some attractive and restrictive traits of the natural landscape. Topographic properties of terrain influence the distribution of environmental phenomena and the nature of environmental processes. In rural areas, steep slopes present many challenges which make it impracticable for building houses. Steep slopes have greater requirements in terms of structural planning and costs. Flat land is easier and cheaper to build on since less time and expenses are incurred in getting the land suitable for building. Owing to these reasons, terrain slope is

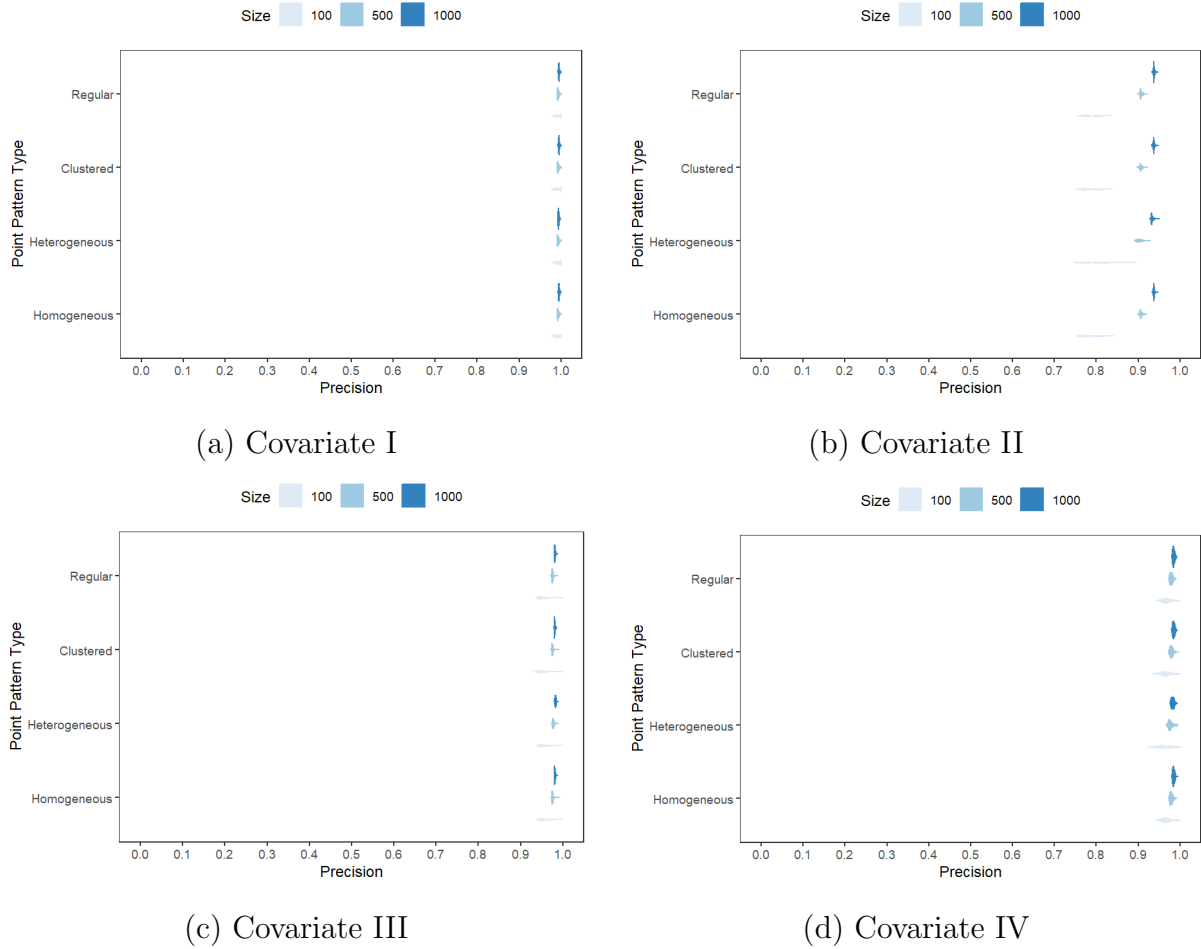


Figure 2.4: Violin plots of the precision of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

useful for describing terrain viable for (new) house locations and is used here as a spatial covariate to characterise land that is suitable for building.

The algorithm is applied to each of the villages and the results presented in the right pane of Figure 2.10. Opening and closing (mathematical morphological operations) are applied to the resultant windows to remove noise and smooth the window edges. For each village, the algorithm detects and removes regions with high terrain slope. Regions with terrain slope values close to covariate values at observed point locations have been identified by the algorithm as low terrain slope areas and are retained in the final window.

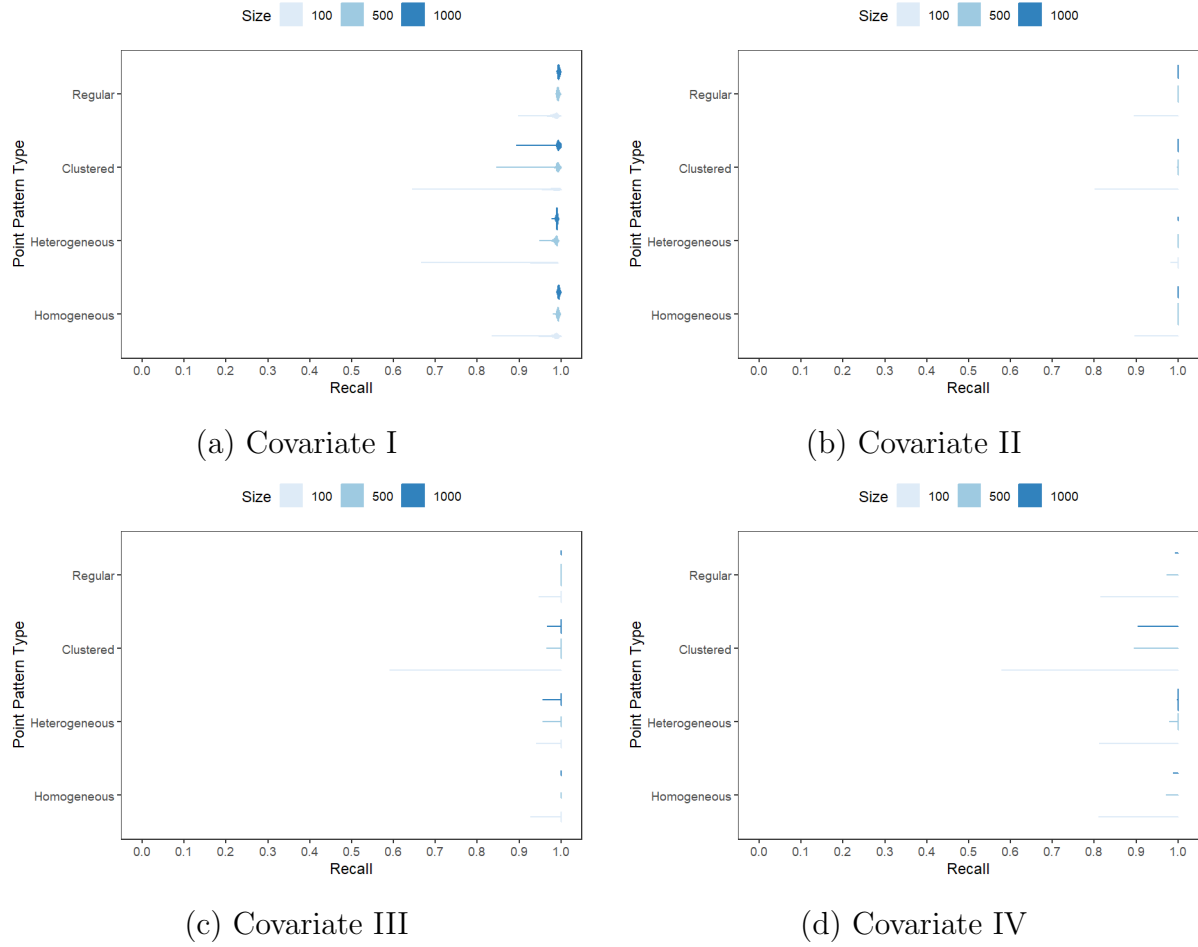


Figure 2.5: Violin plots of the recall of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

Even though there are no realisations of the point process in certain areas, the algorithm accounts for the possibility of a point occurring there as long as it satisfies the definition of viable land that is a function of covariate values at observed locations in the pattern.

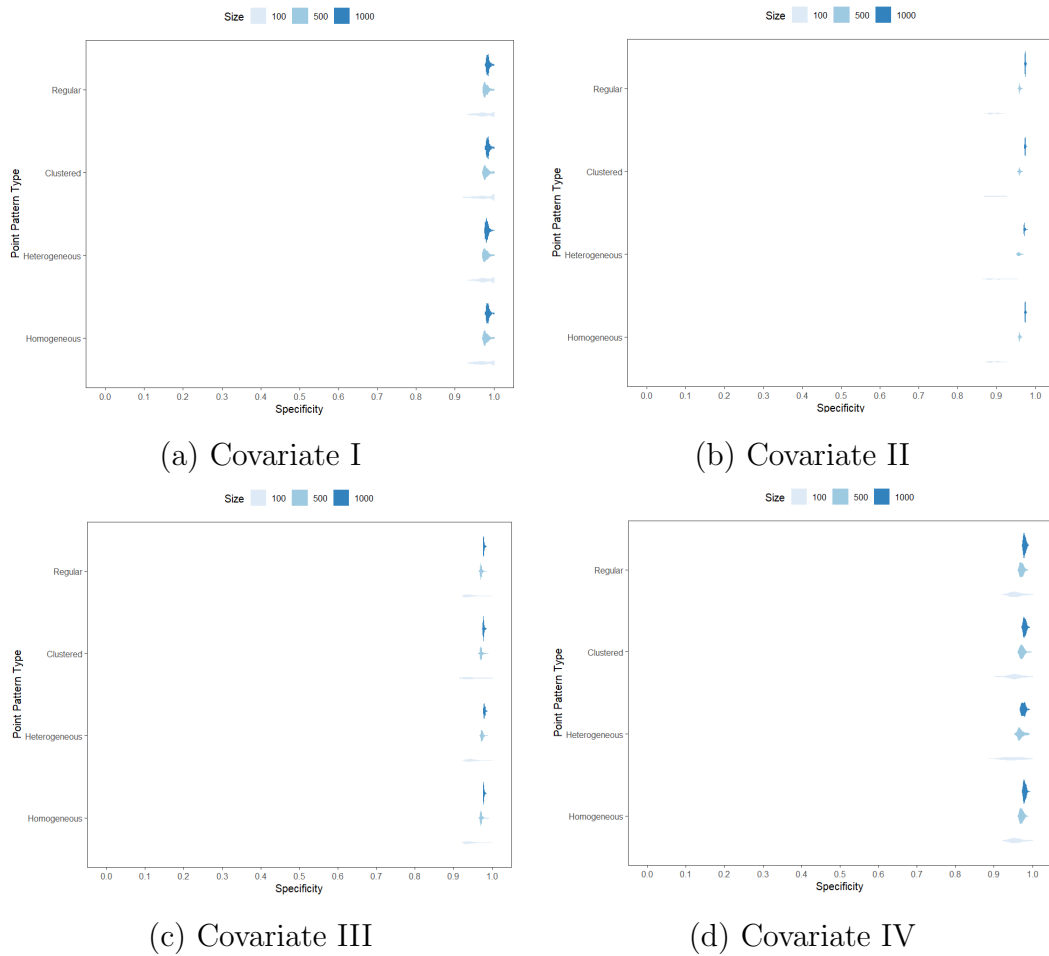


Figure 2.6: Violin plots of the specificity of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

2.5.1 Landscape and terrain features

Consider household locations in Iserere. Figure 2.11³ depicts a Google Earth satellite image of Iseresere. The geographical locations of households in this village are denoted by red dots.

³Google 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]

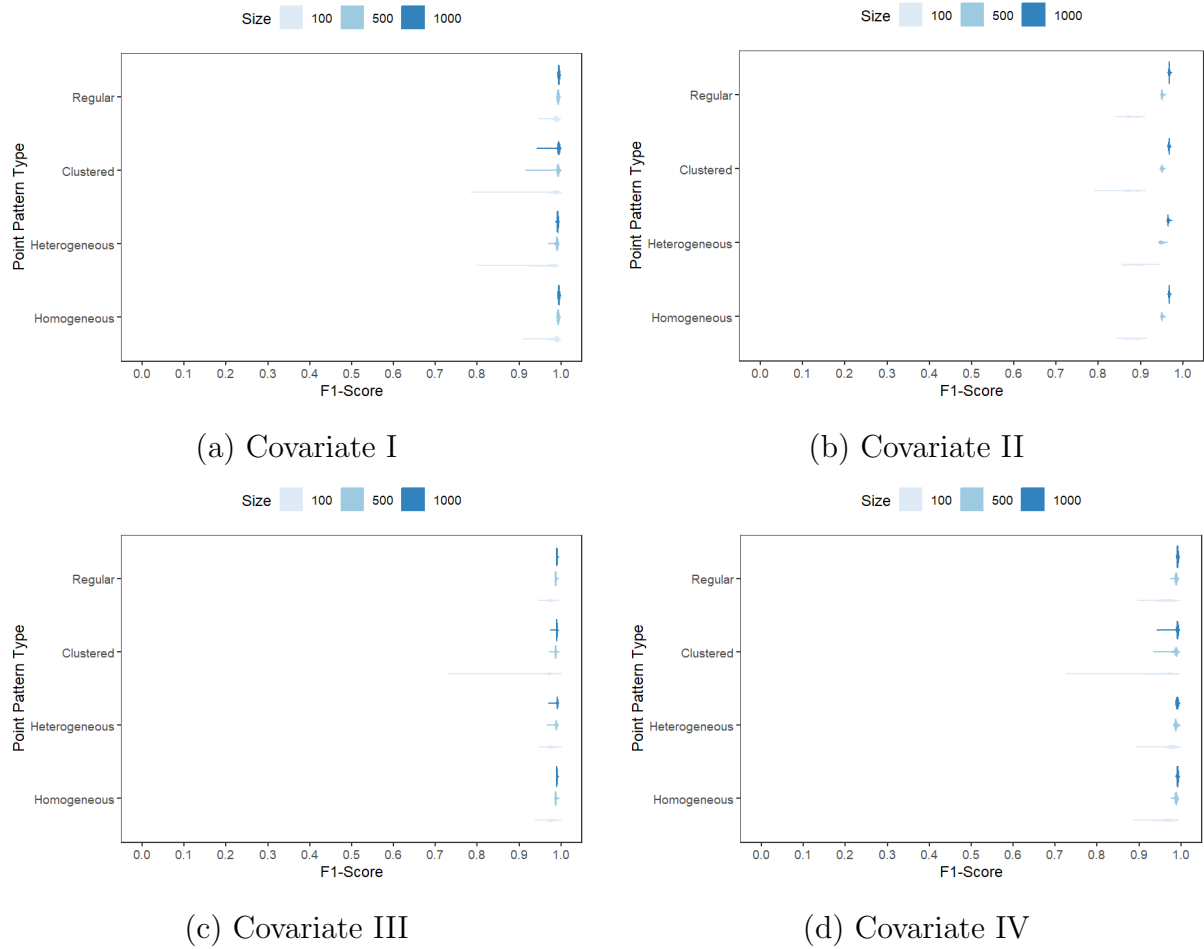


Figure 2.7: Violin plots of the F1-score of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

Figures 2.12⁴, 2.13⁵ and 2.14⁶ show enlarged images of different sites in the village.

⁴Google 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]

⁵Google 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

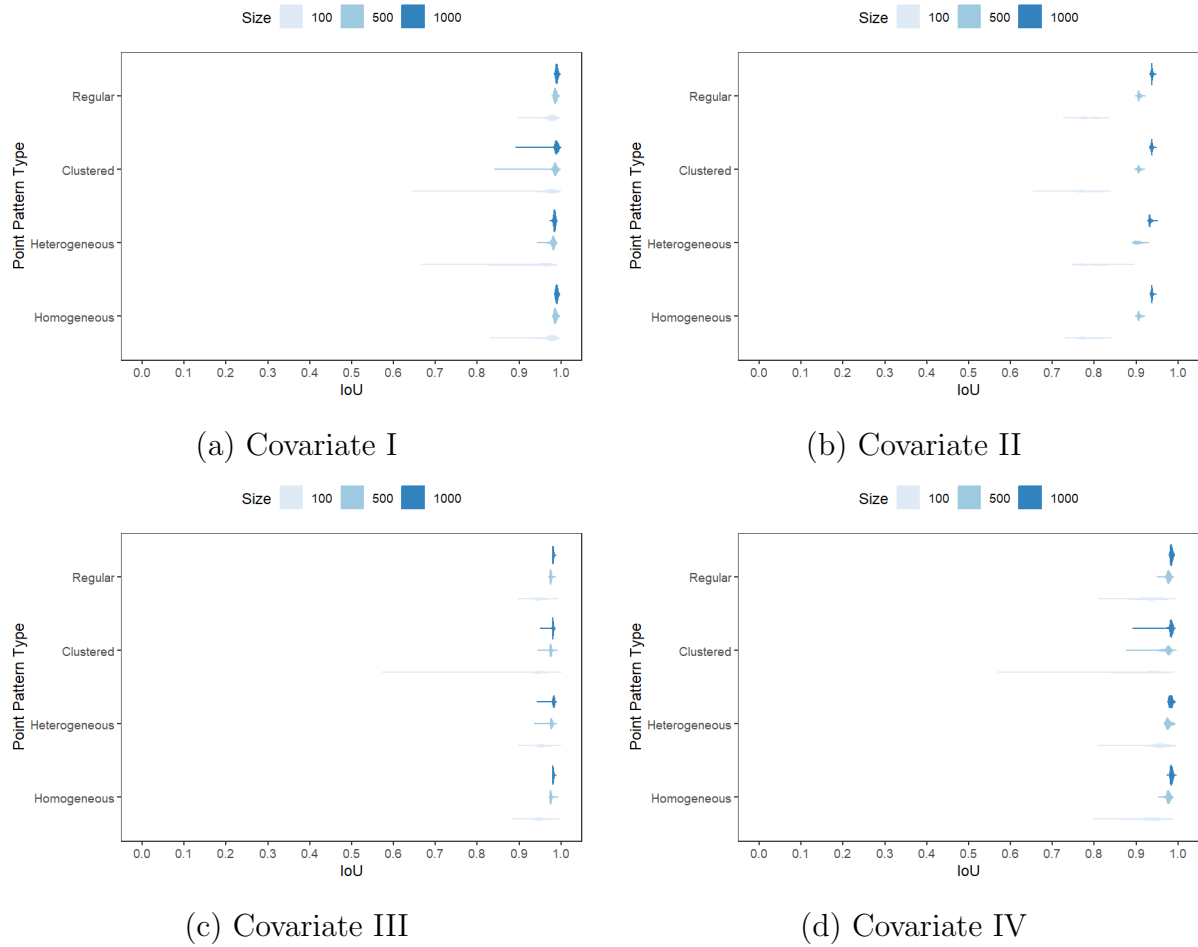


Figure 2.8: Violin plots of the intersection over union of the window selection algorithm for point patterns of different types and sizes across Covariates I - IV.

earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. $1^{\circ}35'10.67''\text{S}$, $34^{\circ}22'39.62''\text{E}$. 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. $1^{\circ}34'31.19''\text{S}$, $34^{\circ}22'12.20''\text{E}$. Eye alt 4.43km. Maxar Technologies 2020. <http://www.earth.google.com>[January 20, 2020]

⁶Google earth V 7.3.2.5776. (July 13, 2018). Mara province, Tanzania. $1^{\circ}35'24.53''\text{S}$, $34^{\circ}21'40.57''\text{E}$. 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. $1^{\circ}33'30.73''\text{S}$, $34^{\circ}22'03.92''\text{E}$. 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. $1^{\circ}35'35.08''\text{S}$, $34^{\circ}21'59.23''\text{E}$. Eye alt 3.46km.

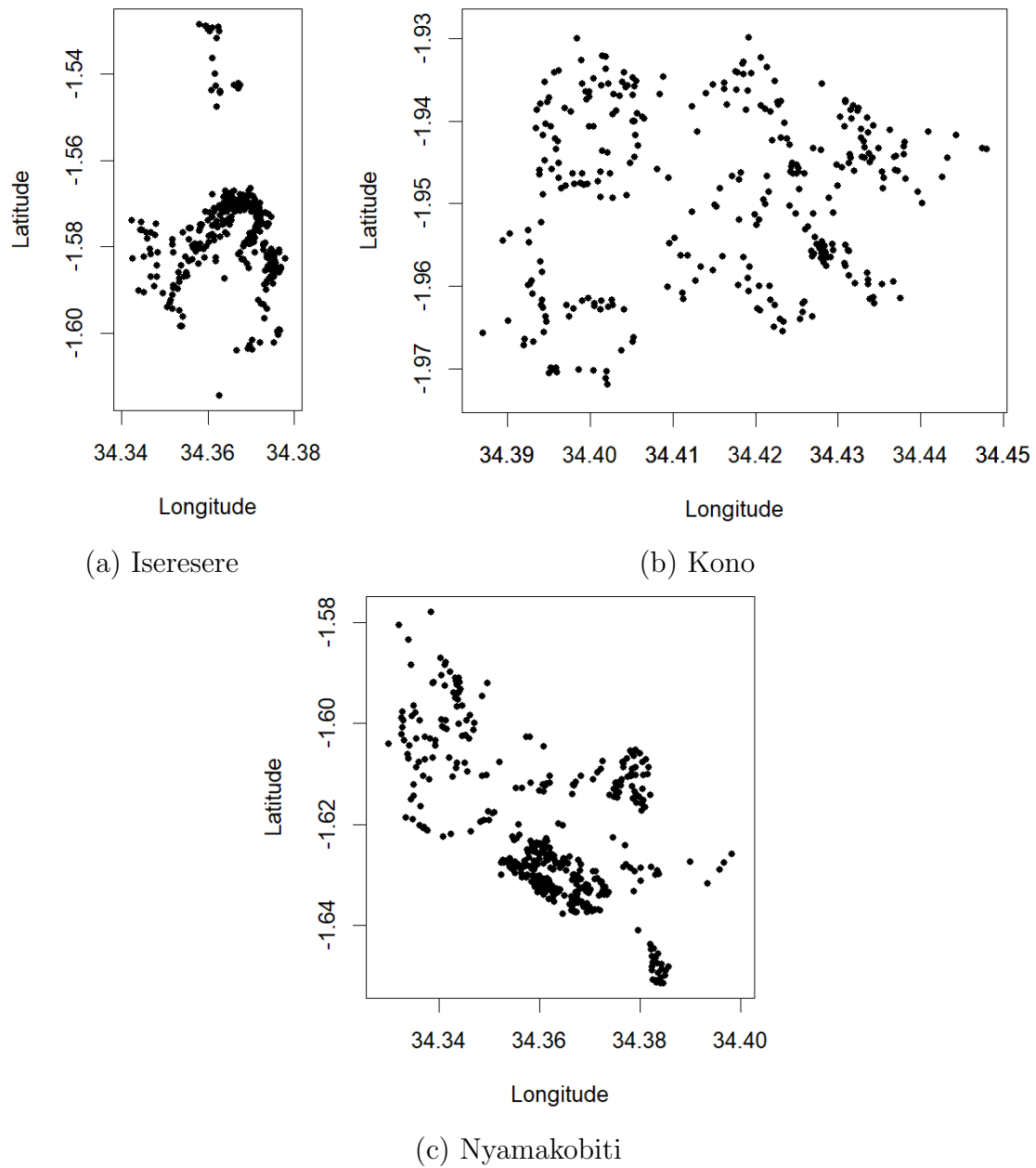
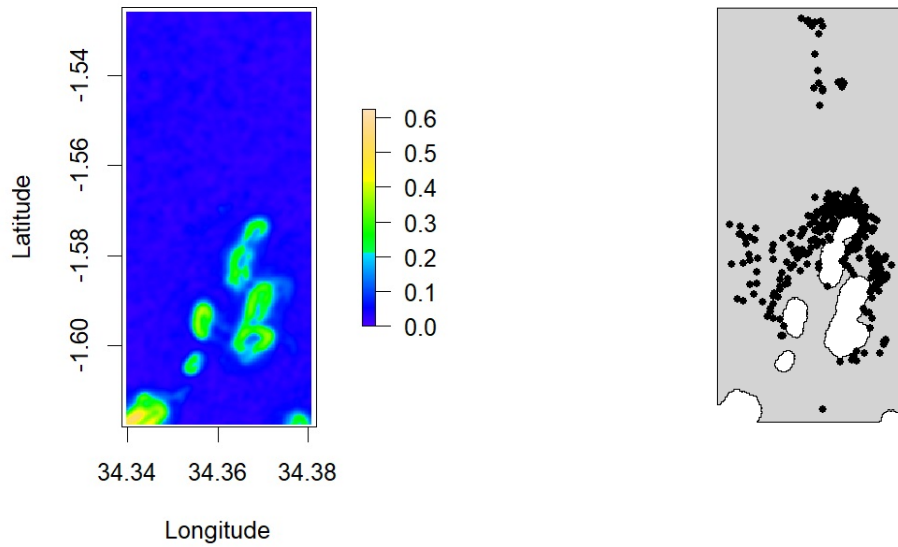


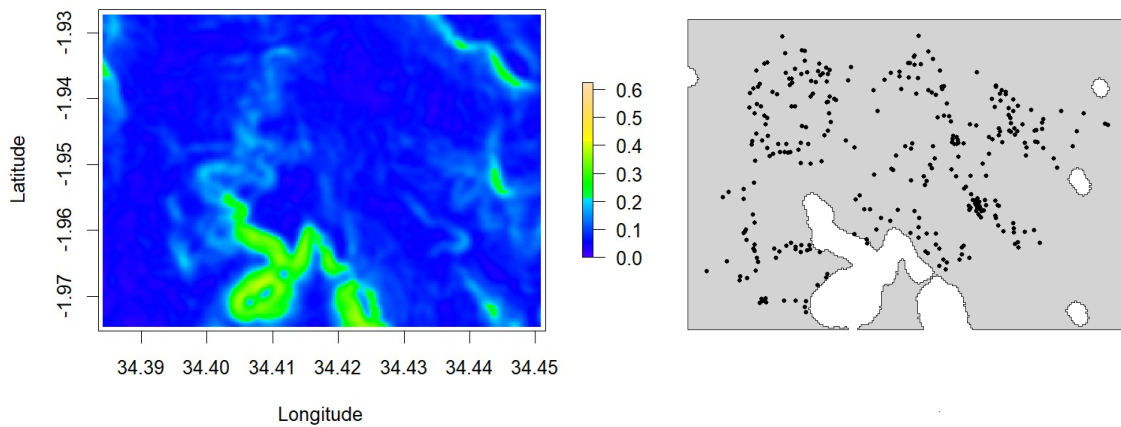
Figure 2.9: Household locations for 3 villages in Tanzania's Mara province.

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 in the

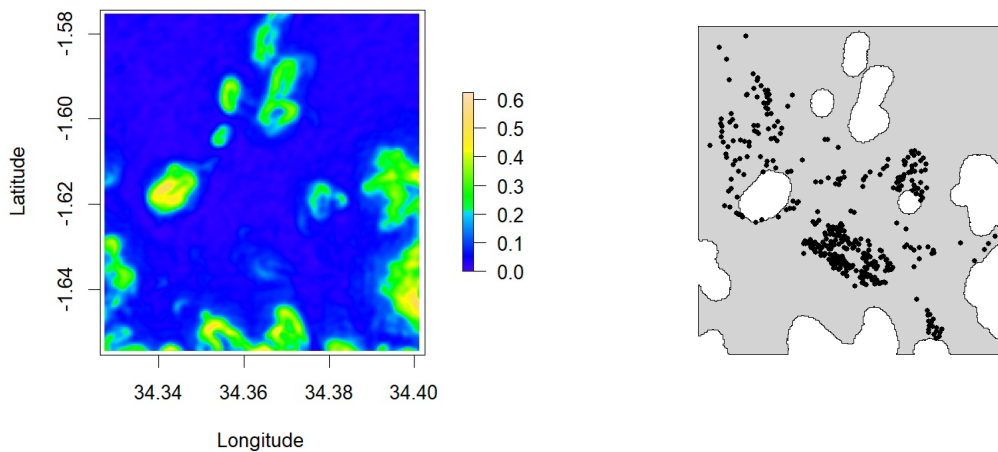
Maxar Tecnhonogies 2020. <http://www.earth.google.com>[January 20, 2020]



(a) Iseresere



(b) Kono



(c) Nyamakobiti

Figure 2.10: Constructed window (right) and terrain slope (left) for villages in Tanzania's Mara province.

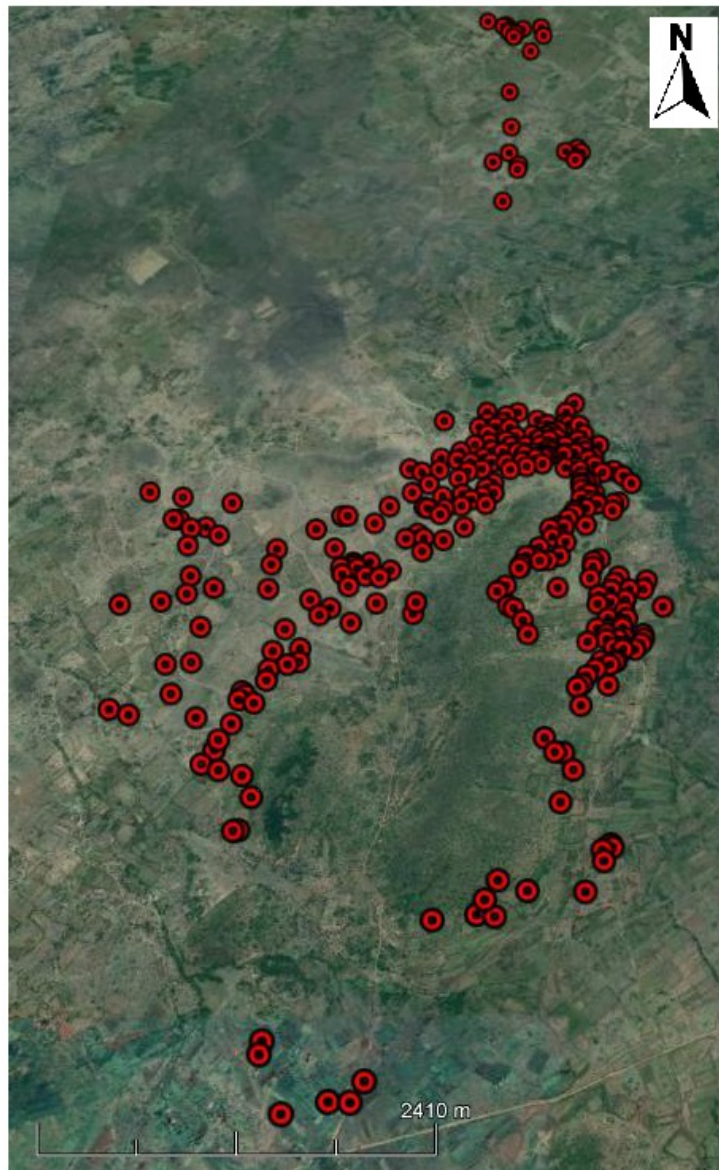


Figure 2.11: Google Earth satellite image of Iseresere village in Tanzania's Mara province, with household locations indicated with red dots.

figures. We observe that households are spread along the edge of terrain with high relief. The occurrence of households 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). Further inspection shows 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.

The algorithm shows a strong ability to identify terrain that are suitable for household locations to occur on. Thus the algorithm has admitted a data-driven domain selection approach that aligns relatively well with the underlying processes generating the point pattern.

2.6 Discussion

Point pattern analysis requires knowledge of the window in which locations are observed, as it is directly involved in inferring the properties of the underlying point process [9]. The window is often predefined before analysis. When unknown, it is typically approximated using the smallest bounding rectangle or convex hull of the data locations. Choosing an overly large window, however, can distort estimates of the point pattern properties. For instance, point density may be underestimated because the total area of the window is artificially inflated by empty spaces where points are unlikely to occur. Also, empty regions in a convex window may lead to misleading conclusions about clustering.

Several estimation techniques to infer the window based on observed data locations exist but are typically derived by assuming that the point pattern is realised from a homogeneous Poisson process and that the unknown domain is convex [60, 58]. Other solutions exist that do not require a convexity assumption, but still impose an assumption about the underlying point process model such as in [57]. The assumptions used to derive the results may not necessarily hold in some spatial applications. Additionally, these methods do not incorporate the effect of covariates that may influence where points on the domain can be observed. In real world applications, the distribution of points may be constrained by some underlying process, expressed as a covariate, resulting in more complex spatial windows.

When covariate information is available, the dependence of the point pattern on the covariate should be investigated. Parametric models that incorporate this dependence

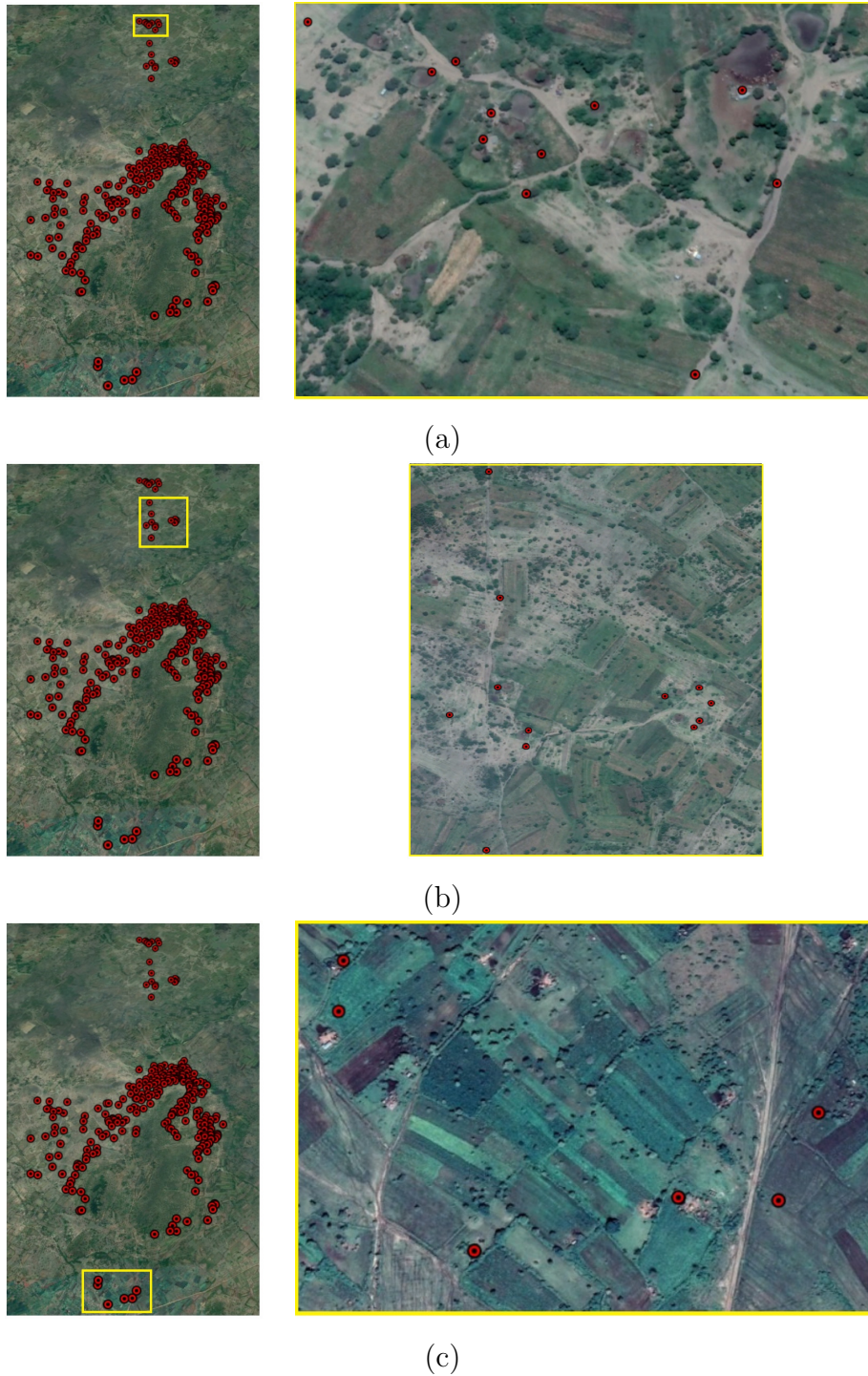


Figure 2.12: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province.

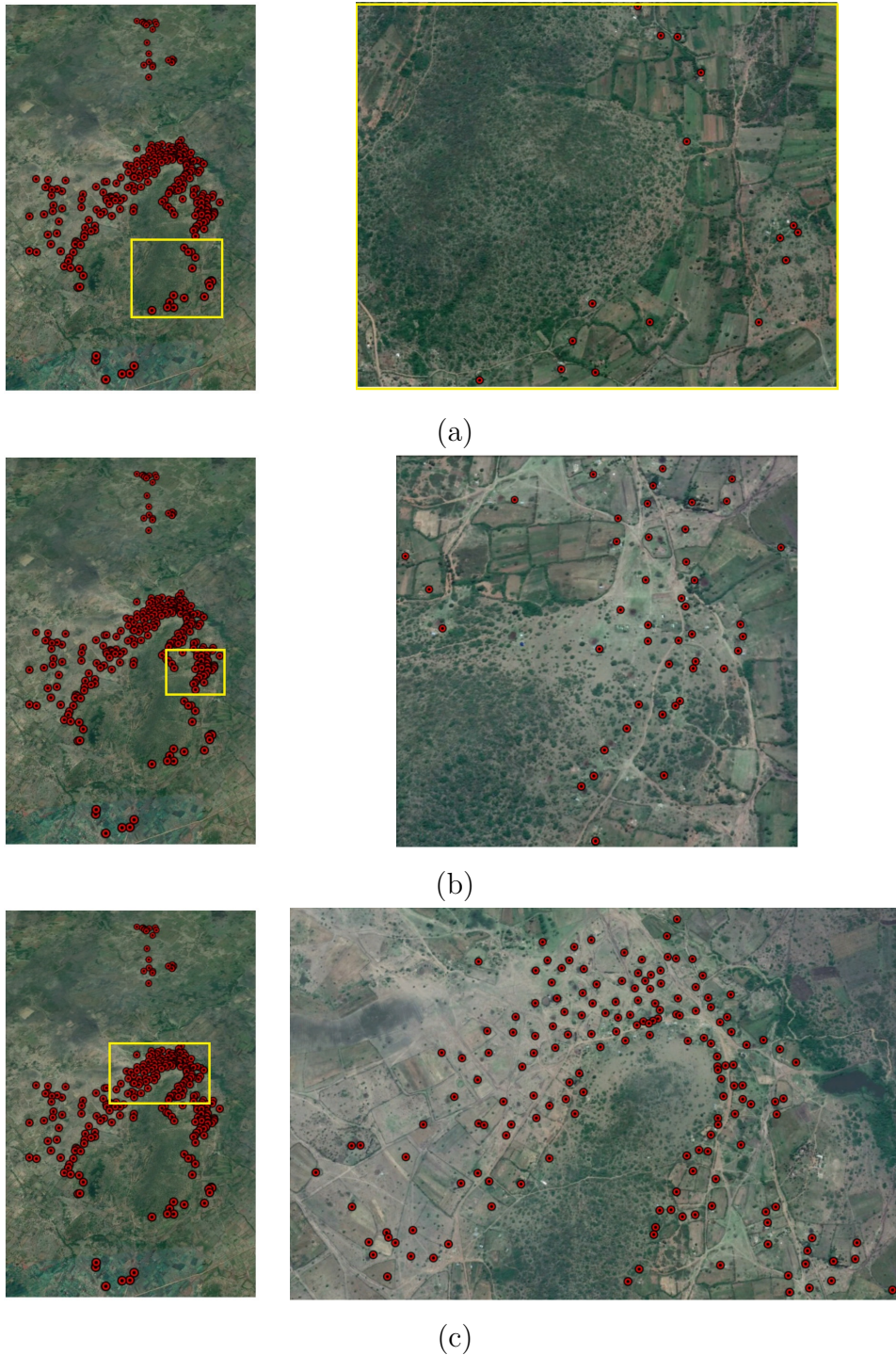


Figure 2.13: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).

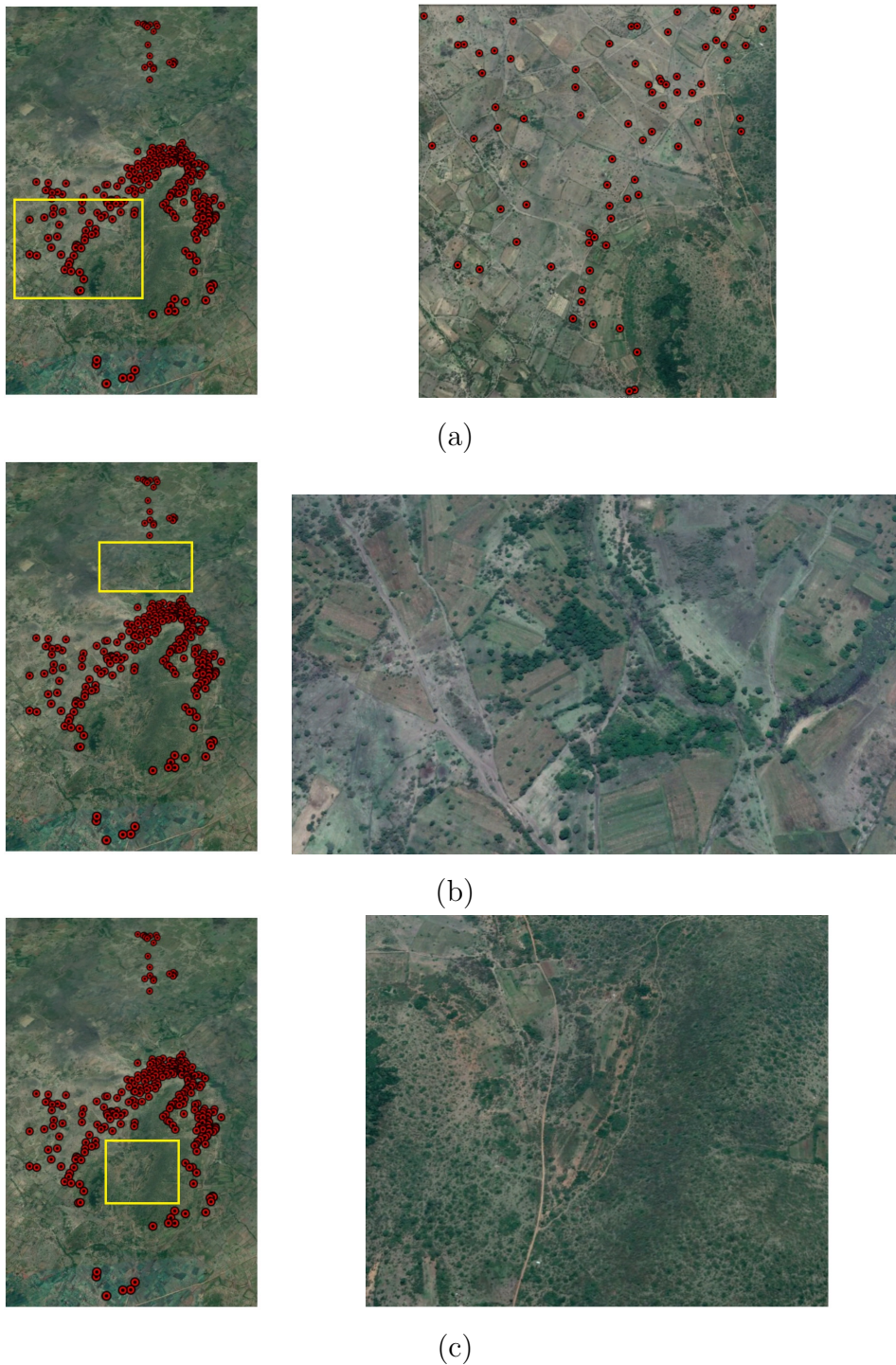


Figure 2.14: Google Earth satellite image (left) and enlarged image (right) of different sites in Iseresere village in Tanzania's Mara province (continued).

and formal hypothesis testing procedures, under parametric assumptions, are well developed. Nonparametric methods have received some attention, albeit minimal, which include extending the kernel smoothed intensity estimate to allow for covariate affects. A treatment of these methods can be viewed in [6, 9]. While these methods quantify the relationship between points and the covariate, they still rely on an incorrectly specified spatial window.

In this research, a method is proposed to estimate the window domain of point patterns using covariate data. The technique does not rely on an assumption of convexity or the underlying structure of the point process from which the point pattern is realised. The investigation of the dependence of the point pattern on the spatial covariate should precede the implementation of this algorithm. It is desirable to use covariates believed to impact the distribution and abundance of the object of interest, or that is correlated to them, since the proposed method makes use of the notion that the covariate values at data points are a random subset of potential data values within the point pattern. When the placement of points is governed by covariates, the covariate distribution at observed locations will deviate significantly from that at randomly chosen points that are realised from a process not tied to the covariate. Thus, the covariate values at observed locations provide a reference data set from which "outlying regions" in the window can be determined.

The algorithm proposed in this work makes use of a distance-based outlier scoring. Essentially, the sparsity of a data point's local neighbourhood is used to indicate its outlier status. An outlier is thus surrounded by relatively few other data points within a defined distance. The performance of the outlier-detection algorithm is influenced by certain covariate distributions, e.g., in data sets with multiple clusters or subgroups, points located between clusters may be incorrectly flagged as outliers, even though they are normal relative to their cluster. This may warrant the use of an initial clustering to identify clusters and a subsequent application of the distance-based outlier detection within each cluster. In future work, the extent to which the distribution of the covariates influence on the performance of the window selection algorithm could be investigated.

In a the simulation study, the algorithm performs well for most point process models with the exception of small clustered point patterns. In the case of the latter, the spatial

arrangement of points can vary significantly between different realisations, which leads to large variability in the performance of the window estimation technique. A marked improvement in performance, however, can be observed as the size of the point pattern increases.

An application to rural household locations is considered with the use of elevation data from a DEM. The algorithm performed 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 unlikely. In other cases it may be appropriate to include other covariates that characterise a feature of the landscape that is unsuitable for building new houses, such as areas with rivers, dams or marshes. In future work the algorithm could be extended to allow for an ensemble of spatial covariate effects.

2.7 Conclusion

Herein, we presented a new algorithm for selecting the spatial point pattern domain without the restriction of convexity. The algorithm works by using a reweighted repeatedly sampled nearest neighbour distance outlier score and Otsu's method for automatic threshold selection on covariates at observed point locations. Using a function or feature of the spatial covariate in regions at observed points in the pattern, the proposed algorithm constructs a nonconvex window. The robustness of the algorithm to various point pattern types and covariates was tested via a simulation study. The algorithm shows good performance in most settings with the poorest performance observed for small clustered point patterns. An application to rural village households in Tanzania's Mara province was considered. Remotely sensed data from a DEM is used as a covariate in this case. When the movement between points is constrained to a nonconvex window, the Euclidean distance will not give a measure representative of the path between points. Consequently, the Euclidean shortest path distance on the nonconvex window is a measure more suited for specifying the distance between points represented by the surface than other measures and thus should be used in any further spatial analysis. In future work the algorithm could be extended to allow for an ensemble of spatial covariate effects. One could also investigate other automatic threshold selection techniques.

Chapter 3

Kernel-smoothing of point patterns constrained on nonconvex domains

In many application fields of point pattern analysis, the aim is to determine a smooth-varying function of intensity. This chapter is based, in part, on our published work in [36]. In the chapter, the proposed modified kernel-smoothed intensity estimator was used as a validation technique for the window selection algorithm proposed therein. Herein we extend the original discussion and include a comparison of the proposed estimator to its only other known competitor, the diffusion smoother of [7].

3.1 Abstract

In traditional kernel smoothing of spatial point patterns, a point's contribution to the intensity estimate is determined by the Euclidean distance. Using Euclidean distance to quantify proximity between points constrained on nonconvex spatial domains, however, can introduce artefacts in the intensity estimate. Herein, a graph-theoretic approach is used through the application of a visibility graph to determine distance on a connected space in nonconvex spatial domains. Using the visibility graph, we develop a kernel-smoothed intensity estimator that allows for path-based estimation that respects obstacles in the domain and the domain boundaries. The proposed method offers a versatile tool for applications in ecology, urban planning, and spatial epidemiology where

such spatial constraints are common.

Keywords: nonconvex spatial domain, kernel smoothing, Euclidean distance, point pattern, visibility graph

3.2 Introduction

In an analysis of spatial point pattern data, kernel smoothing [18, 9] is typically performed to estimate a spatially-varying density or intensity of points. The kernel smoothed intensity estimate assigns kernel weights as a function of the Euclidean distance between points. While the Euclidean distance is often used as a metric to quantify proximity between points in a point pattern, its application on nonconvex windows leads to misleading results that incorrectly represent the spatial constraints imposed by the window geometry [36].

Euclidean distance measures the straight-line distance between two points, assuming a direct, unobstructed path [34]. For convex windows, the Euclidean distance is appropriate since the path between all pairs of points in this domain are unobstructed. In nonconvex¹ windows, however, the shortest path between points may be constrained by holes in the window or the window boundaries. Thus, the actual path must circumvent these obstacles.

In ecological studies, for example, animal movements are restricted by landscape features (e.g., rivers, mountains). Thus, the Euclidean distance does not capture the cost or difficulty of traversal, potentially leading to incorrect conclusions about habitat use or species interactions [49, 70]. The spatial structure of habitats play an important role in shaping ecological processes. It influences population dynamics, species interactions, and organism movement. Fragmented habitats can significantly affect how populations are distributed and the pathways available for species migration or dispersal [22].

The effect of modelling non-Euclidean movement in highly structured ecological networks using the Euclidean distance has also been investigated in [70]. In the study, it was demonstrated that using ecological distances² produced unbiased estimates of

¹A polygon P is considered nonconvex if $\exists x, y \in P$ such that $\alpha x + (1 - \alpha)y \notin P, \forall \alpha \in (0, 1)$.

²distances that minimised the cost of traversal on a discretised space where movement is allowed

abundance³, and better characterised the true home range geometry. In contrast, the Euclidean distance produced biased abundance models.

Other empirical studies have demonstrated the advantages of using alternative distance measures over Euclidean distance in nonconvex domains. For example, in epidemiological research, accounting for the actual paths of human movement within urban environments has improved the understanding of disease spread and transmission dynamics [21, 56].

To address the shortcomings of the Euclidean distance when used to quantify proximity between points restricted to a path connected space, one approach used is the length of the shortest path between points on a network or graph [44, 78, 71]. This metric provides a more accurate representation of the spatial configuration in complex environments where movement is inhibited by physical barriers. Despite the need for the appropriate characterisation of the spatial structure of planar point patterns in nonconvex windows, methodology for analysis of point pattern data in such domains is not well developed.

Herein, we apply a graph theoretic approach through the use of a weighted visibility graph to develop a kernel-smoothed intensity estimate that allows for path-based estimation that respects obstacles in the domain and the domain boundaries. To this end, the remainder of this document has been organised as follows. Section 3.3 provides some definitions and a presentation of the proposed adapted kernel-smoothed intensity estimate. An application to rural village household locations in Tanzania is given in Section 3.4 and concluding discussions in Sections 3.5 and 3.6.

3.3 Methodology

In Sections 3.3.1 and 3.3.2, formal definitions are provided for point pattern data sets and their first-order intensities, respectively. The weighted visibility graph is defined in Section 3.3.3 and its use to develop a path-based kernel-smoothed intensity estimate in Section 3.3.4.

between connected adjacent cells

³the total number of individuals of a species or type in a specific area, ecosystem, or habitat

3.3.1 Spatial point pattern data

A point pattern [9, 13, 30] data set comprises a set of locations, $\{x_1, \dots, x_n\}$, observed in a spatial domain $W \subset \mathbb{R}^2$, where $n \geq 0$ is not predetermined. A point process, $\{X_1, \dots, X_n\}$, is a stochastic mechanism from which the point pattern is realised. Understanding the characteristics of a point pattern is typically the aim of a spatial analysis. An analysis may involve assessing the departure of a point point pattern from stationarity and isotropy. A point process is considered stationary if it is invariant under translations and isotropic if it is invariant under rotations.

Estimating the properties of a point pattern typically involves the use of distance-based statistics. These statistics inherently depend on the definition of the window; it plays a role when calculating global density metrics, adjusting for boundary effects in intensity estimation [9, 18], and measuring distances to assess spatial dependence and correlation [9]. Thus it is important to use an appropriate measure of proximity that captures the structure of the window.

3.3.2 First-order intensity

The intensity function is an important measure in point pattern analysis [32, 39]. It characterises the first-order properties of a point process, describing how the average number of points varies over different locations in space [14, 18, 23].

The first-order intensity [14, 18] at a location $u \in W$ is defined as

$$\lambda(u) = \lim_{|du| \rightarrow 0} \frac{E[N(du)]}{|du|},$$

where du is a small region around u , $E[\cdot]$ is the expected value operator, $|du|$ is the area of du , and $N(du)$ is the number of points in the region du . For a stationary isotropic point process, the intensity function is a constant λ that does not vary with space.

The intensity function is conventionally estimated using kernel smoothing via the estimator [18, 65]

$$\hat{\lambda}(u) = \sum_{i=1}^n \kappa_{\sigma}(\|u - x_i\|), \quad (3.1)$$

where $\|\cdot\|$ is the Euclidean distance, σ is the bandwidth, $\kappa_{\sigma}(u) = \sigma^{-2}\kappa(\frac{u}{\sigma})$, and $\kappa(\cdot)$ is a kernel weighting function, usually a probability density on $\mathbb{R}^2$. Due to edge effects,

the estimator is biased. This bias can be corrected for by using the uniformly corrected kernel estimator [18]

$$\hat{\lambda}^{(U)}(u) = \frac{1}{e_{\sigma}(u)} \sum_{i=1}^n \kappa_{\sigma}(\|u - x_i\|), \quad (3.2)$$

or the Jones-Diggle corrected kernel estimator [31]

$$\hat{\lambda}^{(J)}(u) = \sum_{i=1}^n \frac{1}{e_{\sigma}(x_i)} \kappa_{\sigma}(\|u - x_i\|), \quad (3.3)$$

where $e_{\sigma}(u) = \int_W \kappa_{\sigma}(u - y)dy$ and $e_{\sigma}(x_i) = \int_W \kappa_{\sigma}(x_i - y)dy$ are bias correction terms.

The Euclidean distance, used by the estimators in Equations 3.1 to 3.3, does not provide a representative measure of proximity between points restricted on nonconvex windows. The result of disregarding the nonconvex structure of the window results in artefacts and spurious estimates. This is illustrated in Figure 3.1. In the figure, the contribution of a single point to the intensity estimate, as a function of the Euclidean distance, is shown. It can be observed that the contribution of this point to the intensity estimate is diffused radially from its center regulated by the linear distance. Thus, the contribution to the intensity estimate is high even when the path between points are constrained by the window boundary. Owing to this, an alternative distance metric is required. To find a representative distance measure on a path connected space, we apply the notion of visibility through the use of a weighted visibility graph.

3.3.3 Weighted visibility graph

A weighted visibility graph [25] of a polygon P , denoted by $G(P)$, is an ordered triple (V, E, w) comprising

- $V = \{v_1, \dots, v_m\}$, a set of the vertices⁴ of P ,
- $E = \{(v_i, v_j) | v_i, v_j \in V \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, m\}$, a set of edges between vertices that are intervisible (in a geometric context, two vertices v_i and v_j , $v_i \neq v_j$, are visible to each other if the line segment connecting them does not intersect the

⁴In the case of a polygon with holes, the vertices of the hole boundaries are included in the visibility graph vertex set.

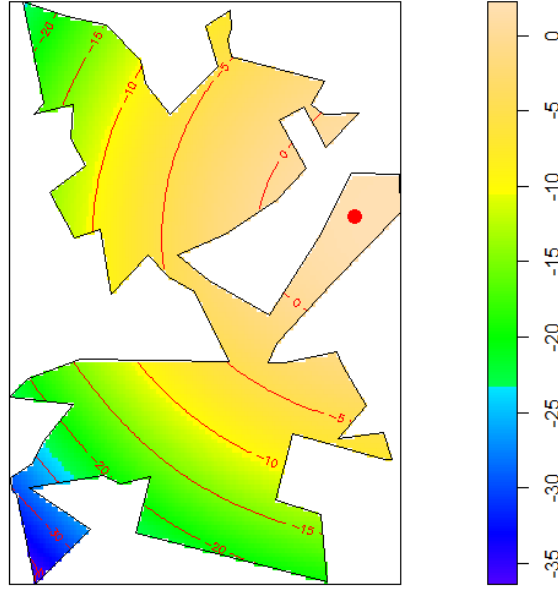


Figure 3.1: Illustration of the effect of using the Euclidean distance when estimating the kernel smoothed intensity function on a nonconvex window, shown for a single point event.

outer boundary of P or hole boundaries, that is $\alpha v_i + (1 - \alpha)v_j \in P$, $\forall \alpha \in (0, 1)$), and

- $w : E \mapsto \mathbb{R}^+$, a function mapping edges to the value of the Euclidean distance between the vertices it connects.

A representative distance measure between two points in a point pattern, one that quantifies the length of a path that does not cross the window's boundary and holes in its domain, can be found by determining the shortest path on the weighted visibility graph of the window.

A path [25, 33] from x to y is a finite sequence of vertices $v_0 = x, v_1, \dots, v_{k-1}, v_k = y$; the path begins at the point x and proceeds from vertex to vertex in the graph until it reaches y . The sum of all edge weights along the shortest path,

$$d(x, y) = \sum_{h=0}^{k-1} w(v_h, v_{h+1}), \quad (3.4)$$

provides a measure of distance within the spatial domain that avoids obstacles, such as

boundaries and holes. Figure 3.2 shows the shortest path (red line) between two points (green dots) along the visibility graph (blue lines) of a nonconvex polygon.

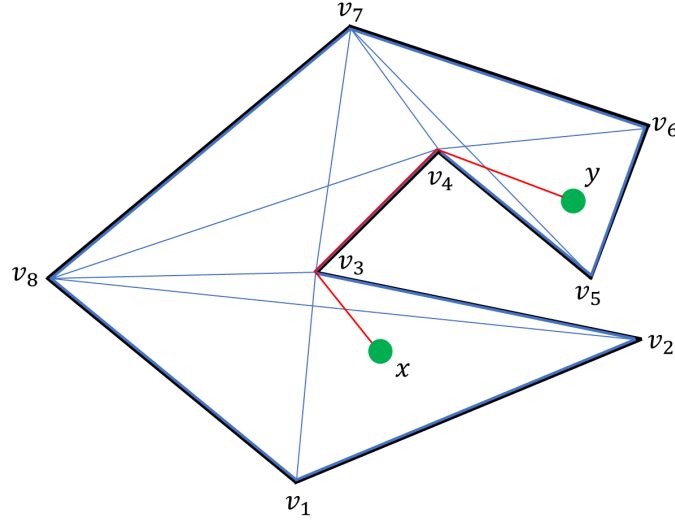


Figure 3.2: Shortest path (red line) between two points (green dots) and visibility graph (blue lines) on a nonconvex polygons.

To incorporate visibility into point pattern analysis, we use the length of the shortest path traversed between two points in the window along the visibility graph. This distance is also known as the Euclidean shortest path (ESP) distance. This formulation allows the use of a path-dependent distance that accounts for the geometric structure of nonconvex windows that constrain the movement of points.

3.3.4 Modified kernel-smoothed intensity estimate

For a point pattern $\{x_1, \dots, x_n\}$ in $W \subset \mathbb{R}^2$ we define the modified kernel intensity estimator based on the visibility graph of W , $G(W)$, as

$$\hat{\lambda}^{(V)}(u) = \sum_{i=1}^n \kappa_{\sigma}(d(u, x_i)), \quad (3.5)$$

where $d(u, x_i)$ is the sum of all edge weights along the shortest path on $G(W)$ between u and x_i , i.e., the Euclidean shortest path distance, σ is the bandwidth, $\kappa_{\sigma}(u) = \sigma^{-2} \kappa(\frac{u}{\sigma})$,

and $\kappa(\cdot)$ is a kernel weighting function on $\mathbb{R}^2$. We also modify the bias-corrected estimators and express them as

$$\hat{\lambda}^{(VU)}(u) = \frac{1}{e_\sigma(u)} \sum_{i=1}^n \kappa_\sigma(d(u, x_i)), \quad (3.6)$$

for the uniformly corrected kernel estimator, and

$$\hat{\lambda}^{(VJ)}(u) = \sum_{i=1}^n \frac{1}{e_\sigma(x_i)} \kappa_\sigma(d(u, x_i)), \quad (3.7)$$

for the Jones-Diggle corrected kernel estimator, where $e_\sigma(u) = \int_W \kappa(d(u, y)) dy$. In the case of convex polygons, these estimators are equivalent to those derived using the Euclidean distance.

The choice of bandwidth is an important consideration in the estimation of the kernel smoothed intensity function. It is directly responsible for the approximation accuracy, i.e., 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. A good choice for the bandwidth would be a value that gives a smooth estimate for the intensity function and that captures the distribution of the data. The methods detailed in [9, 18] can be reverted to for the determination of the appropriate bandwidth.

Figure 3.3 shows the result of applying Equations 3.1 and 3.5 to two different point patterns, one on a convex window and the other on a nonconvex window. A Gaussian kernel is used for the intensity estimation. In the case of the convex window, the intensity estimates fitted using Equations 3.1 and 3.5 produce the same result since the length of the shortest path along the visibility graph on this domain is equal to the Euclidean distance. This can be seen in the left panel of Figures 3.3(b) and 3.3(c). In the right panel of Figure 3.3(b) it can be observed that the intensity estimates on the nonconvex window, calculated using Equation 3.1, disregards the window boundaries and incorrectly allocates higher intensity in areas devoid of points, whereas in the case of the intensity estimate calculated using Equation 3.5 shown in the right panel of Figure 3.3(c), the intensity estimate is more representative. The result of applying the intensity estimator in the case of the Euclidean distance and the length of the shortest path along the visibility graph will follow similarly for the edge-corrected counterparts.

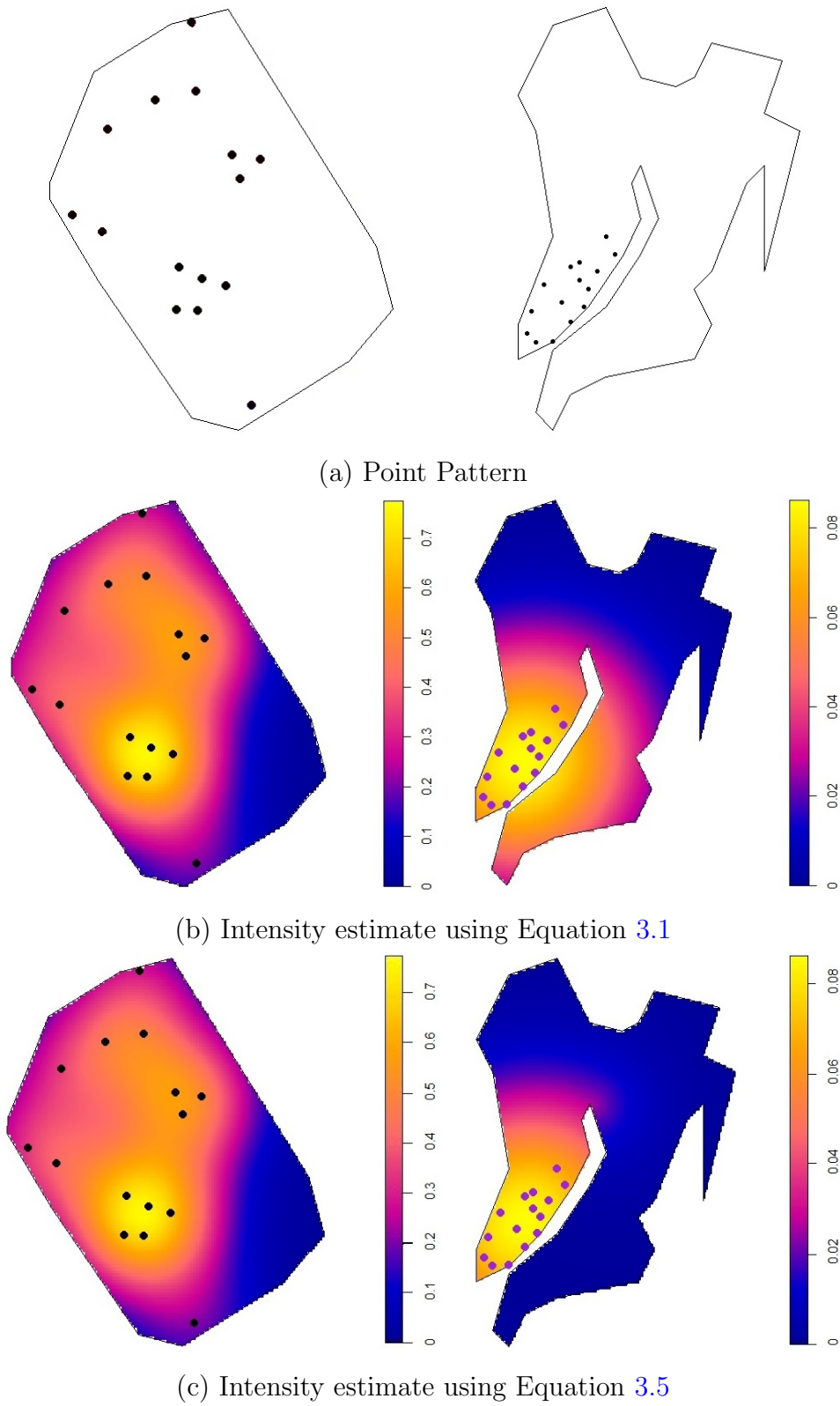


Figure 3.3: Point pattern plots and intensity estimates on convex (left) and nonconvex (right) windows fitted using Equations 3.1 and 3.5, respectively.

Figure 3.4 compares the intensity estimates of a point pattern, simulated from a homogeneous Poisson process with $\lambda = 0.5$, computed using Equations 3.3 and 3.7. Estimation is done using a Gaussian kernel. The pointwise differences illustrated in Figure 3.4(d) are computed by subtracting the estimated intensity values in Figure 3.4(b) from the estimated intensity values in Figure 3.4(c). In areas of the window where points are inter-visible, the values of the intensity estimates are relatively similar. Larger differences are observed in the regions of the window where the presence of the boundary separates point locations that are in close proximity to each other.

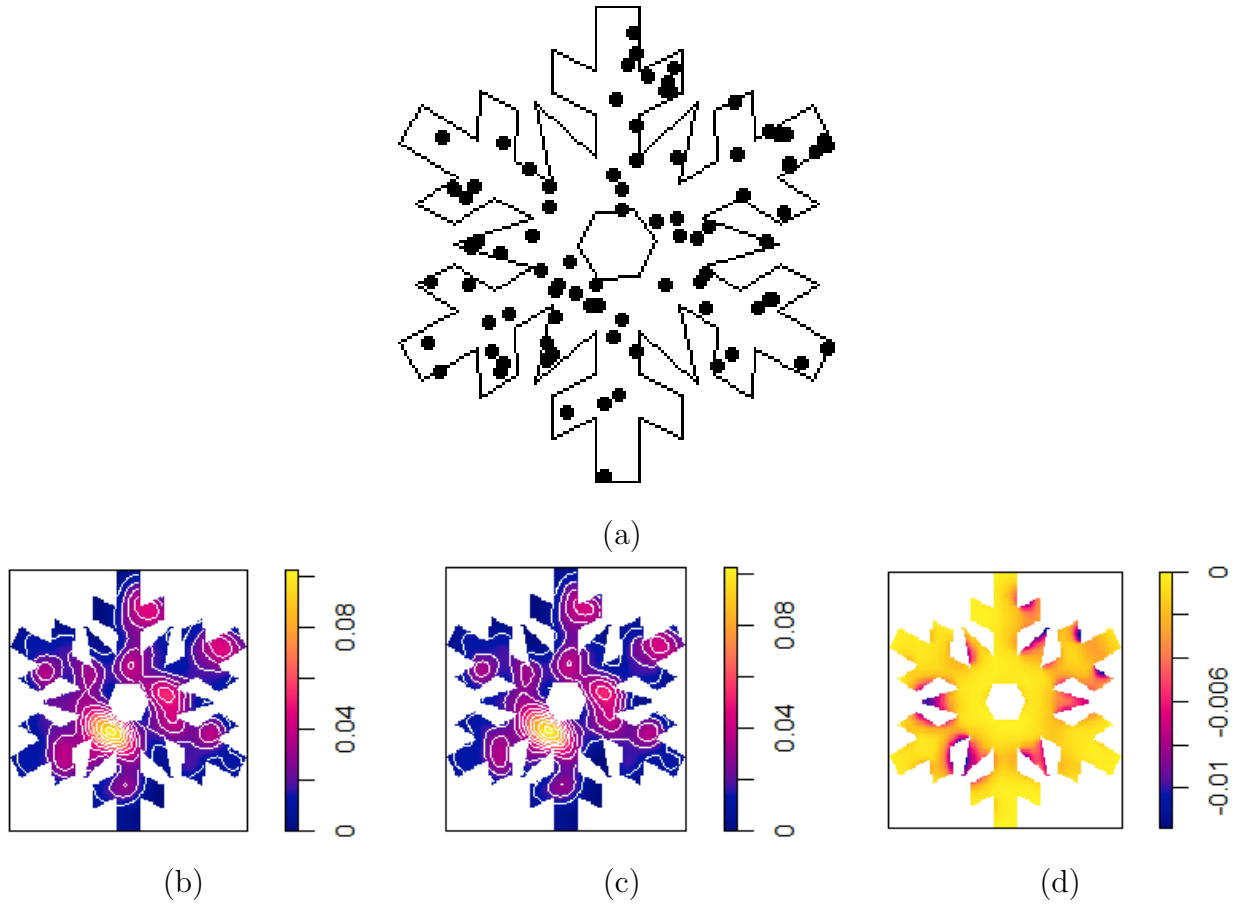


Figure 3.4: Intensity estimates of simulated point pattern. (a) Point pattern simulated from a homogeneous Poisson process with intensity 0.5. (b) Intensity estimate using Equation 3.3. (c) Intensity estimate using Equation 3.7. (d) Pointwise difference between intensity estimate using Equations 3.3 and 3.7.

Empirical results

In this section we compare the modified kernel smoothed intensity estimator based on the Euclidean shortest path (ESP) distance with its only other known competitor, a diffusion (DF) smoother proposed in [7]. For this task, 100 point patterns are simulated in a nonconvex L-shaped and irregular window from point processes with inhomogeneous intensity shown in the left panel of Figure 3.5. We perform 100 simulations due to the large computation time required to construct and run a query on the visibility graph. For each simulated point pattern, the modified kernel smoothed intensity estimate and diffusion smoother are applied. A numerical approximation of the integrated squared error (ISE),

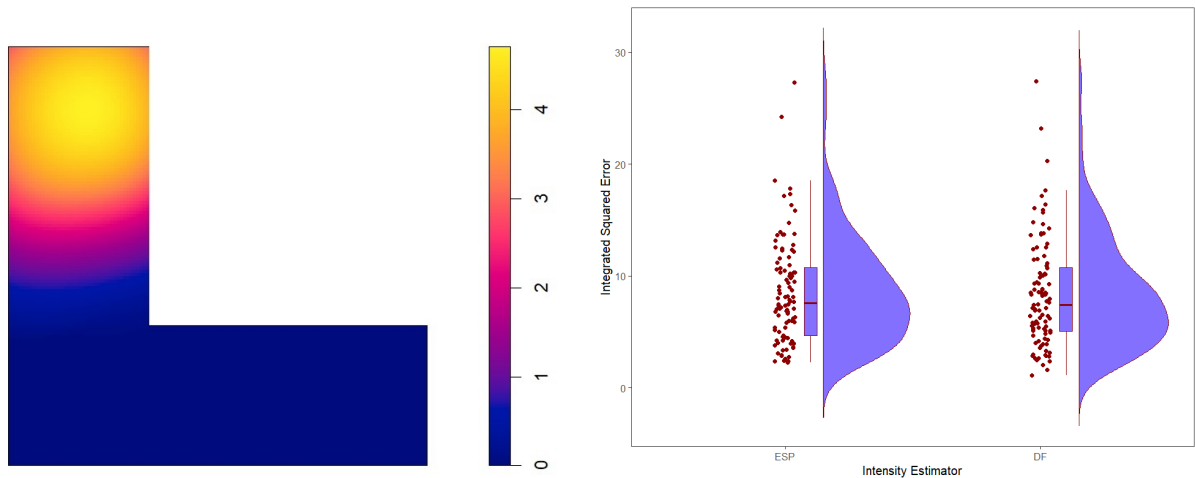
$$ISE(\hat{\lambda}) = \int_W \left(\hat{\lambda}(u) - \lambda(u) \right)^2 du, \quad (3.8)$$

is then calculated for each estimator and the average across the simulations determined to obtain a value for the mean integrated squared error (MISE). The results are presented in the raincloud plot⁵ [2] in right panel of Figure 3.5 and in Tables 3.1 and 3.2.

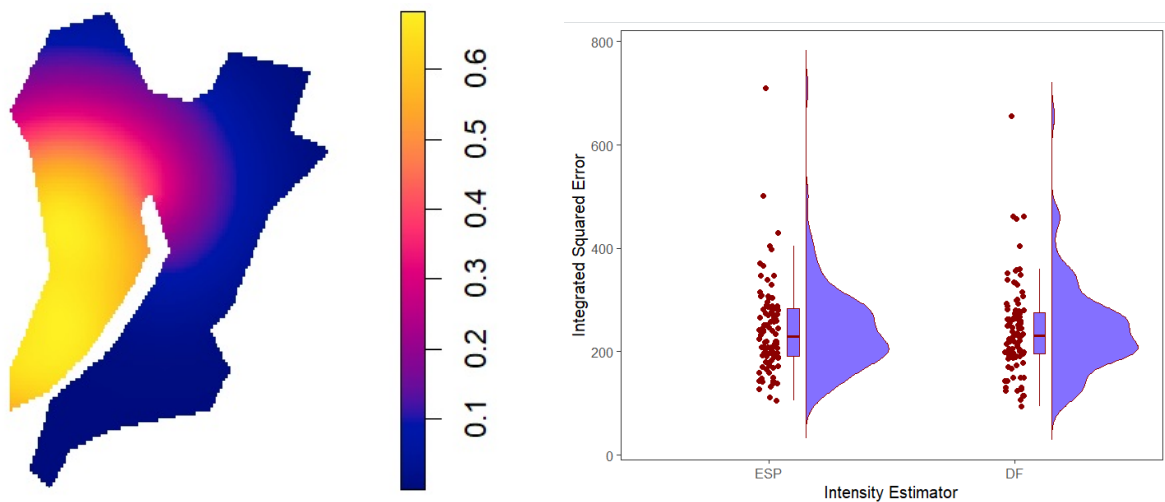
There is no significant difference between the performance of the modified kernel smoothed intensity estimator and the diffusion smoother. The modified kernel smoothed intensity estimator exhibits a slightly higher MISEs, but slightly lower standard deviation when compared to the diffusion smoother. Consequently, the modified kernel estimator remains a good candidate for estimation tasks of intensity for point patterns restricted on nonconvex windows.

The provision of the new modified kernel intensity estimate, determined through the use of a visibility graph, allows for better characterisation of the spatial relationships of points constrained within nonconvex domains. Consequently, we have addressed the “tunnelling” effect produced by the traditional kernel smoothed intensity estimate that makes use of the Euclidean distance to quantify proximity relationships.

⁵A raincloud plot is a data visualization that combines a violin plot, box plot, and raw data points to display the distribution, summary statistics, and individual observations of a dataset.



(a) L-shaped window



(b) Irregular window

Figure 3.5: Heterogeneous intensity function of a point process on two different nonconvex windows (left) and corresponding raincloud plot (right) of ISE values in the case of the Euclidean shortest path (ESP) intensity estimator and diffusion (DF) smoothing.

3.4 Application

We now apply the modified kernel smoothed intensity estimator to rural household locations in Tanzania. The data used is from a census of the Serengeti District in Tanzania's

Table 3.1: Summary statistics of the ISE values when using the intensity estimators based on the Euclidean shortest path (ESP) distance and diffusion (DF) smoothing on the L-shaped nonconvex window.

	MISE	Standard Deviation
ESP	8.5094	4.6889
DF	8.3216	4.8385

Table 3.2: Summary statistics of the ISE values when using the intensity estimators based on the Euclidean shortest path (ESP) distance and diffusion (DF) smoothing on an irregularly shaped nonconvex window.

	MISE	Standard Deviation
ESP	242.8119	85.3729
DF	242.3102	85.6489

Mara province⁶. The census comprises georeferenced data of 35 947 households distributed across 88 villages. Three villages are considered, namely Magatini, Majimoto and Hekwe with households that number 100, 336 and 235 respectively. Figure 3.6 shows Google Earth plots of the household locations for each village with terrain features of the landscape. In rural areas, steep slopes have greater costs in terms of traversal. Thus, the mountainous regions in the domain of the point patterns are obstacles that must be circumvented in order for travel between different household locations. As a preprocessing step, the window selection algorithm presented in [36] is used to construct nonconvex windows with terrain elevation as spatial covariates. The result is shown in the left panel of Figure 3.7. The right panel of Figure 3.7 shows the modified kernel smoothed intensity

⁶The data was provided by Katie Hampson, <http://www.gla.ac.uk/researchinstitutes/bahcm/staff/katiehampson/>, <http://www.katiehampson.com/#intro>, and approved for use by the Faculty of Natural and Agricultural Science Research Ethics committee under the reference NAS33/2019.

estimate. A Gaussian kernel is used and the adapted Jones-Diggle bias correction term applied. The bandwidth is chosen using the criterion defined by Diggle [18]. The modified intensity estimates produces plots that captures the spatial distribution of points restricted by mountainous regions. The influence of a point spreads along a connected path within the defined domain with tunnelling effects absent from the intensity plots.

3.5 Discussion

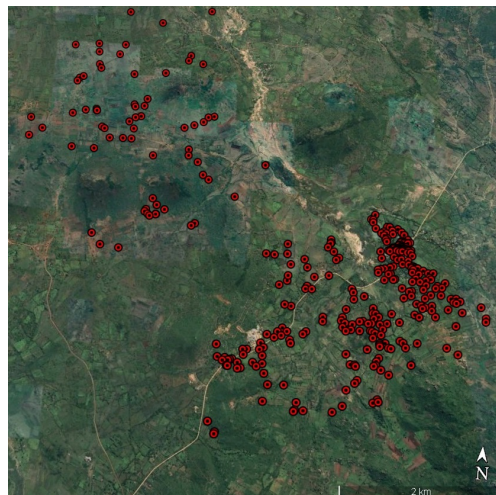
In spatial point pattern analysis, the first-order properties of a point process are traditionally defined using a function of distance between points [9, 18]. The choice of an appropriate metric to quantify proximity between points is important since this has a direct influence on the estimation of the properties of a pattern [36]. When points are observed in nonconvex window domains that constrain their movement, the Euclidean distance is not an appropriate choice for distance since this disregards the obstacles imposed by the window boundaries.

In this research, a modified kernel-smoothed intensity estimator is introduced. The purpose in proposing this estimator is to address the misrepresentation of intensity when points are confined on nonconvex windows. The traditionally kernel-smoothed intensity estimator [18, 9, 38] makes use of the Euclidean distance to govern the contribution of a data point to the intensity estimate. When points are restricted to a continuous path within the study window, the result is an overestimation of their contribution to the intensity estimate. The proposed estimator accounts for the constraints imposed by the window boundary and obstacles in the window on the movement of points. Through the use of a weighted visibility, the proposed estimator admits an estimate of intensity that is adapted to the window domain.

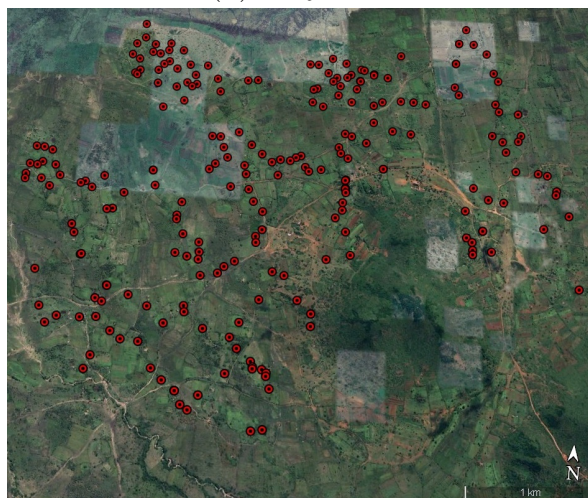
Diffusion smoothing [7] has been developed as another framework alternate to traditional smoothing. Similar to our proposed modified smoother, it produces intensity estimates that are intrinsically adapted to the study domain, i.e., it accounts for the domain boundaries and other obstacles in the domain. The empirical results in the simulation demonstrate that the diffusion smoother and the path-connected kernel smoothed intensity estimate are near identical in performance with slight differences in the values



(a) Magatini

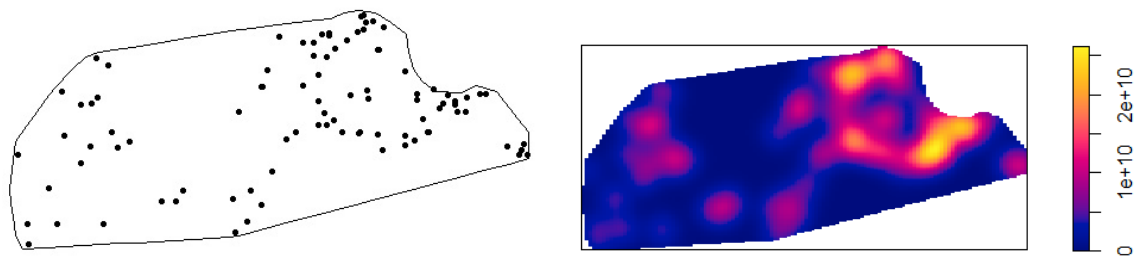


(b) Majimoto

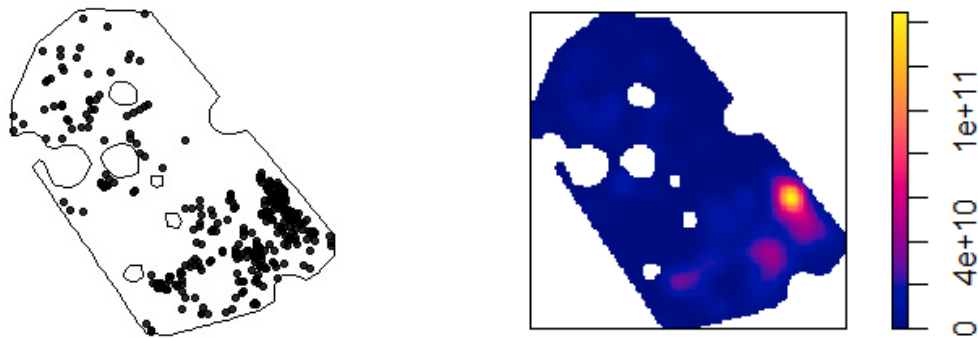


(c) Hekwe

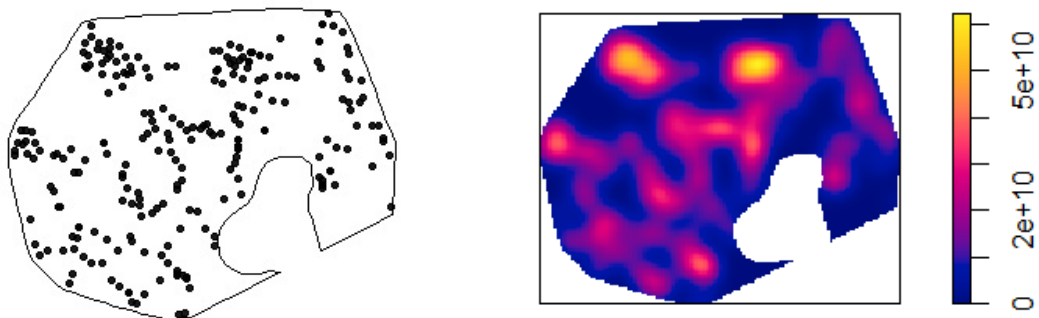
Figure 3.6: Google Earth plot of village household locations for three villages in Tanzania's Mara province with terrain map



(a) Magatini



(b) Majimoto



(c) Hekwe

Figure 3.7: Point pattern plot (left) and modified kernel smoothed intensity estimate plot (right) for villages in Tanzania's Mara province, on nonconvex windows constructed using terrain slope covariate data

of the MISE and error standard deviation.

Despite the advantages of the modified kernel smoothed intensity estimate, it does have a limitation that makes it unscalable to very large data sets. The determination of a visibility graph is computationally expensive and the computation time increases

with domains that are more complex and have a larger number of vertices [26, 25]. For example, performing a visibility query (i.e., determining if a line of sight exists between two points) for a set of fixed coordinates in a grid on a window with many obstacles can be computationally expensive as more vertices must be considered when determining interpoint visibility. Thus, for future work, more efficient techniques will be developed to optimise the computational time for constructing the weighted visibility graph.

Notwithstanding the limited coverage of bandwidth selection in this research, it remains an important consideration in the estimation of the kernel-smoothed intensity function. The choice of bandwidth influences the approximation accuracy, with large bandwidth leading to oversmoothing and small bandwidth sizes resulting in under-smoothing. Several strategies exist for determining a bandwidth based on rules of thumb [63, 66, 9] or other optimisation criteria [1, 15, 18, 9]. Future work could consider the effect of the bandwidth on the kernel smoothed intensity estimate and its extension to non-Euclidean distance smoothers.

3.6 Conclusion

Herein, we have provided a framework to incorporate the nonconvex structure of window domains to appropriately capture the spatial structure imposed by the windows geometry. The problem of finding a representative distance measure that respects the window boundaries and avoids holes in the window domain was solved by determining distance through the use of a weighted visibility graph. This distance gives a more representative measure of proximity on a connected path in a nonconvex domain than the Euclidean distance. We illustrated the usefulness of using this distance to quantify proximity via the kernel smoothed estimator and showed it is a good candidate against its other only known competitor, the diffusion smoother. We applied this modified kernel smoothed intensity estimate to rural village household locations with visibility determined along a path the circumvents mountainous regions identified by the algorithm of [36]. A drawback of using the visibility graph is the computational costs require to compute it. For future work, we will consider an implementation strategy that can improve on the computational time when determining the visibility graph.

Chapter 4

Assessing point pattern interactions constrained on nonconvex domains using visibility graphs

4.1 Abstract

Second-order quantities are commonly used to analyse point patterns for clustering, randomness, or regularity. Relationships between points are typically quantified using the Euclidean distance as a proximity measure. The application of this distance on nonconvex domains, however, may misrepresent spatial relationships when points are constrained by obstacles and the domain's boundary. Herein, we use distance along the shortest path of a weighted visibility graph to modify several second-order quantities and illustrate their usefulness in better quantifying spatial interactions between points inhibited by their environments.

4.2 Introduction

In point pattern analysis, second-order properties refer to statistical measures that describe spatial interactions between pairs of points in a point process [61, 9, 13]. These properties represent how the points are distributed relative to one another - whether

they are clustered, regular, or random.

Second-order quantities are intimately related to counts based on pairs of points. The relationships between pairs of points are determined based on proximity measures, namely the Euclidean distance. This distance is commonly used in several different instances, e.g., when assessing pairwise interactions between distinct pairs of data points using the K - and L -functions, nearest-neighbour relationships of data points via the G -function, and empty-space distance¹ in the F -function [61, 9, 39].

The Euclidean distance may lead to misleading results in nonconvex domains since it does not correctly capture the spatial restrictions imposed by the geometry of the window [36]. An illustration of this is provided in Figure 4.1 for the typical types of distances, i.e., pairwise, nearest neighbour, and empty space, that are used in second-order summary statistics. The red points comprise a point pattern (a set of locations) contained in a window (represented by the blue shaded region). When the window geometry is disregarded, the summary functions may falsely produce evidence of clustering at given ranges since the distance is underestimated, inflating the number of points that are considered neighbours at given scales. Thus, to analyse spatial patterns in these contexts, the domain geometry needs to be taken into account as interactions between points occur on a physically connected path on the window.

The limitations of the Euclidean distance as a measure of proximity between points in other settings have motivated adaptations of the second-order summary functions. In the case of the K -function, for example, Okabe et al. [45, 44], proposed an extension of the function to networks by defining distance as the shortest path between two points on the linear network. This formulation deals with spurious K -function estimates, for point patterns on a road, that may falsely produce evidence of clustering at short-ranges owing to the concentration of points along the road, or regularity at long ranges due to the separation of different roads. Empirical studies have also been conducted to evaluate the effect of using a planar K -function to detect spatial interactions on a network space that constrains data points such as in traffic analysis and urban crime distribution [77, 35].

Herein, we provide a framework to quantify interactions in point pattern data constrained in nonconvex spatial domains via the use of a weighted visibility graph to deter-

¹The distance from a fixed location in the window to the nearest data point.

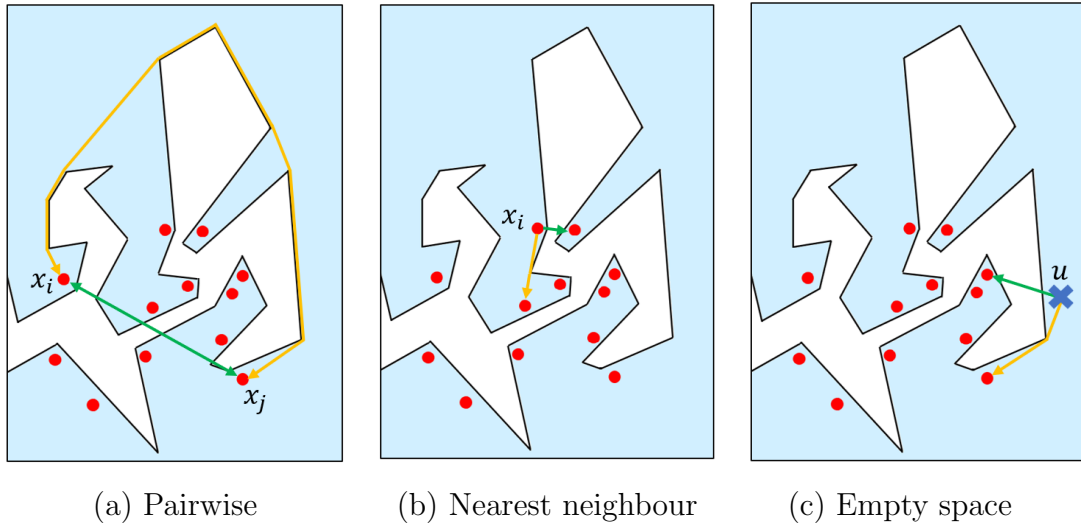


Figure 4.1: An illustration showing the effect of using the Euclidean distance (green arrows) and shortest path distance (orange arrows) to quantify proximity on nonconvex windows when computing several types of distances used in second-order quantities

mine distance on a path connected space in the spatial window. The organisation of this chapter is as follows. Section 4.3 presents some definitions and background theory. The proposed methodology to modify second-order quantities is proposed in 4.4. Results are presented in 4.5 and concluding remarks in Section 4.7.

4.3 Methodology

4.3.1 Spatial point pattern data

A point pattern [9, 13, 30] data set comprises a set of locations, $x = \{x_1, \dots, x_n\}$, observed in a spatial domain $W \subset \mathbb{R}^2$, where $n \geq 0$ is not predetermined. A point process, $X = \{X_1, \dots, X_n\}$, is a stochastic mechanism from which the point pattern is realised. Understanding the characteristics of a point pattern is often the aim of a spatial analysis and requires knowledge of both the point locations and the window on which the points are observed.

4.3.2 Second-order quantities

Ripley's K -function

The second-order intensity [14] of a point process is defined as

$$\lambda_2(u, v) = \lim_{\substack{|du| \rightarrow 0 \\ |dv| \rightarrow 0}} \frac{E[N(du) \times N(dv)]}{|du| \times |dv|}$$

where du and dv are small regions around the points u and v respectively, and their areas are denoted by $|du|$ and $|dv|$. The second-order intensity is an approximate probability that there is at least one point of the point process in each of the regions du and dv . For a stationary isotropic spatial point process, the reduced second-moment measure, termed the K -function, is related to the second order intensity via the expression

$$K(r) = \frac{1}{\lambda} \int_0^{2\pi} \int_0^r \lambda_2(u, \theta) du d\theta$$

where $r > 0$ is the distance from points in the pattern. The K -function measures the relative expected number of points within a distance r of a randomly chosen point. If a point process exhibits clustering, each point is typically surrounded by more points from the same cluster, resulting in a relatively high value of $K(r)$ for small values of r . Conversely, if a point process is randomly dispersed in space, each point is more likely to be isolated, leading to a lower value of $K(r)$ for small r . Under complete spatial randomness, $K(r) = \pi r^2$. When $K(r) > \pi r^2$, this is an indication of clustering, while $K(r) < \pi r^2$ suggests that points are more regularly spaced or exhibit repulsion. The edge-corrected empirical K -function [39, 9, 61], $\hat{K}(r)$, is commonly used as an estimator of $K(r)$ and is defined as

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n I(\|x_i - x_j\| \leq r) e_{ij}(r)$$

where $I(\cdot)$ is an indicator function, $\|\cdot\|$ is the Euclidean distance, $|W|$ is the area of the window W , and $e_{ij}(r)$ is a weight to correct for edge effects, i.e., the proportion of the circumference of a circle centered at x_i , passing through x_j , and contained in W .

Empty-space function F

The empty-space function F [39, 9] for a stationary point process X in $W \subset \mathbb{R}^2$ is the distribution function of the distances from a fixed point $u \in W$ to the nearest point in X . Mathematically, this is defined as

$$F(r) = P\left(\min_{X_i \in X} \|u - X_i\| \leq r\right),$$

where $r \geq 0$ and the distance $\min_{X_i \in X} \|u - X_i\|$ is termed the empty-space distance from a fixed location u to the point X_i in the point process. If X is completely spatially random, the theoretical F -function is

$$F_{CSR}(r) = 1 - \exp(-\lambda\pi r^2)$$

where λ is the intensity of the stationary point process. An estimator of the F -function is typically determined using the empirical cumulative distribution function of the empty-space distances with respect to the observed data point locations, i.e., $x = \{x_1, \dots, x_n\}$ and is given by

$$\hat{F}(r) = \frac{1}{m} \sum_{j=1}^m I\left(\min_{x_i \in x} \|u_j - x_i\| \leq r\right)$$

where the u_j 's are fixed locations, usually the centers of the pixels in a grid on W . In the presence of edge effects a border corrected estimate [39, 9]

$$\hat{F}_{EC}(r) = \frac{\sum_{j=1}^m I\left(\min_{x_i \in x} \|u_j - x_i\| \leq r\right) \times I(b_j > r)}{\sum_{j=1}^m I(b_j > r)},$$

can be used where b_j is the distance from u_j to the boundary of the window. The border corrected estimate of the F -function effectively computes an empirical cumulative distribution function based on the fixed locations that have an empty-space distance that is smaller than the distance to the boundary for a given r .

To identify clustering, regularity, or complete spatial randomness, the empirical F -function is compared against the theoretical F -function. An $\hat{F}(r) > F_{CSR}(r)$ provides an indication of regularity since for a given distance r , the empty-space distances are shorter than expected under completely spatially randomness. When $\hat{F}(r) < F_{CSR}(r)$, this suggest that there may be clustering since the empty-space distances are larger than expected under completely spatially randomness.

Nearest-neighbour function G

The nearest-neighbour function G [39, 9] for a stationary point process X in W , with intensity λ , is defined as

$$G(r) = \frac{1}{\lambda|W|} E \left[\sum_{X_i \in X} I \left(\min_{X_j \in X \setminus \{X_i\}} \|X_i - X_j\| \leq r \right) \right].$$

The G -function is a measure that characterises the distribution of distances from each point in a point pattern to its nearest neighbour. Under complete spatial randomness the theoretical G -function is

$$G_{CSR}(r) = 1 - \exp(-\lambda\pi r^2)$$

which is equivalent to the empty-space function for the same process, i.e., $G_{CSR} \equiv F_{CSR}$. The empirical cumulative distribution function of the nearest neighbour distances based on the n observed locations x is used as an estimator and is defined as

$$\hat{G}(r) = \frac{1}{n} \sum_{i=1}^n I \left(\min_{x_j \in x \setminus \{x_i\}} \|x_i - x_j\| \leq r \right).$$

The edge effect border corrected G -function [39, 9] is given by

$$\hat{G}_{EC}(r) = \frac{\sum_{i=1}^n I \left(\min_{x_j \in x \setminus \{x_i\}} \|x_i - x_j\| \leq r \text{ and } b_i \geq r \right)}{\sum_{i=1}^n I(b_i \geq r)}$$

where b_i is the distance from a data point x_i to the boundary of the window W . If a point process is clustered, the nearest neighbour distances will typically be smaller than they would be under complete spatial randomness. Thus a $\hat{G}(r) > G_{CSR}(r)$ provides an indication of clustering. In the case of a regular point process, the nearest neighbour distances would be larger than expected under complete spatial randomness, so $\hat{G}(r) < G_{CSR}(r)$ suggests regularity.

The J -function

The J -function [73, 39, 9] is a summary statistic that combines the properties of the nearest-neighbour distance function G and the empty-space function F . For a stationary

point process X , the J -function is defined as

$$J(r) = \frac{1 - G(r)}{1 - F(r)},$$

where $r \geq 0$ and $F(r) < 1$. Under complete spatial randomness $F \equiv G$ resulting in $J(r) \equiv 1$. Values $J(r) < 1$ provide evidence of clustering in a point pattern and values $J(r) > 1$ indicate regularity.

The functions described in Section 4.3.2 rely on the Euclidean distance which does not give a representative measure of proximity on nonconvex spatial domains that limit the movement of points. The result of disregarding the constraints of the nonconvex structure of the window results in artefacts in spatial statistical summaries and spurious estimates. Owing to this, we propose using a representative distance measure on a path connected space, through the application of a weighted visibility graph.

4.3.3 Weighted visibility graph

A weighted visibility graph [25] of a polygon P , denoted by $G(P)$, is an ordered triple (V, E, w) comprising

- $V = \{v_1, \dots, v_m\}$, a set of the vertices² of P ,
- $E = \{(v_i, v_j) | v_i, v_j \in V \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, m\}$, a set of edges between vertices that are intervisible (in a geometric context, two vertices v_i and v_j , $v_i \neq v_j$, are visible to each other if the line segment connecting them does not intersect the outer boundary of P or hole boundaries, that is $\alpha v_i + (1 - \alpha)v_j \in P$, $\forall \alpha \in (0, 1)$), and
- $w : E \mapsto \mathbb{R}^+$, a function mapping edges to the value of the Euclidean distance between the vertices it connects.

A representative distance measure between two points in a point pattern, one that quantifies the length of a path that does not cross the window's boundary and holes in its

²In the case of a polygon with holes, the vertices of the hole boundaries are included in the visibility graph vertex set.

domain, can be found by determining the shortest path on the weighted visibility graph of the window.

4.3.4 Shortest path distance

A path [25, 33] from x to y is a finite sequence of vertices $v_0 = x, v_1, \dots, v_{k-1}, v_k = y$; the path begins at the point x and proceeds from vertex to vertex in the graph until it reaches y . The sum of all edge weights along the shortest path,

$$d(x, y) = \sum_{h=0}^{k-1} w(v_h, v_{h+1}),$$

provides a measure of distance within the spatial domain that avoids obstacles, such as boundaries and holes. We use the length of the shortest path traversed between two points in the window along the visibility graph to determine distance that accounts for the geometric structure of nonconvex windows that constrain the movement of points.

4.4 Visibility graph-based adaptations

4.4.1 Visibility-based pairwise, nearest-neighbour, and empty-space distance

Consider $W \subset \mathbb{R}^2$, the spatial domain in which a planar point pattern $x = \{x_1, \dots, x_n\}$ is observed. The visibility graph of W , $G(W)$, is defined as the triplet $G(W) = (V_W, E_W, \rho_W)$ where

- $V_W = \{v_1, \dots, v_q\}$, is the set of vertices of W ,
- $E_W = \{(v_i, v_j) | v_i, v_j \in V, \alpha v_i + (1 - \alpha)v_j \in W, \forall \alpha \in (0, 1), \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, q\}$, and
- $\rho_W : E \mapsto \mathbb{R}^+$ is the function $\rho(v_i, v_j) = \|v_i - v_j\|$.

A path, $P(u_i, u_j)$, from u_i to u_j , $u_i, u_j \in W$, along the visibility graph of W is the sequence of vertices traversed on the graph with a

- vertex set $V_{u_i, u_j} = V_W \cup \{u_i, u_j\}$,
- an edge set $E_{u_i, u_j} = \{(v_i, v_j) | v_i, v_j \in V_{u_i, u_j}, \alpha v_i + (1 - \alpha)v_j \in W, \forall \alpha \in (0, 1), \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, q + 2\}$, and
- a weight function $\rho_{u_i, u_j} : E_{u_i, u_j} \mapsto \mathbb{R}^+$ is the function $\rho(v_i, v_j) = \|v_i - v_j\|$.

The shortest path between two points $u_i, u_j \in W$ is the path that minimises the sum of weights along the edges traversed on the graph, i.e.,

$$P_S(u_i, u_j) = \arg \min_{P(u_i, u_j)} \left(\sum_{h=0}^{k_P-1} \rho(v_h, v_{h+1}) \right)$$

where k_P is the number of vertices traversed on path P , $v_0 = u_i$, and $v_{k_P} = u_j$.

We define the visibility-based pairwise distance as

$$d(x_i, x_j) = \rho(P_S(x_i, x_j)) = \sum_{h=0}^{k_{P_S}-1} \rho(v_h, v_{h+1})$$

where $v_0 = x_i, v_{k_{P_S}} = x_j$, and k_{P_S} is the number of vertices traversed on the shortest path $P_S(x_i, x_j)$ between x_i and x_j . The visibility-based nearest-neighbour distance of a point x_i is defined as

$$d(x_i, x \setminus x_i) = \min_{x_j \in x \setminus \{x_i\}} \rho(P_S(x_i, x_j)) = \min_{x_j \in x \setminus \{x_i\}} \left(\sum_{h=0}^{k_{P_S}-1} \rho(v_h, v_{h+1}) \right)$$

where $v_0 = x_i, v_{k_{P_S}} = x_j$, and k_{P_S} is the number of vertices traversed on the shortest path $P_S(x_i, x_j)$ between x_i and x_j . The empty-space distance is defined as

$$d(u, x) = \min_{x_i \in x} \rho(P_S(u, x_i)) = \min_{x_i \in x} \left(\sum_{h=0}^{k_{P_S}-1} \rho(v_h, v_{h+1}) \right)$$

where $v_0 = u, v_{k_{P_S}} = x_i$, and k_{P_S} is the number of vertices traversed on the shortest path $P_S(u, x_i)$ between u and x_i . If W is convex, the visibility condition is satisfied for all combinations of vertices in the vertex set of the visibility graph. Thus, in this setting, the length of the shortest path is identical to the Euclidean distance. These distances now form the basis on which the K , F , G , and J functions are calculated.

4.5 Results

In this section, we consider the result of applying the modified K , F , G , and J -functions to a simulated point patterns. Clustered, random, and regular point patterns are considered. The point patterns are generated from a Matern cluster process, a homogeneous Poisson point process, and a simple sequential inhibition process and are depicted in Figures 4.2(a), 4.3(a), and 4.4(a), respectively. To simulate the point patterns we make use of the functions `rpoispp`, `rSSI`, and `rMatClust` in the `spatstat` package in R. We assume that the point patterns are complete, thus the true distances are observable in the window with no edge effects present. Simulation envelopes [9] are determined under the assumption of spatial randomness. The shortest path distances along the visibility graph, as outlined in Section 4.4, are used and the point patterns along with their correspond empirical function values are determined, in the Figures.

It is observed in Figure 4.2 that all the functions provide evidence of clustering. For the K -function, in Figure 4.2(b), this is evidenced by the observed empirical function values have plotting above the simulation envelopes. This is echoed as well through the other second-order summary functions with the observed empirical values of the F -function (Figure 4.2)(c) plotting below the simulation envelopes across differing scales plotting below the simulation envelopes, the G -function, in Figure 4.2(d) is above the simulation envelope, showing that points are typically more clustered than what would be expected under spatial randomness.

The observed random point pattern, depicted in Figure 4.3(a), has observed function values, Figures 4.3(b) through to 4.3(e), that plot within the simulation envelopes. This supports the notion that the points are realised from a point process that is spatially random. In the last case of the regular pattern, the functions indicate the present of regularity a different scales, i.e., since K is below the simulation envelopes, F is above, G is below, and J is above the envelopes. Ultimately, this demonstrates the ability of the adapted second-order quantities considered detecting the presence of different types of relationships between points at varying scales.

4.6 Discussion

Proximity measures play an important role in determining the second-order characteristics of a point pattern. The K -, F -, G -, and J -function [9] aid in analysing spatial dependence, clustering, or dispersion, but their effectiveness depends on using appropriate distance metrics that reflect true spatial interactions.

It is typical to use the Euclidean distance when computing these summaries. This distance measure, however, may not accurately capture the underlying structure of point patterns in certain settings. For instance, when dealing with point patterns constrained on irregular spaces such as in urban environments, roads, or terrain with obstacles. In the case of linear networks efforts have been made to circumvent this limitation by adapting measures to correctly account for the network structure [44, 77, 35]. For planar point patterns, there is limited literature which offers non-Euclidean alternatives to spatial statistical methods such as in [42].

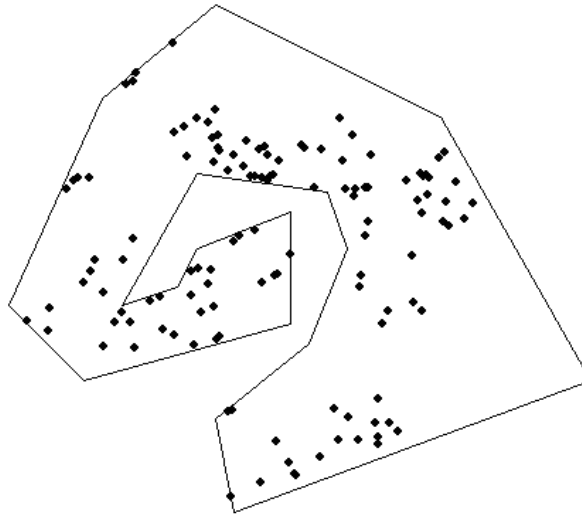
The modifications, herein, allow for a more representative depiction of point interactions. In these cases, distances calculated along a weighted visibility graph provide a realistic measure of accessibility in restrictive environments. By incorporating these representative distances, second-order summary functions can better capture the true interactions between points, offering reliable insights into a point pattern's spatial structure under the given constraints. It was demonstrated through experimental results that the adjusted functions are able to detect clustering, randomness, and regularity within point patterns. A great limitation in characterising distances along connected physical paths (determined by line of sight in the case of the visibility graph) is that it is computationally expensive. Future work will consider efficient implementation strategies to leverage the power of the adjusted measures without requiring a heavy demand on computational resources.

A possible extension of the current methodology could be to involve covariates, such as terrain features, infrastructure, or demographic data into distance calculations. For example, in an ecological setting, the distribution of species may be influenced by terrain slope or proximity to water, and traditional distance metrics fail to account for these factors [74, 50, 28]. Representative distances, derived from models that integrate such covariates could thus allow for more accurate analyses by reflecting both the interaction

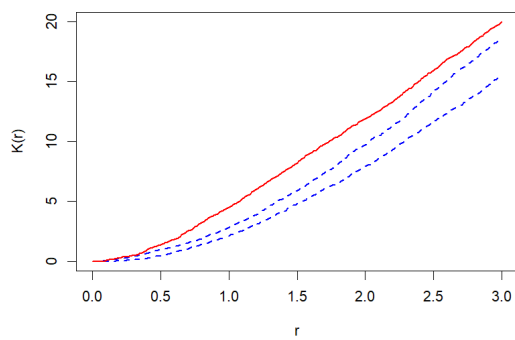
between points and the environmental context that shapes their distribution .

4.7 Conclusion

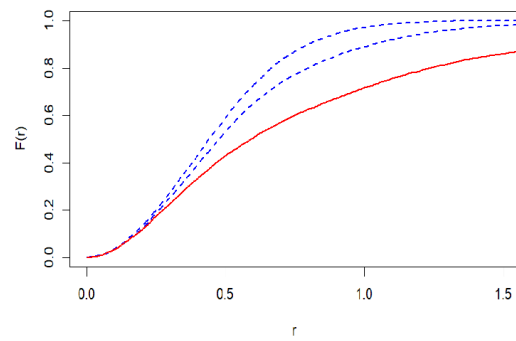
Herein, we considered second-order property measures typically used in point pattern analysis to assess the departure of a point pattern from spatial randomness. The functions are based on proximity relationships that are described through distance distributions. Since the Euclidean distance does not account for obstructed paths on a connected space in nonconvex domains, we proposed a modification of the description of the pairwise, nearest-neighbour, and empty-space distances using shortest-path distances on a weighted visibility graph.



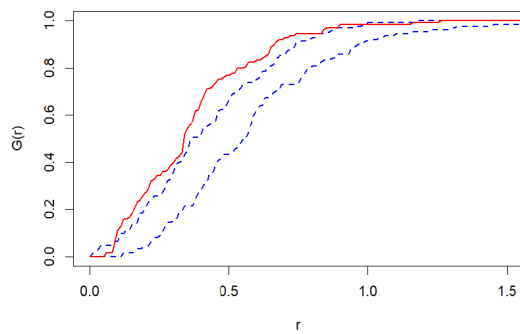
(a) Clustered point pattern



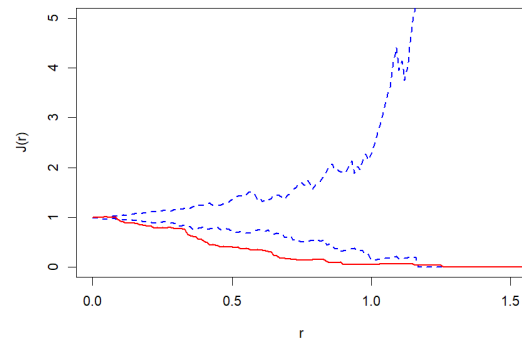
(b) K -function



(c) F -function

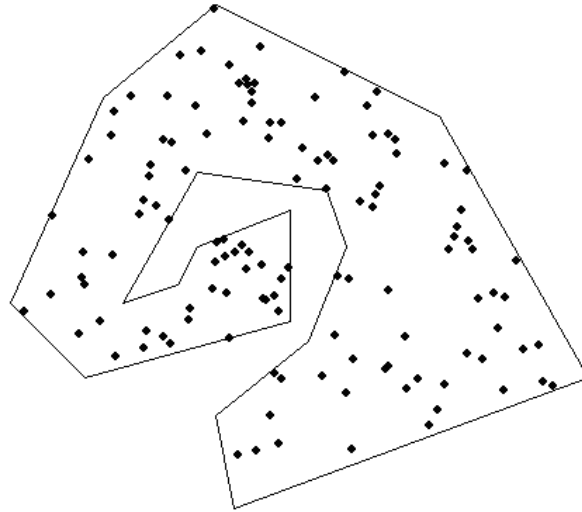


(d) G -function

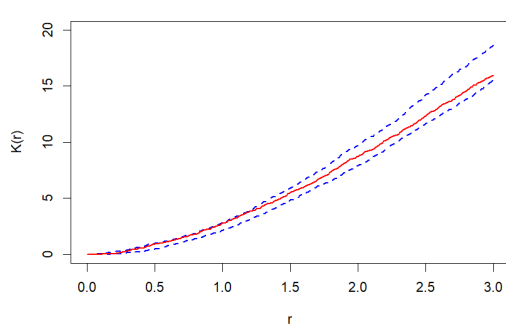


(e) J -function

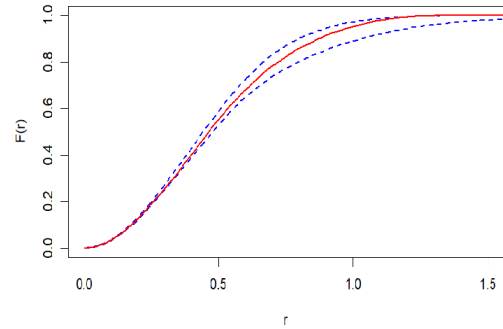
Figure 4.2: Modified K , F , G , and J -function estimates for a simulated clustered point pattern.



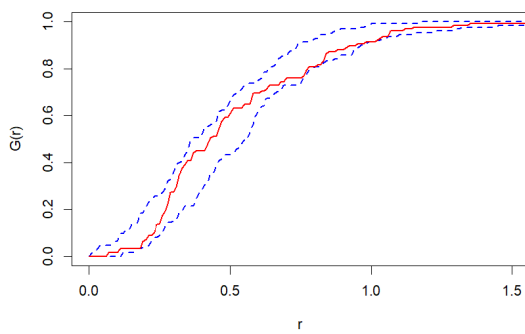
(a) Random point pattern



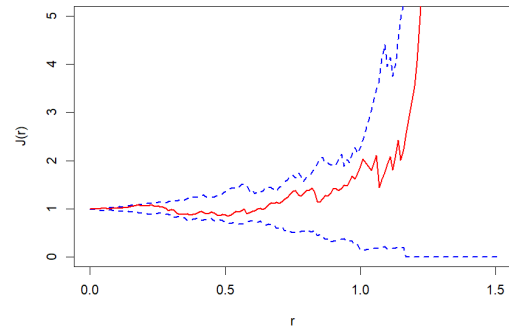
(b) K -function



(c) F -function

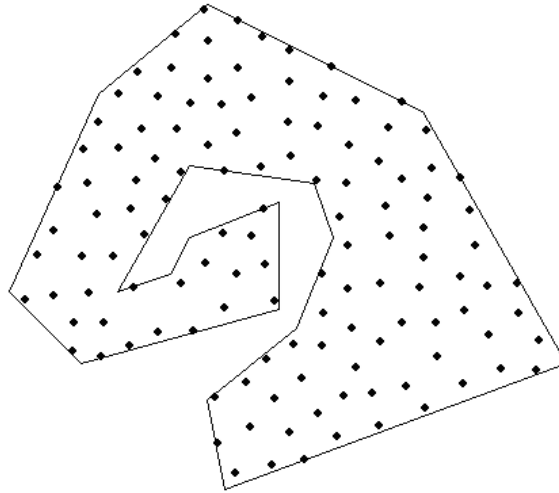


(d) G -function

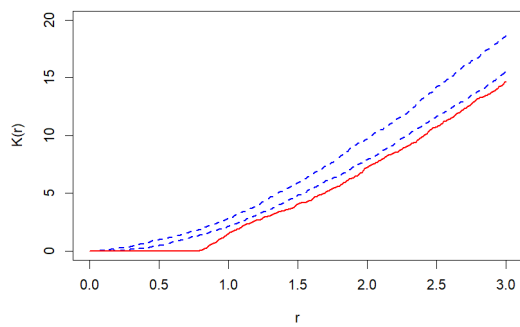


(e) J -function

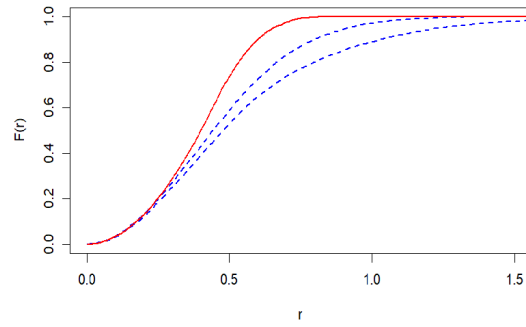
Figure 4.3: Modified K , F , G , and J -function estimates for a simulated random point pattern.



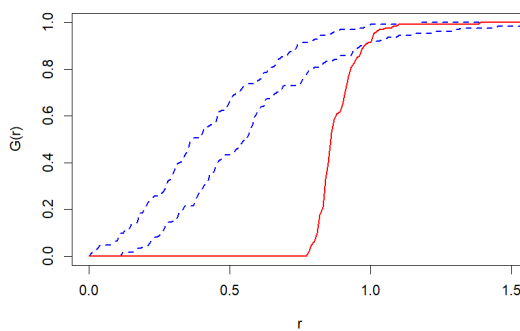
(a) Regular point pattern



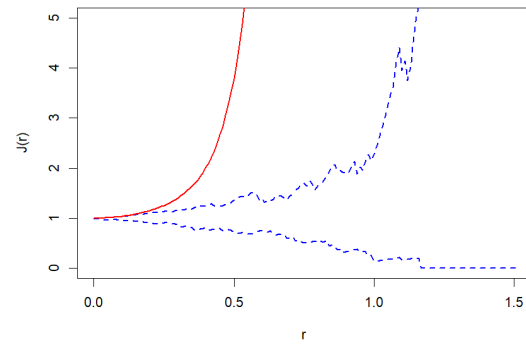
(b) K -function



(c) F -function



(d) G -function



(e) J -function

Figure 4.4: Modified K , F , G , and J -function estimates for a simulated regular point pattern.

Chapter 5

A framework for analysing point patterns on nonconvex domains using visibility graphs and multidimensional scaling

The work presented in this chapter is based on an article submitted to the South African Statistical Association 2024 unpublished postgraduate chapter competition for which it received third prize.

5.1 Abstract

A visibility graph comprises a set of points, represented as vertices in space, that connect those that are inter-visible with edges. Shortest-path distances computed on the visibility graph provide a representative model for describing proximity relationships in point patterns restricted on nonconvex domains; since they account for obstacles such as holes and the window boundaries. The usefulness of applying visibility graphs to modify statistical point pattern summary measures has been demonstrated in our previous works. Herein, we propose a strategy to embed visibility graph distances into Euclidean metric space using multidimensional scaling. The aim of which is to simplify analyses, harness

Euclidean distance methods that are already well-developed, and retain (as much as possible) the true proximity relationships on a nonconvex spatial domain. The kernel smoothed intensity estimate and K -function are determined in this new spatial context and are used as validation.

Keywords: nonconvex spatial domain, kernel smoothing, Euclidean distance, point pattern, visibility graph

5.2 Introduction

The properties of a point pattern are traditionally determined using distance- and density-based measures [14, 9, 61, 18, 39]. The choice of distance metric is crucial for accurately characterising the spatial relationships and dependencies within a given study area. While the Euclidean distance is often used to quantify proximity between objects in a point pattern, its application on nonconvex domains leads to misleading results that do not correctly represent the restrictions imposed by the geometry of the window [36].

Capturing proximity relationships using a visibility graph [25] in point patterns constrained on nonconvex windows is valuable and has been applied in [36] to modify the kernel smoothed intensity estimate. Visibility graphs [25] are mathematical graphs that describe a network of intervisible locations. They encode spatial relationships by considering each vertex of an obstacle or boundary as points and connecting them with edges based on lines of sight between points. Using the visibility graph allows us to handle complex geometric structures in real-world applications, where traditional Euclidean metrics may not be valid; especially when correctly describing spatial interactions is important for understanding movement or connectivity.

Despite its usefulness in appropriately quantifying proximity relationships, constructing the visibility graph and using it in point pattern analysis algorithms can be computationally intensive. For example, performing a visibility query (i.e., determining whether a line of sight exists between two points) for a set of fixed coordinates in a grid during the process of kernel smoothing in a domain with many obstacles can be computationally expensive as the complexity, number of pixel cells, and domain obstacles increase [25]. Techniques developed to try address the computational challenges of optimising

the visibility graph construction include [47, 26, 29, 25, 69]. While visibility graph construction optimisation algorithms are invaluable for reducing computational resources in some route planning applications [64, 76], their utility in point pattern analysis is constrained by scalability issues.

Owing to the scalability issues, we are interested in embedding the distances computed on the visibility graph in a Euclidean metric space to leverage the computationally optimal techniques to calculate distance in such spaces. The technique usually employed to do this is multidimensional scaling (MDS) [72, 11]. MDS is used to convert the distances between each pair of q elements in a set to a configuration of q points positioned within an abstract Cartesian space. The embedding seeks to minimise the error between the original distances and the new distances produced in the embedded space.

Herein, we provide a framework to support analysis of point pattern data constrained in nonconvex spatial domains. We apply a graph theoretic approach through the use of a weighted visibility graph to determine distance on a path connected space in the spatial window. The resultant distances are embedded in a Euclidean metric space using MDS to leverage the available statistical summary measures that are well developed for implementation on convex windows.

The remainder of this chapter is organised as follows. Section 5.3 provides some definitions and background theory. Section 5.4 gives a presentation of the proposed framework for analysing point patterns on nonconvex windows. Section 5.5 is dedicated to a discussion of the results of applying the proposed framework. We finish with a discussion in Section 5.6 and concluding remarks in Section 5.7.

5.3 Background Theory

We begin with the definition of a point pattern data set and its first- and second-order properties in Sections 5.3.1 and 5.3.2. Sections 5.3.3 and 5.3.4 follow with a discussion of weighted visibility graphs and shortest-path distances, respectively. Multidimensional scaling is outlined in Section 5.3.5 and the proposed framework for analysis, using multidimensional scaling, is presented in Section 5.4.

5.3.1 Spatial point pattern data

A point pattern [9, 13, 30] refers to a set of points, $x = \{x_1, \dots, x_n\}$, in a defined region of space $W \subset \mathbb{R}^2$. The points represent the locations of objects, entities, or events. The main aim of an analysis is typically to understand the underlying random mechanism, termed a point process and denoted by $X = \{X_1, \dots, X_n\}$, that produced the point configuration. A scientific inquiry into the properties of a point pattern could involve assessing its departure from stationarity and isotropy. A point process is considered stationary if it is invariant under translations and isotropic if it is invariant under rotations.

To analyse point patterns, the data must comprise the specific locations $\{x_i\}$ of points and the window W where points are observed. These components are important since the properties of point patterns inherently rely on the definition of the window and how proximity relationships are determined on it [9, 18, 61, 13].

5.3.2 First- and second-order properties

Point processes are commonly characterised using their first- and second-order intensity measures. The first-order intensity [14, 9, 39] at a location $u \in W$ is defined as

$$\lambda(u) = \lim_{|du| \rightarrow 0} \frac{E[N(du)]}{|du|},$$

where du is a small region around u , $E[\cdot]$ is the expected value operator, $|du|$ is the area of the small region, and $N(du)$ is the number of points in the region du . The intensity is the average number of points per unit area. It is a constant λ for stationary isotropic point processes. The kernel smoothed intensity estimate [14, 9, 39] is commonly used as an estimator of varying intensity. The intensity estimate is given by,

$$\hat{\lambda}(u) = \sum_{i=1}^n \frac{1}{e_{\sigma}(u)} \kappa_{\sigma}(\|u - x_i\|) \quad (5.1)$$

where $\|\cdot\|$ is the Euclidean distance, σ is the bandwidth that controls the smoothness of the estimator, $\kappa_{\sigma}(u) = \sigma^{-2}K(\frac{u}{\sigma})$, and $\kappa(\cdot)$ is a kernel function, usually a probability density on $\mathbb{R}^2$, and $e_{\sigma}(u) = \int_W \kappa(u - y)dy$ is a bias correction term for boundary effects.

The second-order intensity [14, 9, 61, 39] of a point process is defined as

$$\lambda_2(u, v) = \lim_{\substack{|du| \rightarrow 0 \\ |dv| \rightarrow 0}} \frac{E[N(du) \times N(dv)]}{|du| \times |dv|}$$

where du and dv are small regions around the points u and v respectively, and their areas are denoted by $|du|$ and $|dv|$. The second-order intensity is an approximate probability that there is at least one point of the point process in each of the regions du and dv . For a stationary isotropic spatial point process, the reduced second-moment measure, termed the K -function, is related to the second order intensity through the expression

$$K(r) = \frac{1}{\lambda} \int_0^{2\pi} \int_0^r \lambda_2(u, \theta) du \, d\theta$$

where r is the distance (usually Euclidean) from points in the pattern. The K -function measures the relative expected number of points within a distance r of a randomly chosen point. If a point process exhibits clustering, each point is typically surrounded by more points from the same cluster, resulting in a relatively high value of $K(r)$ for small values of r . Conversely, if a point process is randomly dispersed in space, each point is more likely to be isolated, leading to a lower value of $K(r)$ for small r . Under complete spatial randomness, $K(r) = \pi r^2$. When $K(r) > \pi r^2$, this is an indication of clustering, while $K(r) < \pi r^2$ suggests that points are more regularly spaced or exhibit repulsion. The edge-corrected empirical K -function, $\hat{K}(r)$, is commonly used as an estimator of $K(r)$ and is defined as

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n I(\|x_i - x_j\| \leq r) e_{ij}(r) \quad (5.2)$$

where $I(\cdot)$ is an indicator function and $e_{ij}(r)$ is a weight to correct for edge effects, e.g., a border or isotropic correction [61, 39, 9]. For the case of an isotropic correction, the distances are reweighted using the proportion of the circumference of a circle centered at x_i , passing through x_j , and contained in W . Additional second order summary quantities are available that make use of the nearest-neighbour and empty-space distances. The reader is referred to [9, 39] for a discussion of these.

Equations 5.1 and 5.2 are commonly used summary measures (among others) that use the Euclidean distance to characterise the first- and second-order nature of a point

pattern. The Euclidean distance, however, is not a representative measure of proximity on nonconvex domains since it assumes an unobstructed connected path between points. Consequently, spurious estimates are produced and artefacts arise. Owing to this, we propose using a weighted visibility graph to find representative proximity relationships.

5.3.3 Weighted visibility graph

A weighted visibility graph [25] of a polygon P , denoted by $G(P)$, is an ordered triple (V, E, w) comprising

- $V = \{v_1, \dots, v_m\}$, a set of the vertices¹ of P ,
- $E = \{(v_i, v_j) | v_i, v_j \in V \text{ and } v_i \neq v_j \text{ for } i, j = 1, \dots, m\}$, a set of edges between vertices that are intervisible (in a geometric context, two vertices v_i and v_j , $v_i \neq v_j$, are visible to each other if the line segment connecting them does not intersect the outer boundary of P or hole boundaries, that is $\alpha v_i + (1 - \alpha)v_j \in P$, $\forall \alpha \in (0, 1)$), and
- $w : E \mapsto \mathbb{R}^+$, a function mapping edges to the value of the Euclidean distance between the vertices it connects, i.e., $w(v_i, v_j) = \|v_i - v_j\|$ for $(v_i, v_j) \in E$.

A representative distance measure between two points in a point pattern, one that quantifies the length of a path that does not cross the window's boundary and holes in its domain, can be found by determining the shortest path on the weighted visibility graph of the window.

5.3.4 Shortest path distance

A path [25, 33] from x to y is a finite sequence of vertices $v_0 = x, v_1, \dots, v_{k-1}, v_k = y$; the path begins at the point x and proceeds from vertex to vertex in the graph until it reaches y . The sum of all edge weights along the shortest path,

$$d(x, y) = \sum_{h=0}^{k-1} w(v_h, v_{h+1}),$$

¹In the case of a polygon with holes, the vertices of the hole boundaries are included in the visibility graph vertex set.

provides a measure of distance within the spatial domain that avoids obstacles, such as boundaries and holes. Figure 5.1 shows the shortest path (red line) between two points (green dots) in a nonconvex polygon. The visibility graph (blue lines) of the polygon is also shown.

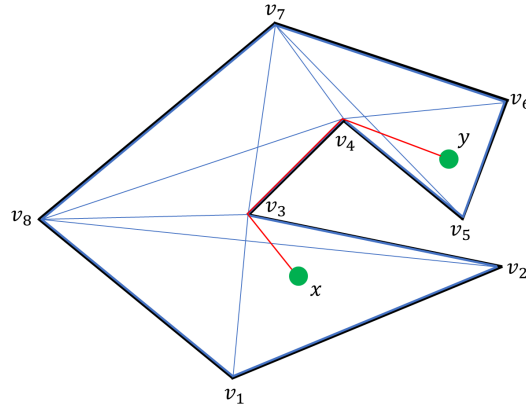


Figure 5.1: Shortest path (red line) between two points (green dots) and visibility graph (blue lines) on a nonconvex polygons.

To incorporate visibility into point pattern analysis, we embed the metric space² with a shortest path distance metric into Euclidean space using multidimensional scaling. This method simplifies the interpretation of complex, path-dependent, distances while preserving the original space’s geometric and topological properties. It also allows for the application of point pattern techniques suited for convex spaces whilst improving estimation of first- and second-order properties of point patterns constrained on nonconvex domains.

5.3.5 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) [72, 11] is a technique used to convert the distances between each pair of q elements in a set into a configuration of q points positioned within an abstract Cartesian space.

²A metric space (X, d) is a set X with a metric d that defines a notion of distance between its elements.

Given a distance matrix $D = [d_{ij}]$, where $d_{ij} = d(x_i, x_j)$ is the distance metric between the points x_i and x_j in the set, and a specified number of dimensions, $p < q$, the MDS algorithm positions each object within a p -dimensional space (a reduced-dimensional representation) to maintain the original distances between objects as closely as possible. Metric MDS is applied to quantitative dissimilarities and non-metric to qualitative (ordinal) dissimilarities [11]. We deal with the former.

In classical MDS the aim is to minimise the Strain loss function [11, 72, 37],

$$\text{Strain} = \sum_{i < j} (\delta_{ij}^2 - d_{ij}^2) \quad (5.3)$$

where δ_{ij} are the low-dimensional distances produced from the solution. The solution can be obtained analytically for $\delta_{ij} = \|y_i - y_j\|$, $y_i, y_j \in \mathbb{R}^p$ via a matrix eigendecompositions [37]. Usual metric MDS seeks to find an optimal configuration of points that minimises the stress loss function

$$\text{Stress} = \left(\frac{\sum_{i < j} (\delta_{ij} - d_{ij})^2}{\sum_{i < j} d_{ij}^2} \right)^{\frac{1}{2}}. \quad (5.4)$$

A solution can then be obtained using numerical optimisation methods. A thorough treatment of these techniques along with other loss functions is provided in [11]. In this research we consider Sammon mapping [62, 11]. Sammon mapping [62] aims to minimise the error

$$\text{Sammon Stress} = \frac{1}{\sum_{i < j} d_{ij}} \sum_{i < j} \frac{(\delta_{ij} - d_{ij})^2}{d_{ij}}, \quad (5.5)$$

known as Sammon's stress or Sammon's error.

Figure 5.2 is an illustrative example that shows the effect of applying the classical MDS on the shortest path distances of the weighted visibility graph. The new shape produced within the abstract space is not necessarily convex, but the distances are approximately Euclidean and preserved the original distances as much as possible. Thus, there will be slight deviations between the distances computed from the original space and the distances produced in the embedded space.

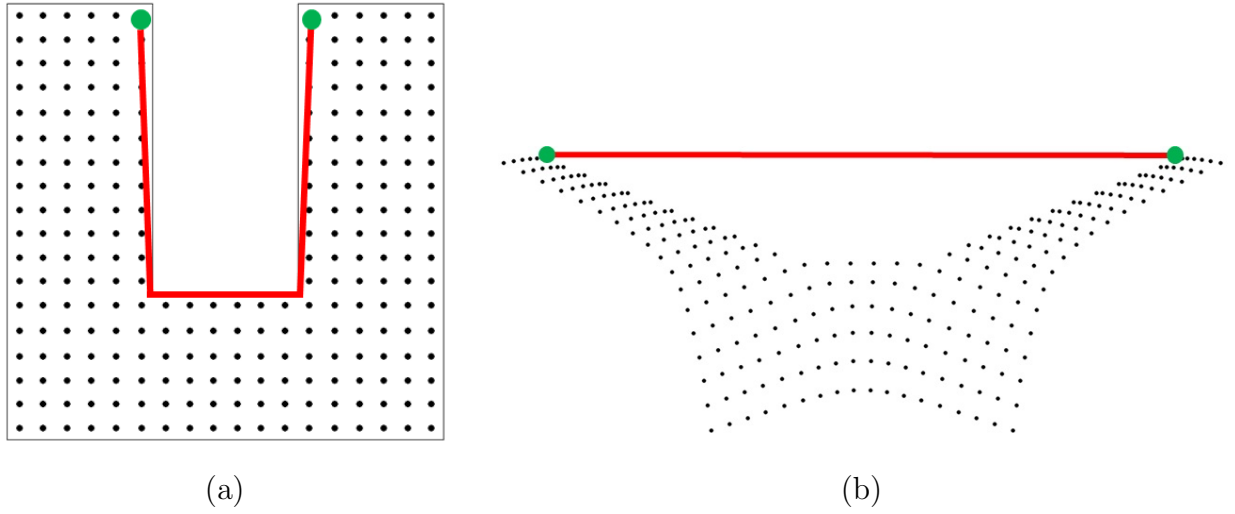


Figure 5.2: An illustration of the result of embedding (a) the shortest path distances on a weighted visibility graph in a Euclidean space. (a) shows the path of the shortest path in the original nonconvex domain and (b) shows the result after applying MDS.

5.4 Methodology

In this section, a methodology is provided for computing spatial measures in a setting where the points in a point pattern are constrained by the boundaries of a nonconvex window polygon.

The MDS procedure works on a finite set of points. To get a discrete representation of the spatial domain, we subdivide the window polygon into a triangular mesh. A mesh is a discretised representation of a geometric domain, often used in computational modelling to approximate complex shapes. Figure 5.3(a) depicts a triangular mesh created on a nonconvex window and Figure 5.3(b) shows the transformed mesh after applying MDS. The centers or vertices of the triangles in the mesh are used as the finite set of points over which the weighted visibility graph is constructed.

The shortest path algorithm is then applied on the resultant weighted visibility graph and the sum of the edge weights on the shortest path used compute the distance matrix. The distance matrix provided from this step is used to embed the points in a Euclidean metric space using metric MDS. The results of applying this procedure to a U-shaped, S-shaped, and irregular window is depicted in Figures 5.4, 5.5, and 5.6. For each windows

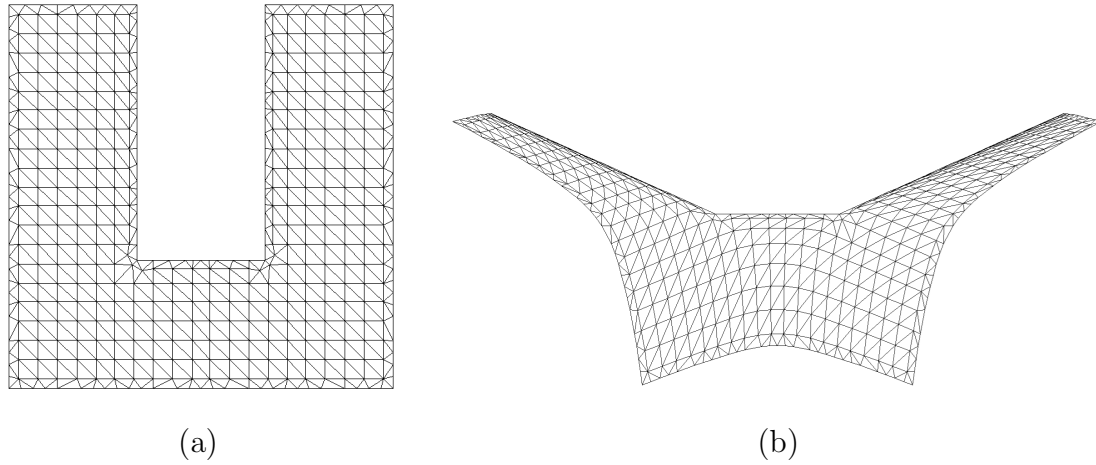


Figure 5.3: A discretisation of a nonconvex polygon using a triangular mesh in (a) and its corresponding representation in Euclidean space after the application of MDS in (b).

a scree plot is shown with the number of dimensions on the horizontal axis and the Stress (i.e., Equation 5.4) on the vertical axis. It can be observed in Figures 5.4(b), 5.5(b), and 5.6(b) that Stress decreases as the number of dimensions increases. There are noticeable “elbows” in the curves, where the decrements in Stress begin to be less pronounced. In the case of the U-shaped window the elbow appears at three-dimensions, the S-shaped window at five-dimensions, and the irregular window at four-dimensions.

A Shepard diagram [11] for the two- and three-dimensional cases are provided in (d) and (e) of each figure respectively. The Shepard diagram is essentially a plot of the weighted visibility graph distances, i.e., the d_{ij} ’s, on the horizontal axis against the corresponding MDS distances δ_{ij} on the vertical axis. The 2D and 3D plots have been oriented so they mostly align with the original window. The line where $d_{ij} = \delta_{ij}$ is shown in red. Observable from these diagrams is the reduction in the spread around the line $d_{ij} = \delta_{ij}$ when the dimensions increase from two to three. This is the typical observed pattern as the number of dimensions increase in all cases we considered. Thus, it is valuable to project to dimensions higher than three-dimensions in terms of optimising precision of estimates based on these distances; even though we may not be able to visual the points in the new space.

Once a decision is made regarding the number of dimensions p in which to project, the observed points $x = \{x_1, \dots, x_n\}$ in the original space, $W \subset \mathbb{R}^2$, are transformed to

Chapter 5. A framework for analysing point patterns on nonconvex domains using visibility graphs and multidimensional scaling 81

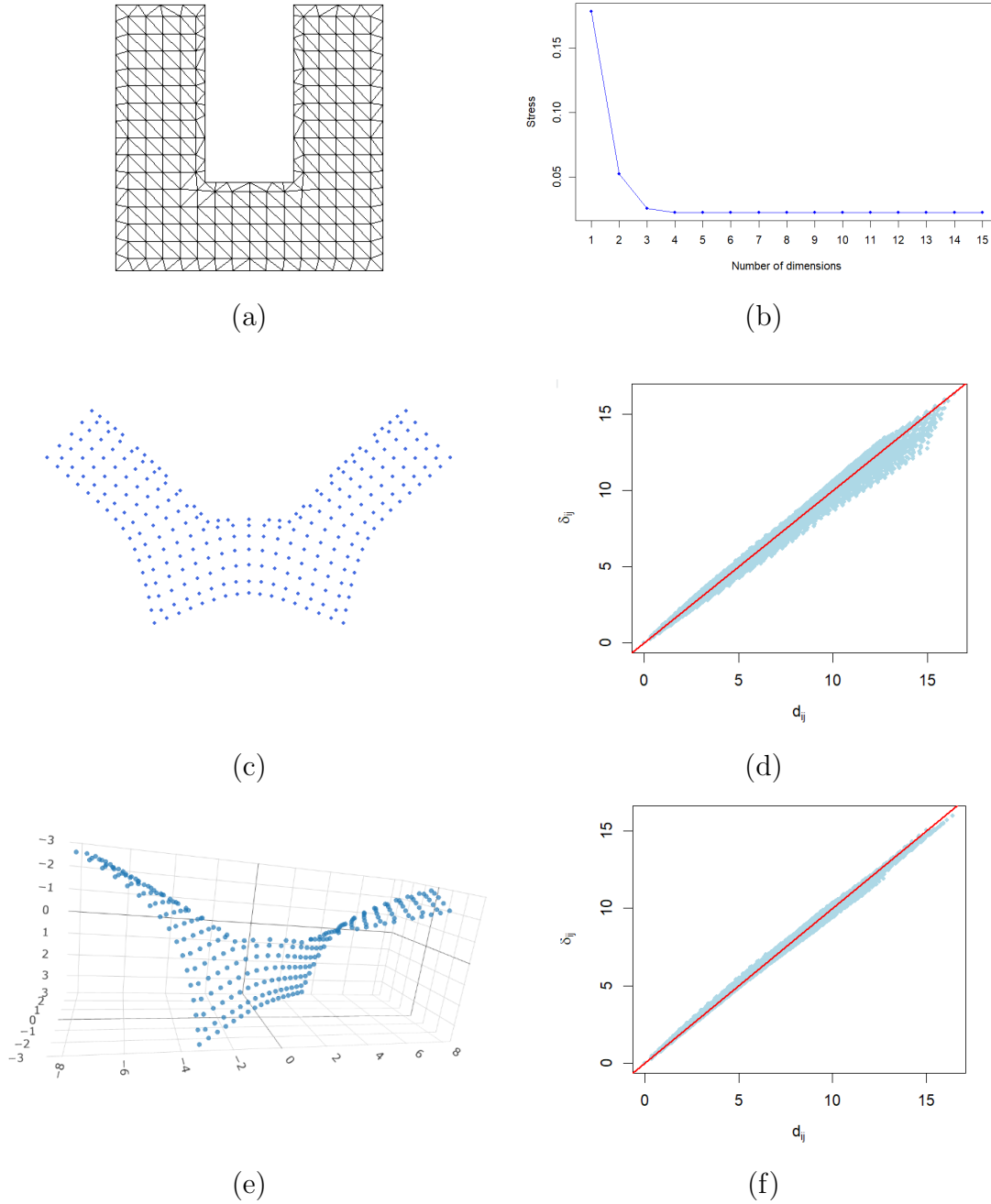


Figure 5.4: The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), a U-shaped window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.

Chapter 5. A framework for analysing point patterns on nonconvex domains using visibility graphs and multidimensional scaling 82

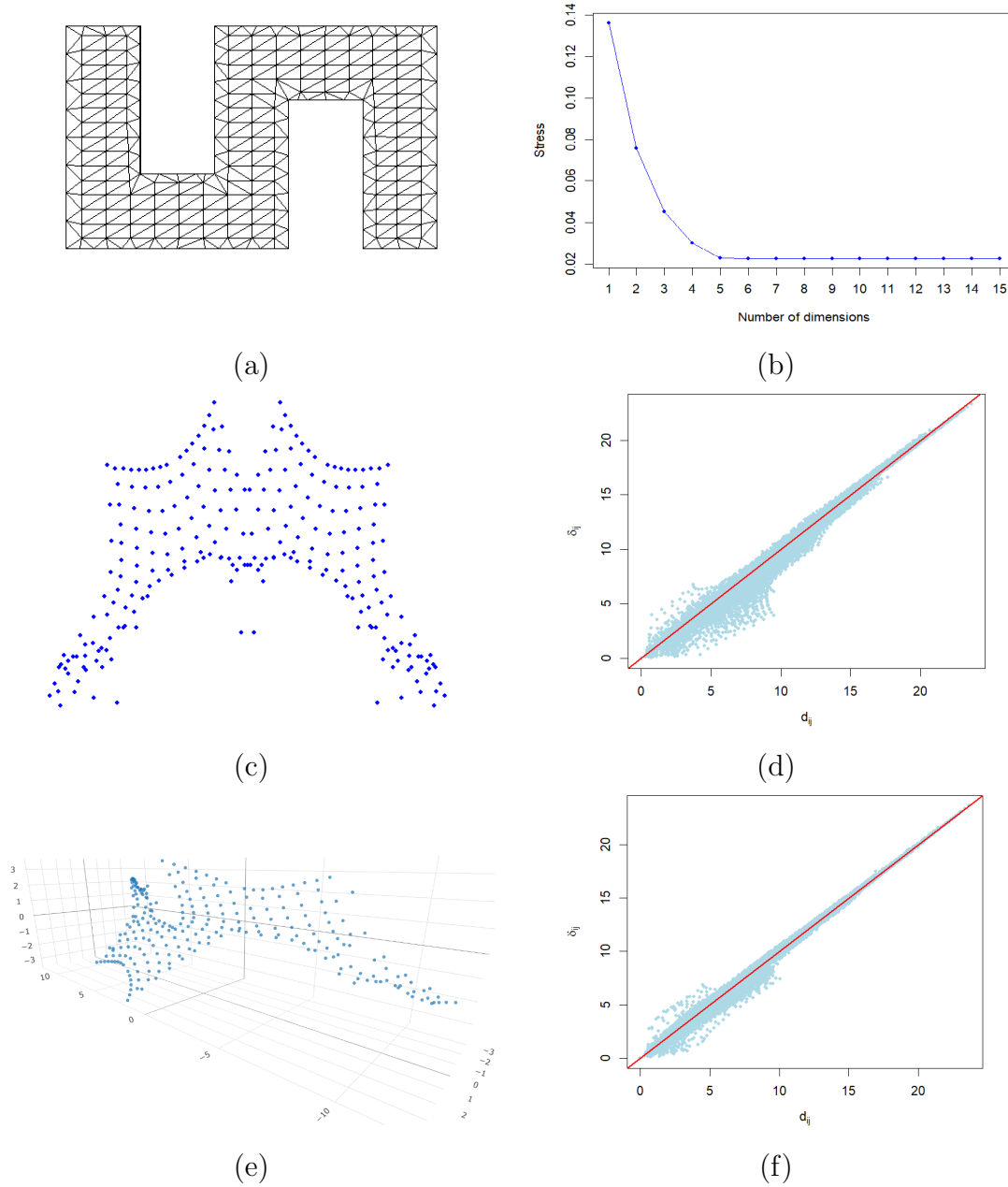


Figure 5.5: The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), an S-shaped window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.

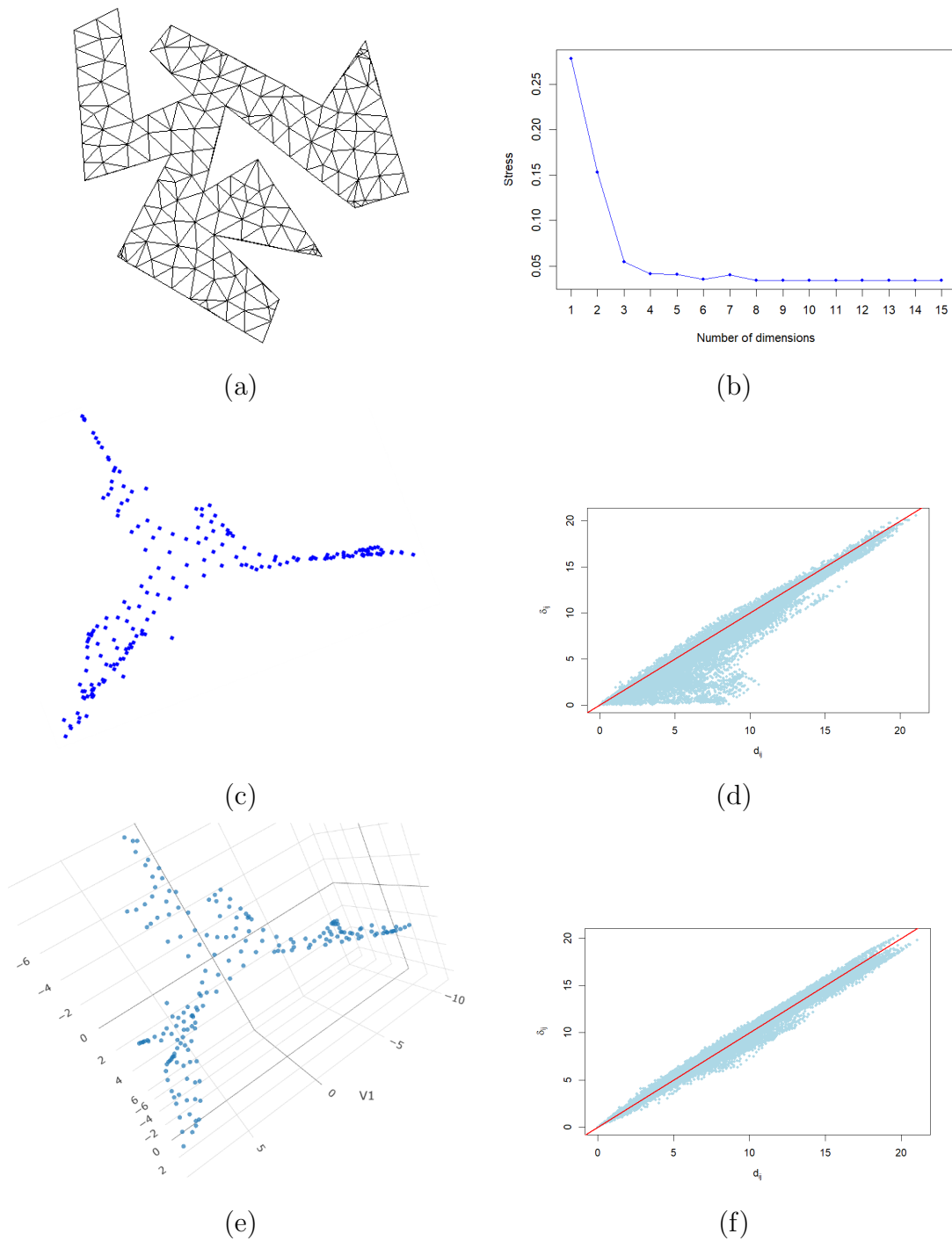


Figure 5.6: The stress determined for metric MDS in different dimensions is shown in (b) for the vertices extracted from the triangular mesh in (a), an irregular window. The representation of the points in two-dimensional and three-dimensional space is provided in (c) and (e) respectively. A plot of d_{ij} against δ_{ij} is provided in each case in (d) for the 2D case and (f) for the 3D case.

the new metric space using the transformation

$$T_l(x_i) = A_l x_i \quad (5.6)$$

where $T_l : \mathbb{R}^2 \mapsto \mathbb{R}^p$ and A_l is a transformation matrix that maps a point x_i from the triangle in which it lies on the triangular mesh of the window W to its counterpart in the new metric space.

To compute the spatial summary measures for the first- and second-order properties of the point pattern, we apply traditional statistical methodology based on the Euclidean distance on this new space and transform it to the original space. An illustration is provided for the kernel smoothed intensity estimate and the K -function.

In this section, we have outlined a framework to embed distances calculated on a weighted visibility graph in Euclidean space. The method starts with discretisation of the original nonconvex window using a triangulation or triangle mesh. The inter-point distances between the vertices on the triangular mesh are computed along the weighted visibility graph. Multidimensional scaling is then applied to the resultant distances to produce a point configuration in a dimensional space admitted by observing the dimension at which the elbow appears on the scree plot. The triangles in the new space are matched with their counterparts in the original space and a local mapping between triangles is performed to preserve local visibility.

5.5 Results

Consider the point pattern depicted in Figure 5.7. The points are realised from a process with inhomogeneous intensity, i.e., there is a higher density of points in the upper right corner of the nonconvex window as opposed to the bottom left corner.

We apply the methodology outlined in Section 5.4 as follows. The window is discretised using an irregular³ triangular mesh. The result is shown in Figure 5.8(a). A weighted visibility graph of the window polygon is constructed. The centers of each triangle in the mesh is then determined and the distance between all pairs of points

³The choice between using a regular or irregular triangular mesh has no observable influence on the result of the final estimates provided

computed using the length of the shortest path in the weighted visibility graph. The resultant distances are embedded in a Euclidean metric space using metric MDS. The result of which was already shown in Figure 5.6. We perform multidimensional scaling using four-dimensions since this was the dimension at which the elbow of the scree plot as observed. For illustrative purposes, the result in two-dimensions is shown in Figure 5.8(b). The new locations of the observed points in the Euclidean metric space are then determined by using a linear mapping from the triangle in which it originally lies to the corresponding triangle in four-dimensional Euclidean metric space. This new domain now forms the basis on which analysis will be performed.

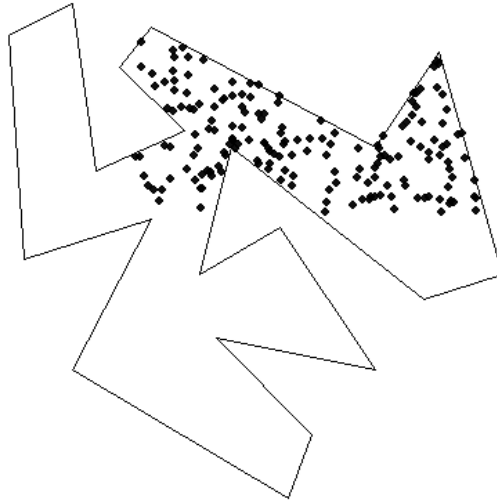


Figure 5.7: A point pattern realised from a point process with inhomogeneous intensity.

In the case of the smoothed estimate of intensity, an estimate at $u \in W$ is determined using the expression,

$$\hat{\lambda}_T(u) = \sum_y \frac{1}{e_\sigma(u)} \kappa_\sigma(\|T_u(u) - T_y(y)\|) \quad (5.7)$$

where $T_y(\cdot)$ and $T_u(\cdot)$ are mappings from the triangle t_y (in which y lies) and t_u (in which u lies), respectively, to the corresponding triangles in Euclidean metric space. The MDS algorithm does not perform the same set of transformation on all the triangles to embed points in Euclidean space, so we do projections between triangles to preserve local visibility since all points in the same triangle are intervisible to each other. The

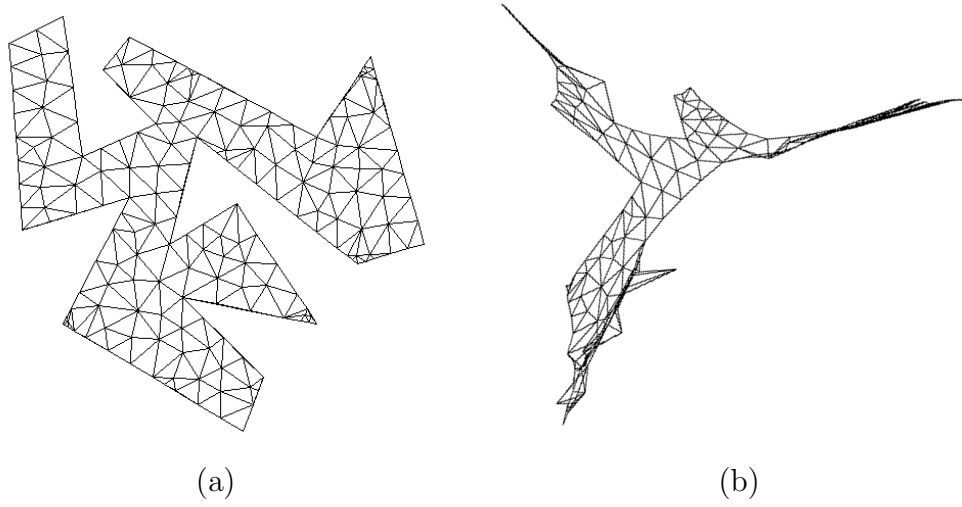


Figure 5.8: A discretisation of a nonconvex polygon using a triangular mesh in (a) and its corresponding representation in Euclidean space after the application of MDS in (b).

result of using this as an estimator of intensity is depicted in Figure 5.9(b) and the result of using Euclidean distance on the original nonconvex window is shown in Figure 5.9(a). For comparison purposes, an uncorrected kernel estimator with a bandwidth value of 1 is used for both estimates. In the case of the estimate computed using Equation 5.7, the influence exerted by a point spreads through a continuous path within the domain. When Euclidean distance is used in this domain, the influence of a point diffuses radially based on a straight line that disregards window boundaries. Thus, the contribution of the point to the intensity estimate remains substantial, even when movement between points is restricted by the domain's boundary.

In the case of the K -function, the empirical function is determined using the expression,

$$\hat{K}_T(r) = \frac{|W|}{n(n-1)} \sum_{i \neq j} I(\|T_i(x_i) - T_j(x_j)\| \leq r) \quad (5.8)$$

where $T_i(\cdot)$ and $T_j(\cdot)$ here again denotes the mapping of the points x_i and x_j from the triangles in which they lie in W to the corresponding triangles in p -dimensional Euclidean metric space.

The right panels of Figure 5.10 show the result of using Equation 5.8 on a clus-

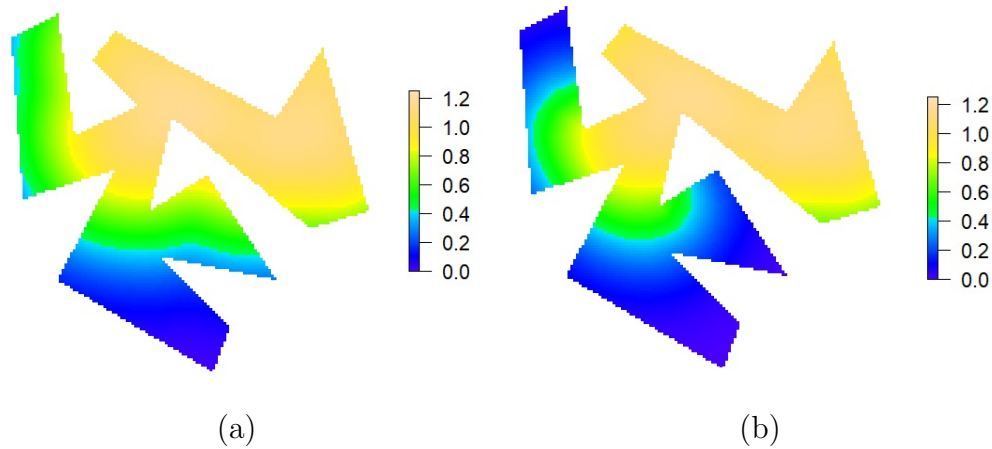
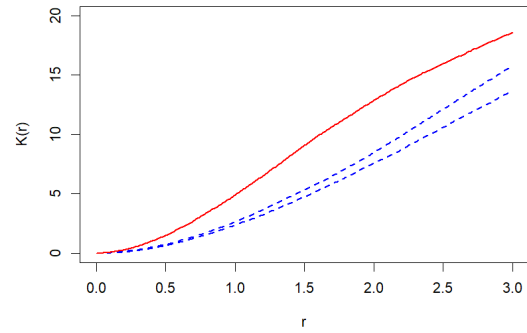
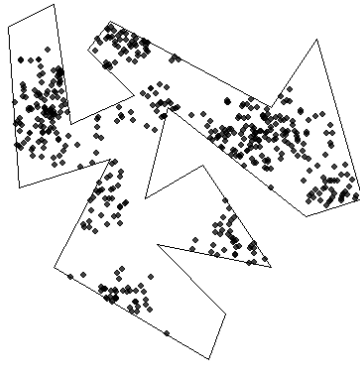


Figure 5.9: A kernel smoothed intensity estimate produced using (a) Equation 5.1 and (b) Equation 5.7.

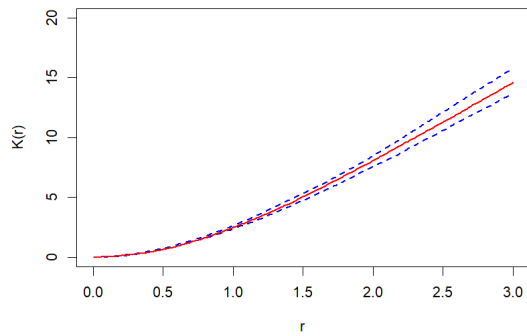
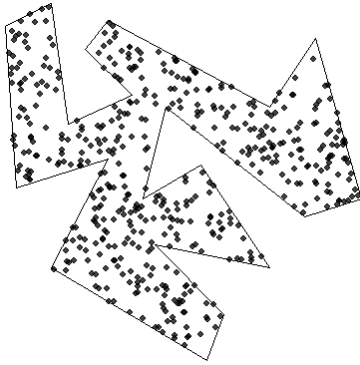
tered, random, and regular point pattern. The corresponding point patterns are given in the left panel of Figure 5.10. Simulation envelopes [9] (blue lines) are computed under the assumption that the pattern is random. The red curve represents the modified empirical K -function values for the given point patterns. In the case of the clustered point pattern, depicted in Figure 5.10(a), the empirical K -function plots above the simulation envelopes. This indicates clustering and demonstrates the functions ability to detect its presence thereof. For the random point pattern in Figure 5.10(b), the empirical K -function is contained within the simulation envelopes. In the case of the regular point pattern shown in 5.10(c), the function plots just below the simulation envelopes illustrating an ability to detect regularity at given scales.

5.6 Discussion

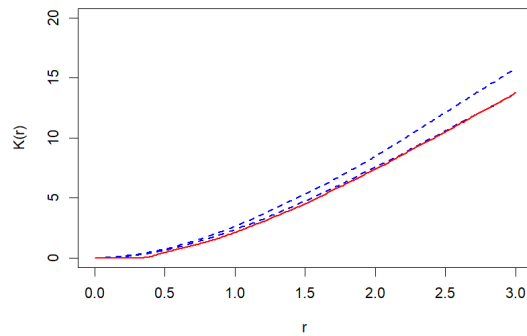
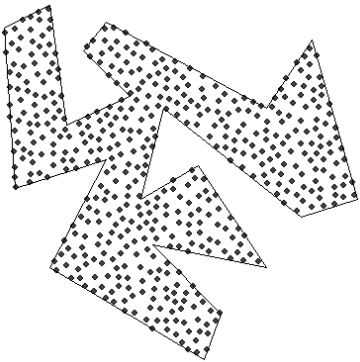
The modified intensity estimate in Figure 5.7 produces an estimate of intensity that more accurately captures the spatial distribution of points contained in the nonconvex window domain. Here, the influence of a point spreads along a connected path within the domain whilst the use of the Euclidean distance results in an underestimation of distance and an incorrect allocation of higher weights to points that cannot be reached by traversing along the straight-line path represented by this distance. The K -function



(a) Clustered point pattern



(b) Random point pattern



(a) Regular point pattern

Figure 5.10: Modified empirical K -function estimates for simulated (a) clustered, (b) random, and (c) regular point patterns.

based on Equation 5.8 and depicted in Figure 5.10 demonstrates an ability to detect departures from randomness, namely clustering and regularity.

In this research, we address the need to adjust point pattern summary measures to accommodate nonconvex spatial domains. The choice of metric used to quantify the proximity of points in a point pattern is important since estimation and prediction are significantly influenced by it. Traditional approaches often make use of the Euclidean distance which assumes an unobstructed linear path between points. This distance measure does not reflect the complexities found in real-world scenarios where nonconvex windows limit the movement of points. In practical applications, point distributions are frequently shaped by the underlying geometric structure of the domain, resulting in more intricate spatial structures.

We have proposed a framework to account for the nonconvex structure of spatial domains using the notion of visibility and a path connected space determined by using the edge weights of the shortest path on a weighted visibility graph. Multidimensional scaling is used to allow the use of traditional statistical methodology that is based on the Euclidean distance. This alternative approach allows for a reduction in computational time that would be required for multiple queries to a visibility graph by preserving local visibility through the use of points in a triangular mesh.

Our findings show that when points are confined to nonconvex spatial domains, kernel-smoothed intensity estimates that apply weights as a function of Euclidean distance do not account for boundaries and gaps within the window. We addressed this issue by using the shortest path distance on the visibility graph, which provides a more appropriate measure of proximity within a connected space on a nonconvex window. In the case of the K -function, we have also obtained good empirical results.

A drawback of the proposed method is its ability to deal with nonconvexity induced by holes in a polygon. In this setting, it is not straightforward to get a Euclidean representation that preserves distances on nonconvex spaces through the application of MDS. Future research will thus extend the current methodology to allow for a Euclidean characterisation that appropriately represents polygons with holes. In the absence of such methodology, the visibility graph can directly be used to compute distances.

5.7 Conclusion

In spatial point pattern analysis, the first- and second-order properties of a point process are traditionally defined using a function of distance between points. The choice of an appropriate metric to quantify proximity between points is important since this has a direct influence on the estimation of the properties of a pattern. When points are observed in nonconvex window domains that constrain their movement, the Euclidean distance is not an appropriate choice for distance since this disregards the obstacles imposed by the window boundaries.

Herein, we have provided a framework to incorporate the restrictions imposed by the nonconvex structure of window domain when determining distance. The problem of finding a representative distance measure that respects the window boundaries and avoids holes in the window domain was solved by determining distance through the use of a weighted visibility graph and multidimensional scaling. This distance gives a more representative measure of proximity on a connected path in a nonconvex domain than the Euclidean distance. We used these ideas to develop a framework for analysing point patterns on nonconvex domains and illustrated their usefulness by using the kernel smoothed estimator and K -function as examples. For future work, we will consider evaluating the robustness of this methodology on different point pattern types and extend the methodology to better deal with polygons with holes.

Chapter 6

Conclusion

Point pattern data arises in a wide scope of application fields. In this thesis we address the methodological challenges inherent in analysing point patterns confined to irregular nonconvex spatial domains. It is customary for spatial structure to be modelled using proximity relationships. Thus the selection of the appropriate window for analysis and a valid distance metric on this window must be afforded particular attention. Nonconvex spatial domains may impede the occurrence of points in given spatial contexts. It is thus important to incorporate the geometric structure of the windows on which points are observed when estimating so called first- and second-order properties. There is methodology that has been developed for non-planar topological spaces such as line networks, but there are methodological challenges in the case of nonconvex irregular planar windows that has not been thoroughly addressed. This gap has motivated this research for which we have made the following contributions.

In Chapter 2, a presentation of novel methodology to estimate nonconvex spatial domains was given. This method is proposed to address the problem of selecting representative nonconvex domains that account for constraints imposed by the geometry of restrictive environments. If a window is unknown and chosen too large, estimation may occur over areas devoid of points. The result is spurious estimation of intensity and inference in areas where the point occurrence of an object or event is unlikely to happen. The method described in Chapter 2 uses a reweighted outlier-scoring scheme that is sensitive to covariate distributions. Robustness was assessed in a simulation study with several

performance metrics used to measure the discrepancy between the estimated window and the simulation extent governed by covariates. Empirical results demonstrated that the window has a strong ability to automatically detect and filter observable features that make the likelihood of a point occurrence improbable. An application in rural Tanzania was considered in which the spatial domain for analysis was estimated.

The restrictions induced by the resultant window geometry necessitated the need to redefine proximity relationships when characterising spatial structures in point pattern analysis. Subsequently, a graph theoretic approach through the use of a visibility graph was proposed to capture proximity relationships within nonconvex geometries. This new characterisation of distance was used in Chapter 3 to extend the use of kernel smoothing, traditionally used in nonparametric intensity estimation, to nonconvex topological spaces that constrain the movement of points. The proposed estimator performs well in incorporating the constraints imposed by nonconvex windows and is a good candidate against its one other known competitor.

In Chapter 4, we considered an extension of second-order quantities. In both the case of the first- and second-order property statistics, we demonstrated the value of using these modified functions as descriptors of point pattern properties when dealing with complex nonconvex geometries.

Finally, in Chapter 5 we provided an implementation strategy that employs localised projections on a triangular mesh and multidimensional scaling to embed visibility-based distances in Euclidean space. Multidimensional scaling is a technique used to represent distances in abstract Euclidean space. In summary this research has contributed the following:

- A novel nonconvex window estimation method that uses distance-based outlier scores of covariates and Otsu thresholding.
- A modified kernel smoothed intensity estimator that incorporates visibility graph-based distance to appropriately capture the nonconvex window geometry constraints.
- A visibility graph-based technique to assess proximity relationships and interactions between points confined to nonconvex domains.

- A framework to analyse point patterns constrained on nonconvex domains using the weighted visibility graph's shortest-path distances embedded in Euclidean space via multidimensional scaling.

Bibliography

- [1] Ian S Abramson. On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics*, pages 1217–1223, 1982.
- [2] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, Jordy van Langen, and Rogier A Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Research*, 4:63, 2021.
- [3] Luc Anselin, Jacqueline Cohen, David Cook, Wilpen Gorr, and George Tita. Spatial analyses of crime. *Criminal Justice*, 4(2):213–262, 2000.
- [4] Luc Anselin, Elizabeth Griffiths, and George Tita. Crime mapping and hot spot analysis. In *Environmental Criminology and Crime Analysis*, pages 119–138. Willan, 2013.
- [5] 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.
- [6] 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.
- [7] Adrian Baddeley, Tilman M Davies, Suman Rakshit, Gopalan Nair, and Greg McSwiggan. Diffusion smoothing for spatial point patterns. *Statistical Science*, 37(1):123–142, 2022.

- [8] Adrian Baddeley, Gopalan Nair, Suman Rakshit, Greg McSwiggan, and Tilman M Davies. Analysing point patterns on networks—a review. *Spatial Statistics*, 42:100435, 2021.
- [9] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, 2015.
- [10] Rémi Boivin and Silas Nogueira de Melo. Comparing global spatial patterns of crime. *Policing: An International Journal*, 42(6):1097–1106, 2019.
- [11] Ingwer Borg and Patrick JF Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media, 2007.
- [12] SN Chiu and D Stoyan. Estimators of distance distributions for spatial patterns. *Statistica Neerlandica*, 52(2):239–246, 1998.
- [13] Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.
- [14] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, 1993.
- [15] 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.
- [16] 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.
- [17] Jon Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Lulu, 2010.
- [18] 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.
- [19] Peter J Diggle, Julian Besag, and J Timothy Gleaves. Statistical analysis of spatial point patterns by means of distance methods. *Biometrics*, pages 659–667, 1976.
- [20] Bradley Efron. The convex hull of a random set of points. *Biometrika*, 52(3-4):331–343, 1965.

- [21] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.
- [22] William F Fagan. Connectivity, fragmentation, and extinction risk in dendritic metapopulations. *Ecology*, 83(12):3243–3249, 2002.
- [23] 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.
- [24] G Gerbier, JN Bacro, R Pouillot, B Durand, F Moutou, and Joel Chadoeuf. A point pattern model of the spread of foot-and-mouth disease. *Preventive Veterinary Medicine*, 56(1):33–49, 2002.
- [25] Subir Kumar Ghosh. *Visibility Algorithms in the Plane*. Cambridge University Press, 2007.
- [26] Subir Kumar Ghosh and David M Mount. An output-sensitive algorithm for computing visibility graphs. *SIAM Journal on Computing*, 20(5):888–910, 1991.
- [27] Jean-Pierre Gilly and André Torre. Proximity relations. elements for an analytical framework. *Industrial Networks and Proximity*, pages 1–16, 2000.
- [28] 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.
- [29] John Hershberger. An optimal visibility graph algorithm for triangulated simple polygons. *Algorithmica*, 4(1):141–155, 1989.
- [30] Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons, 2008.
- [31] M Chris Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

- [32] Yu A Kutoyants. *Statistical Inference for Spatial Poisson Processes*, volume 134. Springer Science & Business Media, 2012.
- [33] Fajie Li and Reinhard Klette. Euclidean shortest paths. In *Euclidean Shortest Paths*, pages 3–29. Springer, 2011.
- [34] Paul Longley and Michael Batty. *Advanced Spatial Analysis: the CASA book of GIS*. ESRI, Inc., 2003.
- [35] Yongmei Lu and Xuwei Chen. On the false alarm of planar K-function when analyzing urban crime distributed along streets. *Social Science Research*, 36(2):611–632, 2007.
- [36] Kabelo Mahloromela, Inger Nicolette Fabris-Rotelli, and Christine Kraamwinkel. Covariate construction of nonconvex windows for spatial point patterns. *South African Statistical Journal*, 57(2):65–87, 2023.
- [37] Kanti V Mardia. Some properties of classical multi-dimesional scaling. *Communications in Statistics - Theory and Methods*, 7(13):1233–1241, 1978.
- [38] Jesper Møller and Rasmus P Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34(4):643–684, 2007.
- [39] Jesper Moller and Rasmus Plenge Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, 2003.
- [40] Marc Moore. On the estimation of a convex set. *The Annals of Statistics*, pages 1090–1099, 1984.
- [41] Mari Myllymäki, Mikko Kuronen, and Tomáš Mrkvička. Testing global and local dependence of point patterns on covariates in parametric models. *Spatial Statistics*, page 100436, 2020.
- [42] J Negre. Non-euclidean distances in point pattern analysis: anisotropic measures for the study of settlement networks in heterogeneous regions. *Mathematics and Archaeology*, pages 369–382, 2015.

- [43] I Newton, Mick Marquiss, DN Weir, and Dorian Moss. Spacing of Sparrowhawk nesting territories. *The Journal of Animal Ecology*, pages 425–441, 1977.
- [44] Atsuyuki Okabe and Kokichi Sugihara. *Spatial Analysis Along Networks: Statistical and Computational Methods*. John Wiley & Sons, 2012.
- [45] Atsuyuki Okabe and Ikuho Yamada. The k-function method on a network and its computational implementation. *Geographical Analysis*, 33(3):271–290, 2001.
- [46] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.
- [47] Mark H Overmars and Emo Welzl. New methods for computing visibility graphs. In *Proceedings of the Fourth Annual Symposium on Computational Geometry*, pages 164–171, 1988.
- [48] Guansong Pang, Kai Ming Ting, and David Albrecht. Lesinn: Detecting anomalies by identifying least similar nearest neighbours. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 623–630. IEEE, 2015.
- [49] Steven L Peck. Quantitative analysis of movement: Measuring and modeling population redistribution in animals and plants. *Ecology*, 80(4):1451–1453, 1999.
- [50] George LW Perry, Ben P Miller, and Neal J Enright. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology*, 187:59–82, 2006.
- [51] Ioannis Pissourios, Pery Lafazani, Stavros Spyrellis, Anastasia Christodoulou, and Myron Myridis. The use of point pattern statistics in urban analysis. In *Bridging the Geographic Information Sciences: International AGILE’2012 Conference, Avignon (France), April, 24-27, 2012*, pages 347–364. Springer, 2012.
- [52] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

- [53] Jean-Paul Rasson, Marcel Remon, 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.
- [54] 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.
- [55] Steven Reader. Using survival analysis to study spatial point patterns in geographical epidemiology. *Social Science & Medicine*, 50(7-8):985–1000, 2000.
- [56] Justin Remais, Adam Akullian, Lu Ding, and Edmund Seto. Analytical methods for quantifying environmental connectivity for the control and surveillance of infectious disease spread. *Journal of the Royal Society Interface*, 7(49):1181–1193, 2010.
- [57] M Rémon. A discriminant analysis algorithm for the inside/outside problem. *Computational Statistics & Data Analysis*, 23(1):125–133, 1996.
- [58] Marcel Remon. 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.
- [59] Mathieu Ribatet. *SpatialExtremes: Modelling Spatial Extremes*, 2022. R package version 2.1-0.
- [60] BD Ripley and J-P Rasson. Finding the edge of a Poisson forest. *Journal of Applied Probability*, 14(3):483–491, 1977.
- [61] Brian D Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266, 1976.
- [62] John W Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100(5):401–409, 1969.
- [63] David W Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.

- [64] Brujal Shah and Satyandra Gupta. Speeding up a* search on visibility graphs defined over quadrees to enable long distance path planning for unmanned surface vehicles. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 26, pages 527–535, 2016.
- [65] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press, 1986.
- [66] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.
- [67] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 2012.
- [68] Radu S Stoica, Elmo Tempel, Lauri Juhan Liivamägi, G Castellan, and Enn Saar. Spatial patterns analysis in cosmology based on marked point processes. *European Astronomical Society Publications Series*, 66:197–226, 2014.
- [69] S Sudarshan and C Pandu Rangan. A fast algorithm for computing sparse visibility graphs. *Algorithmica*, 5:201–214, 1990.
- [70] Chris Sutherland, Angela K Fuller, and J Andrew Royle. Modelling non-Euclidean movement and landscape connectivity in highly structured ecological networks. *Methods in Ecology and Evolution*, 6(2):169–177, 2015.
- [71] Luliang Tang, Zihan Kan, Xia Zhang, Fei Sun, Xue Yang, and Qingquan Li. A network kernel density estimation for linear features in space–time analysis of big trace data. *International Journal of Geographical Information Science*, 30(9):1717–1737, 2016.
- [72] Warren S Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.
- [73] MNM Van Lieshout and AJ1422574 Baddeley. A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361, 1996.

-
- [74] 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.
- [75] Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.
- [76] Gongxing Wu, Incecik Atilla, Tezdogan Tahsin, Momchil Terziev, and LingChao Wang. Long-voyage route planning method based on multi-scale visibility graph for autonomous ships. *Ocean Engineering*, 219:108242, 2021.
- [77] Ikuho Yamada and Jean-Claude Thill. Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2):149–158, 2004.
- [78] 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.