South African Statistical Journal
Vol. 57, No. 2, pp. 65–87
https://doi.org/10.37920/sasj.2023.57.2.1

Covariate construction of nonconvex windows for spatial point patterns

Kabelo Mahloromela, Inger Fabris-Rotelli and Christine Kraamwinkel

Department of Statistics, University of Pretoria, Pretoria, South Africa

In some standard applications of spatial point pattern analysis, window selection for spatial point pattern data is complex. Often, the point pattern window is given a priori. Otherwise, the region is chosen using some objective means reflecting a view that the window may be representative of a larger region. The typical approaches used are the smallest rectangular bounding window and convex windows. The chosen window should however cover the true domain of the point process since it defines the domain for point pattern analysis and supports estimation and inference. Choosing too large a window results in spurious estimation and inference in regions of the window where points cannot occur. We propose a new algorithm for selecting the point pattern domain based on spatial covariate information and without the restriction of convexity, allowing for better estimation of the true domain. A modified kernel smoothed intensity estimate that uses the Euclidean shortest path distance is proposed as validation of the algorithm. The proposed algorithm is applied in the setting of rural villages in Tanzania. As a spatial covariate, remotely sensed elevation data is used. The algorithm is able to detect and filter out high relief areas and steep slopes; observed characteristics that make the occurrence of a household in these regions improbable.

Keywords: Covariate, Euclidean shortest path, Nonconvex, Spatial point pattern, Window selection.

1. Introduction

A spatial point pattern is the mapped locations of objects or events over a region of interest termed a window (Baddeley et al., 2015; Illian et al., 2008). Mapped point pattern data (Illian et al., 2008) usually comprise

- an observation window W,
- the point record of each point x_i in the observation window, $\{x_i, m_{i1}, m_{i2,} ..., m_{ip}\}$, where x_i denotes the coordinates of the observed point location and $m_{i1}, m_{i2,} ..., m_{ip}$ are marks (i.e. observations made on a set of random variables) collected at x_i, and
- covariate information as a spatial measurement $Z(u)$, $u \in W$. It is customary for $Z(u)$ to describe a continuous variable defined at every point in W, whereas marks are only given at the observed points.

Corresponding author: Kabelo Mahloromela (mahloromela.kabelo@gmail.com)
MSC2020 subject classifications: 62GH11, 62P99

A typical aim of spatial point pattern analysis is to expand basic understanding of the first-order and second-order properties of the underlying spatial point process that generated the pattern and extrapolate results to regions where observations have not been made (Bailey and Gatrell, 1995). First-order properties describe how the average number of points varies over different locations in the window. Second-order properties describe how the process values are correlated in space (Bailey and Gatrell, 1995). Information regarding the window W and the point pattern locations $\{x_i\}$ in this window are required for an appropriate analysis of the point pattern. When covariate information is available, an investigation into the dependence of a point pattern on covariate data should be conducted and any dependence quantified using the available techniques in (Baddeley et al., 2015, 2010, 2012; Myllymäki et al., 2020).

In some applications, the selection of W is a non-trivial task. The window W is known when a local phenomenon is studied. Examples include crimes in a city, pores of a metallic foam in a pipe, fungal spots on leaves, fish in a lake, or cell centres in a small biological organ (Illian et al., 2008). Otherwise, it must be chosen using some objective means that reflects a view that it is representative of a larger region, or based on a probability sampling method (Diggle, 2013).

Selection of W must be made carefully; a large choice for W may result in spurious estimation and incorrect conclusions about the first- and second-order properties of the point pattern. The description of first- and second-order properties of point pattern data will rely implicitly or explicitly on the specification of W, i.e. when computing global density measures based on the area of W, when correcting for edge effects in intensity estimation (Baddeley et al., 2015; Diggle, 1985), and when computing distance to quantify spatial dependence and correlation (Baddeley et al., 2015).

In real world applications, sampled points may be defined over an unknown irregular window that has a disconnected or a nonconvex domain. These more general windows result from factors that may make it impossible to observe points in certain regions. Consider the following example. Figure 1 (Dare and Barry, 1990) depicts the distribution of buzzard nests in two upland regions, Snowdonia and Migneint-Hiraethog, in North Wales, UK. Open and closed dots indicate nesting sites outside and inside the study region respectively. The two upland regions are separated by a dashed line and horizontal hatching is used to indicate terrain unsuitable for nesting. In areas unsuited for nesting, points cannot be observed. A similar example can be found in a paper by Newton et al. (1977), based on a study of spacing of territories of *Accipiter ninus*, the sparrowhawk, in twelve areas of Britain. Woodlands suitable for nesting were often separated by large areas of unsuitable woodland, farmland or other open country. The position of the border of the study areas presented a problem as some areas in the woodland were mainly adjacent to rivers and streams, so that estimates of density and habitat proportions were influenced by the amount of open land included on either side. The boundary of the study regions were marked arbitrarily by making a line around the outermost nesting territories at a distance from their centres equal to half the mean inter-territory distance in continuous nesting habitat.

Literature detailing the selection of spatial window domains typically rely on an assumption of convexity of the domain. Ripley and Rasson (1977) proposed a method for window selection through the reconstruction of a compact convex set D from a realisation of a homogeneous planar Poisson process observed over this unknown set D. Given a set of observations from a realisation of a homogeneous planar Poisson point process of unknown intensity within a compact convex set D, they present a technique to determine the compact convex set. Their proposed solution is the dilation

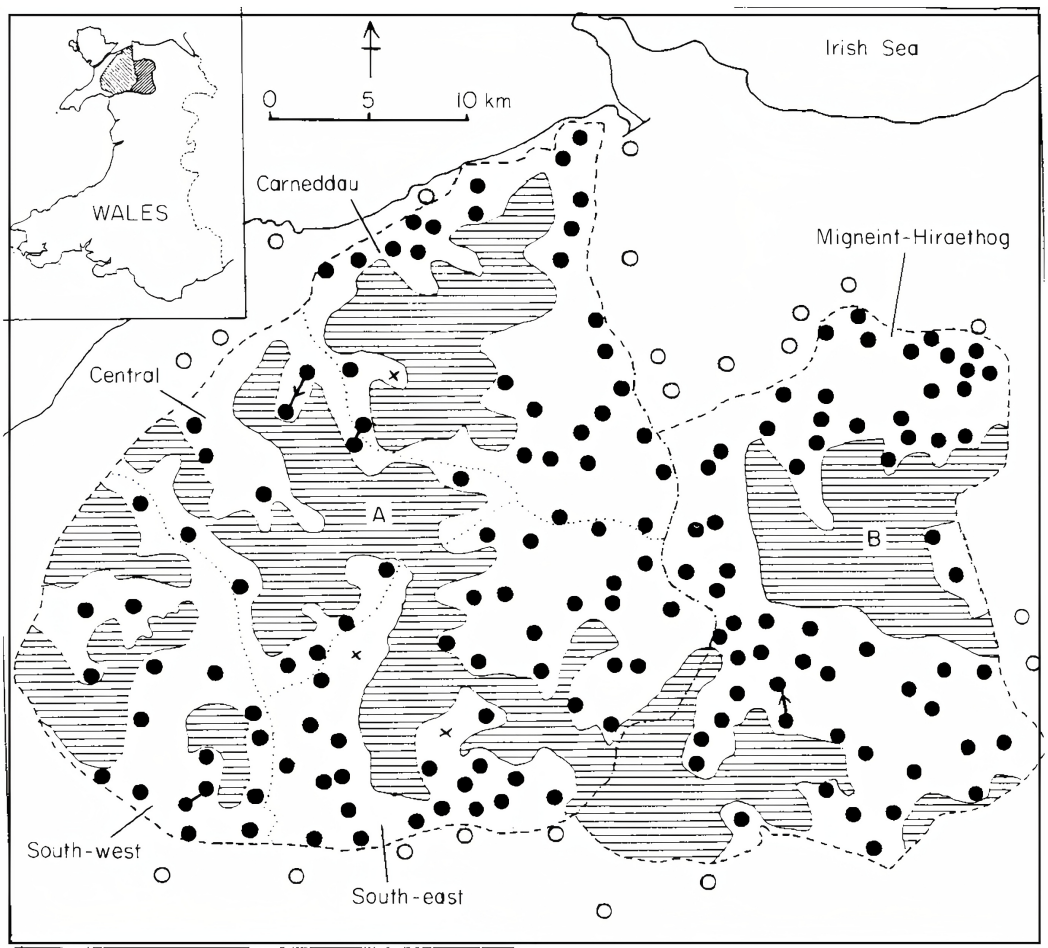


Figure 1. Distribution of buzzard nesting territories.

of the convex hull about its centroid. Moore (1984) presents a method for solving a similar problem, namely estimate D, a compact convex set in $\mathbb{R}^p$, given independent observations $x_1, \ldots, x_n$ sampled uniformly from an unknown D. Rasson et al. (1994, 1996) (also see Remon, 1994) consider how to extend these methods to estimate convex sets with observations that are inside and outside the convex set. These estimates are derived under the assumption of a homogeneous Poisson process and a convex domain; assumptions that may not hold in practice. These estimates also do not incorporate the effect of covariates on the occurrence of a point. Other literature that addresses the determination of a convex hull or convex set from a random set of points include Efron (1965); Dattorro (2010); Phelps (1957). The most common approaches for window selection are the smallest rectangular bounding box and the convex hull.

The objective of this paper is to present a new algorithm to determine nonconvex window domains for spatial point pattern data using covariate information and to illustrate how the proposed nonconvex window aids in improving estimation. In particular, we give focus to the method of intensity

estimation which is used to characterise the first order properties of a spatial point pattern as a validation of the proposed algorithm. This is only one example of a method that is changed by a nonconvex window. To this end, we have organised the paper as follows: In Section 2, we detail the proposed nonconvex window selection algorithm and discuss how the kernel smoothed intensity estimator can be modified for analysis on nonconvex window domains. In Section 3, the nonconvex window construction algorithm is applied to households in a rural setting in Tanzania's Mara province[1] and elevation data from a Digital Elevation Model (DEM) used as a spatial covariate. Section 4 is dedicated to a discussion of the results and Section 5 concludes with a discussion for prospective future work.

2. Methodology

In this section, we propose an algorithm for the construction of a representative nonconvex window using covariate information. A modified kernel smoothed intensity estimator is also presented that makes use of a visibility graph to allocate kernel weights and is provided as an indication of the appropriateness of the constructed nonconvex window.

2.1 Covariate construction of nonconvex window algorithm

The algorithm we propose herein considers the construction of a representative nonconvex window domain that is constructed by incorporating the effect of the covariate on the occurrence of a point; assuming that covariate information is available. 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; see Baddeley et al. (2015, 2010, 2012).

The aim of the algorithm is to select a window domain filtered of all the areas for which it is known that the occurrence of a point is equal to or close to zero, and to determine this based on covariate information. The steps are presented in Algorithm 1, discussed below, and illustrated in Figure 2.

Suppose $x_1, \ldots, x_N$ denote the set of points observed from the unknown, not necessarily convex, domain $W \subset \mathbb{R}^2$ and let $W^* \supset W$. Without loss of generality we suppose that W^* is the smallest bounding rectangular window that contains all the observed points, as shown in the top-left panel of Figure 2. The Minkowski distance (Li and Klette, 2011; Longley and Batty, 2003) between $x = (x_1, x_2)$ and $y = (y_1, y_2)$ is given by

$$||x - y||_p = (|x_1 - y_1|^p + |x_2 - y_2|^p)^{\frac{1}{p}}. \tag{1}$$

For $p = 1$ and 2, the distance is the Manhattan and Euclidean distance respectively. As p tends to infinity, the distance is known as the Chebyshev distance where $||x||_\infty = \max\{|x_1|, |x_2|\}$.

In Step 1 of the algorithm, a moving window M_i at position x_i, given by

$$M_i = \left\{ u \,\middle|\, ||x_i - u||_\infty \leq cd \right\}, \tag{2}$$

[1] This research, entitled *Covariate construction of nonconvex windows for spatial point patterns*, was approved by the Faculty of Natural and Agricultural Science Research Ethics committee at the University of Pretoria under the reference NAS33/2019.

is created, where $d = \min_{i \neq k}\{||x_i - x_k||_2\}$ and $c > 0$ is a positive real number. To choose the value of c, the moving window M_i with varying values of c is moved to each observed point x_i. Points contained in the moving window are systematically chosen[2] and the covariate values at each point evaluated. The values of the covariates are then used to calculate the coefficient of variation. The maximum value of the coefficient of variation is then plotted against the moving window size. The moving window size is then chosen subjectively as the smallest distance before the distance at which a relatively large change in the value of the maximum occurs. The aim in selecting the moving window size in this way is to find the scale c that represents the least variation in covariate values for neighbourhoods of the observed points. Let this appropriate side length be denoted by g. This step is illustrated in the top panel of Figure 2.

In Step 2 of the algorithm, after settling on a value for g, the covariate values at points in the moving window are evaluated and statistically summarised using the variance[3]. Let f_i denote the computed value of the summary measure based on the covariate values in M_i.

In Step 3 we define a set F based on the f_i values. That is, let

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

In Step 4 of the algorithm, the moving window is centred over locations $u_j \in W^*$, $u_j \notin \{x_1, ..., x_n\}$, and is denoted by

$$M_j^* = \left\{ u \,\middle|\, ||u_j - u||_\infty \leq g \right\}. \tag{3}$$

This corresponds to superimposing a grid over W^* with a grid cell size of g.

In Step 5 of the algorithm, the moving window visits each cell and computes a measure based on the covariate values in the cell. Let f_j^* denote the computed summary value based on covariate evaluated at points in M_j^*.

In Step 6 of the algorithm, the true domain W is obtained as

$$W = \bigcup_{\forall j \ni f_j^* \in F} M_j^*.$$

The algorithm uses the moving window to search through overlapping grid cells in the larger domain W^* and, based on a threshold value that conditions on the covariates in the neighbourhood of the observed data points, filter out the regions that are outside the threshold. The thresholds are strictly determined based on the covariate values in the neighbourhood of the observed data points.

[2] The points are chosen such that they span the area covered by the moving window. This can be achieved by selecting points $u = x_i \pm \frac{d}{b}(s_1, s_2)$ where $b, s_1, s_2 \in \mathbb{N}$.

[3] The mean can also be used as a summary measure if the spatial covariate is a surrogate variable that represents a rate of change of an original variable.

Algorithm 1 Covariate construction of nonconvex window

1. Create $M_i = \{u | \, ||x_i - u||_\infty \leq cd \ \}$, where x_i is an observed point and $d = \min_{i \neq k}\{||x_i - x_k||\}$ and $c > 0$ is a positive real number.

2. Select points in M_i, evaluate the value of the covariate at the selected points and use to compute f_i, where f_i is the summary value based on covariates evaluated at points in M_i.

3. Let $F = \{f | \min_i\{f_i\} \leq f \leq \max_i\{f_i\}\}$.

4. Create $M_j^* = \left\{u | \, ||u_j - u||_\infty \leq g\right\}$, where $u_j \in W^*$, $u_j \notin \{x_1, ..., x_n\}$, and g denotes the moving window size.

5. Select points in M_j^*, evaluate the value of spatial covariate at the selected points and use values to compute f_j^*, where f_j^* is the summary value based on covariates evaluated at points in M_j^*.

6. True domain: $W = \bigcup_{\forall j \ni f_j^* \in F} M_j^*$.

2.2 Modified kernel smoothed intensity estimator

The use of nonconvex windows constructed using the algorithm in Section 2.1 necessitates the use of a norm alternative to the Euclidean distance; since this distance does not respect window boundaries. In the case where movement between points is constrained by a connected physical path on the window, it is important to define distance using a measure that is representative of this path since these will affect the estimates used to characterise first- and second-order properties of the pattern.

This section is dedicated to demonstrating the effect of the chosen window on the kernel smoothed intensity estimate. The role of the proposed modified kernel smoothed intensity estimator is used to provide an indication of the appropriateness of the constructed nonconvex window. There are other intensity estimators that are applicable for nonconvex windows that have attractive statistical properties and these can be used in the final estimation of intensity. The reader is directed to Baddeley et al. (2022) for a discussion of these.

2.2.1 Kernel smoothed intensity estimation

An exploratory step in point pattern analysis involves estimating the intensity function. The intensity function characterises the first-order properties of a spatial point process, describing how the average number of points varies over different locations in space (Gatrell et al., 1996). A query into the properties of the intensity of a process is often the aim of a scientific investigation and is a crucial part of point pattern analysis (Kutoyants, 2012; Moller and Waagepetersen, 2003). For example, in forestry surveys the stand density, defined as the number of trees growing per unit area, is an important quantity to be estimated (Ginrich, 1967).

The intensity function $\lambda(x)$ (Gatrell et al., 1996) at position $x \in W$, where W is the spatial window, is defined as the average number of points per unit area and is denoted by the expression,

$$\lambda(x) = \lim_{dx \to 0} \frac{E(N(dx))}{|dx|}, \tag{4}$$

where dx is a small region around x, $E(\cdot)$ is the expected value operator, $|dx|$ is the area for this region, and $N(dx)$ refers to the number of events in the region dx. A point pattern that has constant intensity over varying locations, $\lambda(x) = \lambda$, is termed first-order homogeneous whereas inhomogeneous point

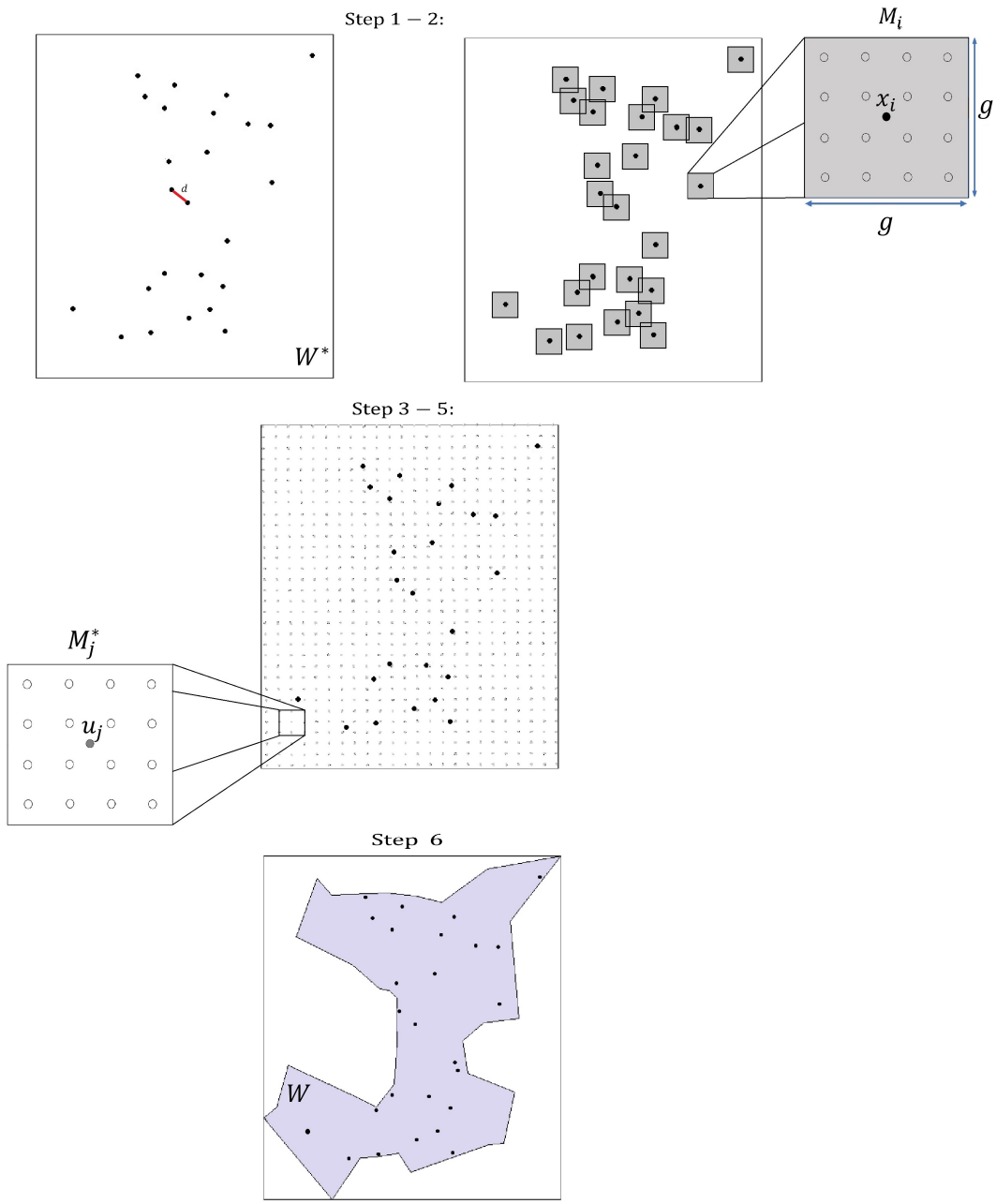


Figure 2. Illustration of the covariate construction of nonconvex window algorithm.

patterns have non-constant intensity functions varying with x. The intensity function can be estimated using the kernel estimator of intensity (Diggle, 1985; Silverman, 1986)

$$\hat{\lambda}(x) = \sum_{i=1}^{n} K_h(x - x_i), \tag{5}$$

where h is the bandwidth, $K_h(x) = h^{-2}K((x - x_i)/h)$, and $K(\cdot)$ is a kernel weighting function, a probability density on $\mathbb{R}^2$. Values of the intensity at certain locations in W are estimated from neighbouring data and the estimates presented. Due to edge effects, the estimator is biased. This can be corrected for by using the uniformly corrected kernel estimator (Diggle, 1985)

$$\hat{\lambda}^{(U)}(x) = \frac{1}{q_h(x)} \sum_{i=1}^{n} K_h(x - x_i), \tag{6}$$

or the Jones-Diggle corrected kernel estimator (Jones, 1993)

$$\hat{\lambda}^{(J)}(x) = \sum_{i=1}^{n} \frac{1}{q_h(x_i)} K_h(x - x_i), \tag{7}$$

where $q_h(x) = \int_W K_h(x - y)dy$ is an edge correction term.

Since the intensity function is estimated on the full extent of W, if W is misspecified, we estimate intensity in areas of the domain for which a point occurrence is non-observable. We show this with the example in Figure 3. The figure depicts the kernel smoothed intensity estimate of a point pattern where point locations (red dots) denote fish in a lake. The point pattern was created by simulating points in a polygon of the Great Bear Lake in Northwest Territories, Canada[4]. The bold irregular boundary represents the separation between lake water and dry land. The estimation is done over a rectangular window domain, the standard smallest bounding rectangle, and fitted with a Gaussian kernel. As seen in the figure, for such a point pattern, estimation of intensity in this window domain will result in the incorrect allocation of positive intensity of fish over areas of dry land. The observation window will obviously be delineated by the boundary separating land and water and would be chosen as such in practice. The importance of making use of covariate information for window determination is clear from this example.

2.2.2 Distance metrics for spatial point patterns

The kernel smoothed intensity estimate in Equation 5 apportions weights based on the Euclidean distance between a point and the point events observed in the point pattern. The expression for the Euclidean distance between two points, $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in $\mathbb{R}^2$, is given by

$$||x - y||_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \tag{8}$$

This distance measure makes two implicit assumptions, (1) the distance is calculated along the shortest physical path formed by the line segment connecting two points in Euclidean space; (2) the

[4] https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-lakes/[Accessed 10 January 2020]

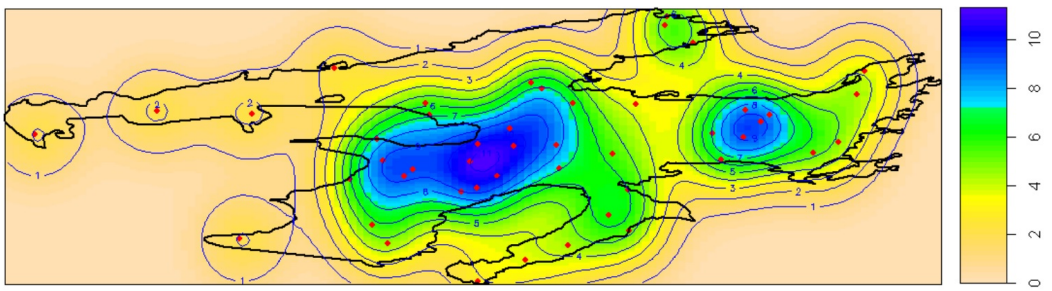


Figure 3. Kernel smoothed intensity estimates of a simulated point pattern of the locations of fish in a lake, estimated on a rectangular window domain and fitted with a Gaussian kernel.

space has no variation in direction and is completely uniform (Longley and Batty, 2003). These assumptions make it possible to use expressions that only require knowledge of the coordinates of the end point locations, thus the actual path between two points is avoided (Longley and Batty, 2003). In cases where the movement of points is constrained by a connected physical path on a nonconvex window, estimating the intensity assuming the Euclidean norm results in a kernel smoothed estimate that does not regard the boundaries of W or holes in this domain.

Consider Figure 4[5], a Google Earth plot of the geographical locations, latitude and longitude, of households in the Hekwe village in Mara province, Tanzania[6]. In Figure 4(b), we see that mountainous regions make it non-viable to move along the blue coloured line segment, indicative of the Euclidean distance, and that a more representative path would follow along the yellow line using a non-Euclidean distance. In instances as shown in Figure 4, this would mean the Euclidean distance does not measure the true distance between points on the physical surface represented by the spatial window. In addition to this, mountainous areas make it likely impracticable to build houses, thus for the given point pattern, intensities should not be registered over these regions.

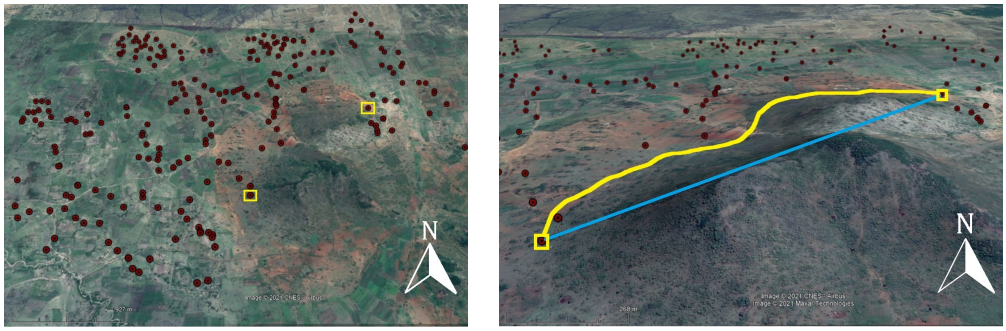
For a convex polygon, the line segment representing the path formed by the Euclidean distance between any two points from the polygon lies entirely in the polygon. On the other hand, a nonconvex polygon has points for which the line segment representing the path formed by the Euclidean distance does not entirely lie within the polygon. An alternative distance metric should thus be chosen for a nonconvex domain that constrain the movement of points.

2.2.3 Path connected distance metric

The problem of finding a representative distance measure between points, one that respects the window boundaries and nonconvex structure, and that avoids holes in the window domain, can be solved by finding the length of the Euclidean shortest path. The Euclidean shortest path (Li and

[5] Google Earth V 7.3.3.7786. (April 14, 2017). Mara province, Tanzania. 1°37'07.33"S, 34°16'26.95"E. Eye alt 5.12km. CNES/Airbus 2021. http://www.earth.google.com[February 18, 2021], Google Earth V 7.3.3.7786. (July 21, 2017). Mara province, Tanzania. 1°37'15.16"S, 34°17'05.01"E. Eye alt 1.96km. Maxar Technologies 2021, CNES/Airbus 2021. http://www.earth.google.com[February 18, 2021].

[6] The 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.



(a) (b)

Figure 4. Google Earth plot of the geographical locations of households (red dots) in Hekwe village in Tanzania's Mara province. (a) Point pattern. (b) Distance between points on adjacent sides of a mountain. The blue line shows the path formed by the Euclidean distance, and the yellow line shows the path formed by a non-Euclidean norm.

Klette, 2011) is the shortest path that can be traversed between two points in a polygon whilst avoiding obstacles (i.e. boundaries and holes). In Euclidean geometry, a path from a point a to a point b (Li and Klette, 2011) , $\mathcal{P}(a, b)$, is a finite sequence of vertices $v_0 = a, v_1, ..., v_{k-1}, v_k = b$; the path starts at the point a and proceeds from vertex to vertex until it reaches b. The Euclidean shortest path is the path between points a and b that has minimum length. The length of path $\mathcal{P}(a, b)$ is given by

$$|\mathcal{P}(a, b)| = \sum_{i=0}^{k-1} ||v_i - v_{i+1}||_2. \tag{9}$$

The Euclidean shortest path is determined by using a visibility graph (Li and Klette, 2011). The visibility graph of a polygon is a graph that connects vertices that are visible to each other. Vertices are visible to each other if the line segment connecting them does not pass through the polygon edges or holes. Each edge of the visibility graph can be labelled with the length of the edge connecting the intervisible vertices of the polygon. A shortest path finding algorithm can then be applied to the visibility graph to find the Euclidean shortest path (Li and Klette, 2011).

Figure 5 shows the Euclidean shortest path (red line) between two points (green dots) in a convex polygon, nonconvex polygon and a polygon with holes. The visibility graph (blue lines) of each polygon is also shown.

An adapted kernel intensity estimator that smooths along the Euclidean shortest path is then available for use and is expressed as

$$\hat{\lambda}(x) = \sum_{i=1}^{n} K_h(|\mathcal{P}(x, x_i)|), \tag{10}$$

where $|\mathcal{P}(x, x_i)|$ is the length of the Euclidean shortest path from point x to point x_i, h is the bandwidth, $K_h(x) = h^{-2}K(|\mathcal{P}(x, x_i)|/h)$, $K(\cdot)$ is a kernel function. The bias-corrected estimators

can also be modified and expressed as

$$\hat{\lambda}^{(U)}(x) = \frac{1}{q_h(x)} \sum_{i=1}^{n} K_h(|\mathcal{P}(x, x_i)|), \tag{11}$$

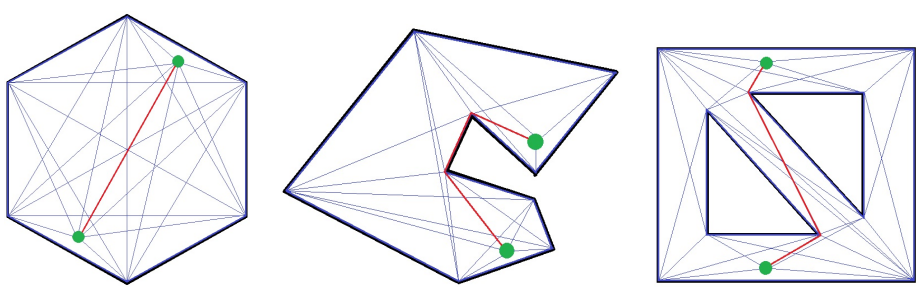
for the uniformly corrected kernel estimator, and

$$\hat{\lambda}^{(J)}(x) = \sum_{i=1}^{n} \frac{1}{q_h(x_i)} K_h(|\mathcal{P}(x, x_i)|), \tag{12}$$

for the Jones-Diggle corrected kernel estimator, where $q_h(x) = \int_W K_h(|\mathcal{P}(x, y)|)dy$. In the case of convex polygons, the Euclidean shortest path distance is equivalent to the Euclidean distance.

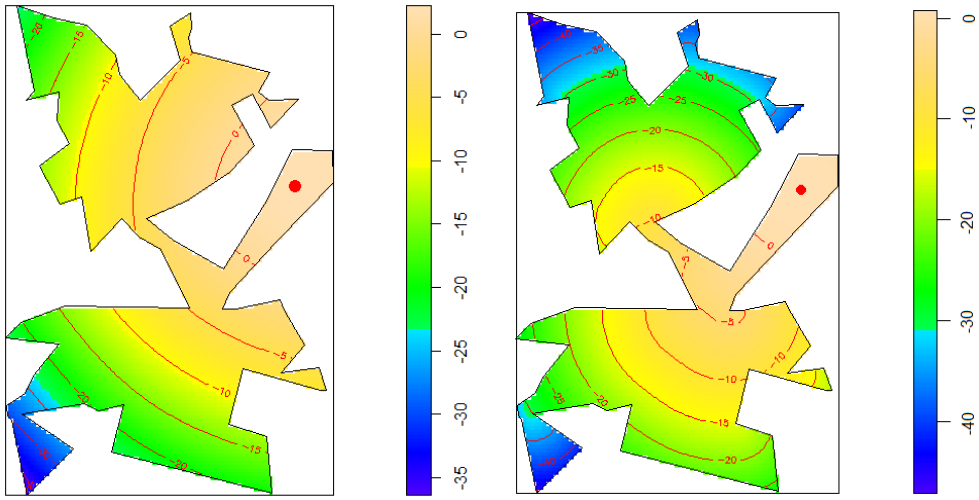
The choice of bandwidth is an important consideration in the estimation of the kernel smoothed intensity function. It is directly responsible for the approximation accuracy. Large bandwidths mask the structure of the data and over-smooth the kernel estimate, whilst small bandwidths under-smooth the data 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. Our goal in providing the proposed modified kernel smoothed intensity estimator is to provide an indication of the appropriateness of the constructed nonconvex window, thus bandwidth selection is out of the scope of what will be considered in this paper. The bandwidth is kept constant for comparisons and consistency. The methods detailed in Diggle (1985) and Baddeley et al. (2015) can be reverted to for the determination of the appropriate bandwidth and the bandwidth effect on the modified kernel smoothed intensity estimate reserved for future work.

Figure 6 illustrates the effect of swapping out the Euclidean distance for the Euclidean shortest path distance on a nonconvex window. Here we see that the weight of influence of the given point diffuses along a connected physical path on the domain. If the Euclidean distance is used as a metric to characterise the distance between two points, the influence of the point event decays linearly as the distance from it increases. Consequently, the contribution of the point event to the final intensity estimate is diffused radially from its centre regulated by the linear distance. This means that the contribution to the kernel smoothed intensity estimate is higher even when the paths between points are constrained by the boundary of the window domain.



(a) Convex polygon (b) Nonconvex polygon (c) Polygon with holes

Figure 5. Euclidean shortest path (red line) between two points (green dots) and visibility graph (blue lines) on three different polygons.

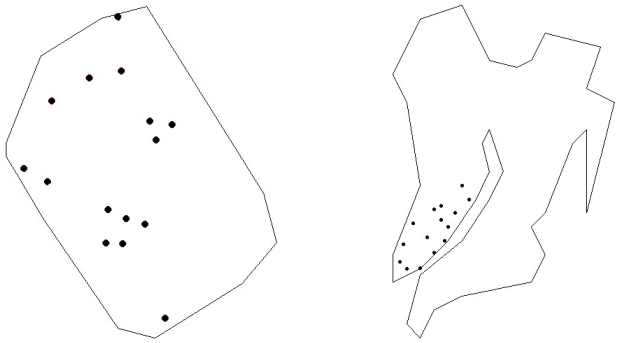


(a) Euclidean distance (b) Euclidean shortest path distance

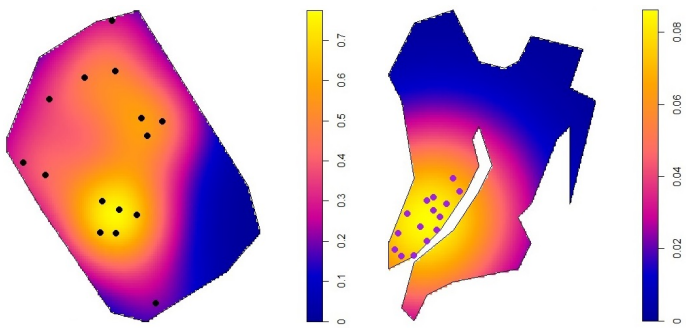
Figure 6. Illustration of the effect of using the Euclidean distance and the Euclidean shortest path distance in estimating the kernel smoothed intensity function shown for a single point event. A Gaussian kernel is used and the influence of a single point plotted on a log scale.

Figure 7 shows the result of applying Equations 5 and 10 to two different point patterns, one on a convex window and one 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 5 and 10 are the same since the Euclidean shortest path distance on this domain will be equal to the Euclidean distance. This can be seen in the left panel of Figures 7(b) and 7(c). In the right panel of Figure 7(b) it can be observed that the intensity estimates on the nonconvex window, calculated using Equation 5, 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 10 shown in the right panel of Figure 7(c), the intensity estimate is more representative of the pattern with its nonconvex window. The result of applying the intensity estimator in the case of the Euclidean distance and Euclidean shortest path distance will follow similarly for the edge-corrected counterparts.

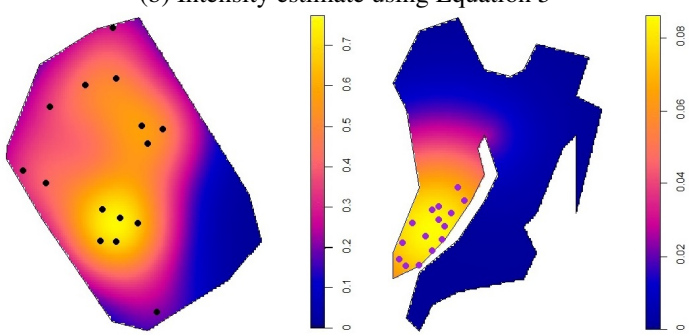
Figure 8 compares the intensity estimates of a point pattern, simulated from a homogeneous Poisson process with $\lambda = 0.5$, computed using the Jones-Diggle corrected intensity estimator in the case when the Euclidean and the Euclidean shortest path distances are used. Estimation is done using a Gaussian kernel and a bandwidth of 3. The pointwise differences illustrated in Figure 8(d) are computed by subtracting the estimated intensity values in Figure 8(b) from the estimated intensity values in Figure 8(c). In areas of the window domain in which points are visible to each other 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.



(a) Point Pattern



(b) Intensity estimate using Equation 5



(c) Intensity estimate using Equation 10

Figure 7. Point pattern plots and intensity estimates on convex (left) and nonconvex (right) windows fitted using Equations 5 and 10, respectively.

3. Results

We now consider an application to rural household locations in Tanzania with significant altitude variation. Algorithm 1, presented in Section 2, is used to construct a window for the locations of houses in an African rural setting. All computations and analysis in this section are done using R Statistical Software R Core Team (2021). Elevation data were processed using the raster package (Hijmans, 2022) and point patterns were created using the spatstat package (Baddeley et al., 2015). Implementation of the algorithm was facilitated by making use of the sp (Pebesma and Bivand, 2005),

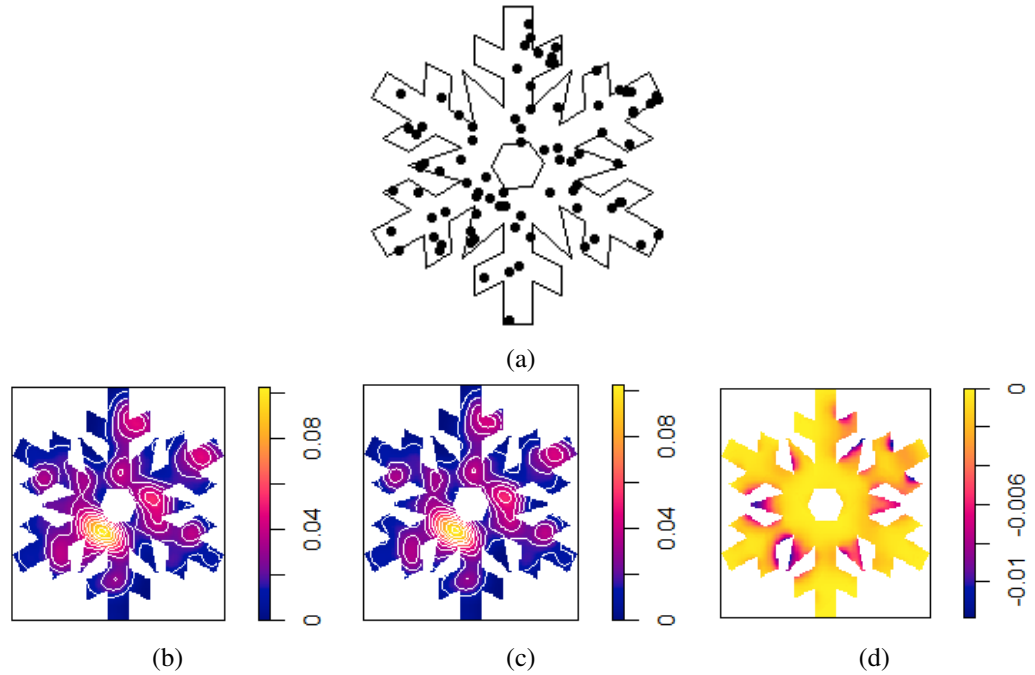


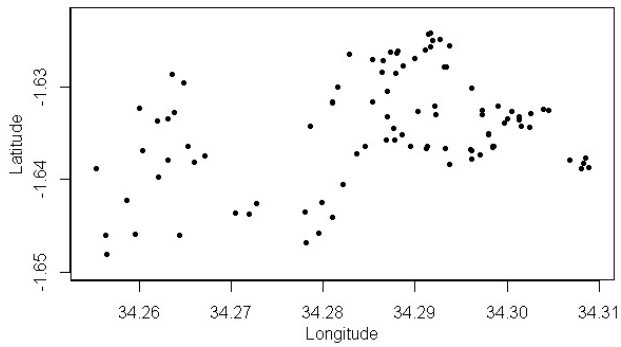
Figure 8. Intensity estimates of simulated point pattern. (a) Point pattern simulated from a homogeneous Poisson process with intensity 0.5. (b) Intensity estimate using Equation 7. (c) Intensity estimate using Equation 12. (d) Pointwise difference between intensity estimate using Equations 7 and 12.

sf (Pebesma, 2018), maptools (Bivand and Lewin-Koh, 2022), and igraph (Csardi and Nepusz, 2006) packages.

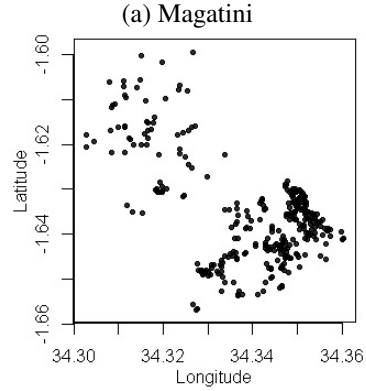
The data used was collected in a census that took place in the Serengeti District, Mara province, situated in Northern Tanzania[7]. The census comprises of georeferenced data for 35 947 households spread across 88 villages. The locations of households are given as latitude and longitude in decimal degree coordinates. Three villages are considered for this paper, namely Magatini, Majimoto and Hekwe with households that number 100, 336 and 235 respectively. The sets of locations of households in the villages form the point patterns. Figure 9 shows plots of the household locations for each village.

Spatial covariate data of terrain elevation for Tanzania is extracted from a Digital Elevation Model (DEM). A DEM, represented as a raster grid, is a matrix of cells, with each cell containing a numeric value representing the elevation (in meters) of the earth's surface above sea level. The grid cells represent a square unit of area and each holds a measurement at sampled points, or estimates at unsampled points, of the terrain elevation value referenced horizontally to a geographic coordinate system. The spatial covariate data were collected in the Shuttle Radar Topographic Mission (SRTM).

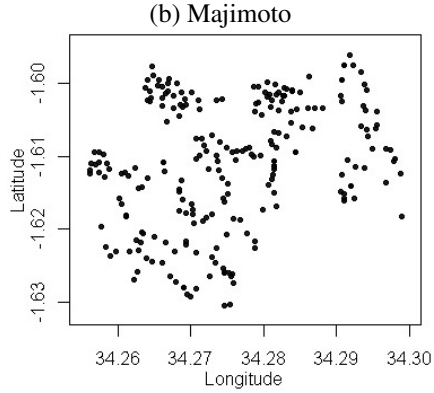
[7] The data was provided by Katie Hampson, http://www.gla.ac.uk/researchinstitutes/bahcm/staff/ katiehampson/, http://www.katiehampson.com/#intro, and approved for use by the Faculty of Natural and Agricultural Science Research Ethics committee under the reference NAS33/2019.



(a) Magatini



(b) Majimoto

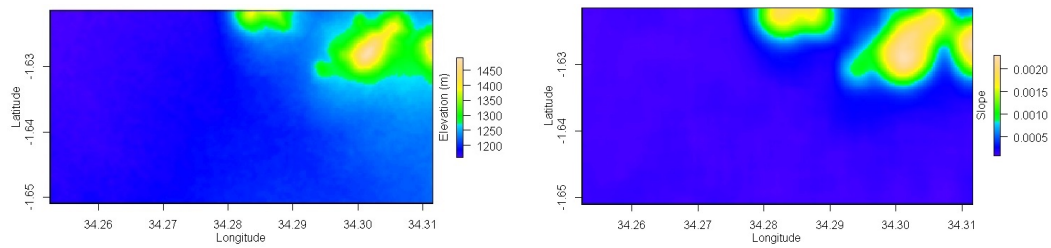


(c) Hekwe

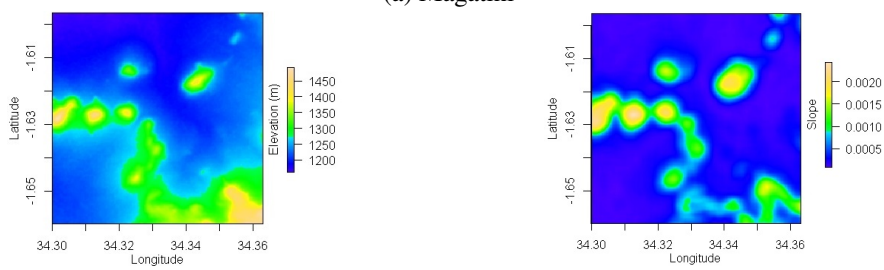
Figure 9. Household locations for three rural villages in Tanzania's Mara province.

The SRTM data were sampled over a grid of 1 arc-second by 1 arc-second (approximately 30m by 30m), with linear vertical absolute height error of less than 16m.

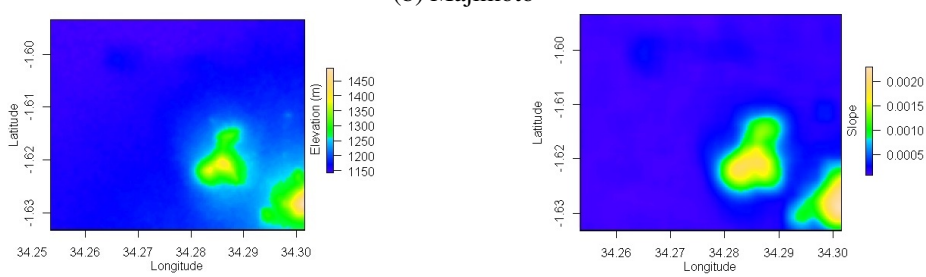
The selection of sites for human settlement in a rural setting is typically guided by some attractive and restrictive traits of the natural landscape. These may include features such as the proximity to water and food, population density and usable land for agriculture and building. Elevation and slope are properties of terrain that influence the distribution of environmental phenomena and the nature



(a) Magatini



(b) Majimoto



(c) Hekwe

Figure 10. Terrain elevation (left) and terrain slope (right) for villages in Tanzania's Mara province.

of environmental processes (Ravibabu and Jain, 2008). 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 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 point occurrence of a household can only be observed in areas that have terrain viable for building. A viable area for building can be characterised by terrain features observed in the neighbourhood of the points in the pattern expressed as the average slope. In places not suitable for building, we expect steeper slopes outside the range of values considered viable for house settlements. Other considerations such as the ease of access to the property have a large influence in determining whether the land is usable. For instance wide, flat, areas elevated high above ground and surrounded by steep ascents are less accessible due to the effort required to move along steep inclines surrounding the landscape.

The terrain slope is computed from the elevation data according to Horn (1981) through the terrain function in the raster package in R (Hijmans, 2022). Gaussian blurring is applied to the computed terrain slope surface to smooth the surface roughness. A 3×3 structuring square kernel is used for this process. The left and right panels of Figure 10 show the terrain elevation and terrain slope for the selected villages respectively.

The relative intensity (Baddeley et al., 2015, 2012) as a function of elevation and terrain slope is calculated for each of the villages as an exploratory step. This is done to confirm the dependence of household locations on the spatial covariates[8]. Figure 11 shows the estimated relative intensity for household locations for each village against terrain elevation and terrain slope. The estimates in each case are modelled relative to the fitted kernel smoothed intensity estimate with no covariate effects. A 95% confidence band is shown in the figure and is computed assuming an inhomogeneous Poisson point process. The estimation was done using the spatstat package Baddeley et al. (2015). In the figure, we see that the relative intensity values against terrain slope, relative to the fitted model with no covariate effects, tend to have larger departures from the value of one than the relative intensity against terrain elevation relative to the fitted model with no covariate effects. This indicates a disagreement with the fitted model with no covariate effects and suggests a possible dependence of the household locations on terrain slope. The figures based on terrain elevation have relative intensity values that are fairly close to one, suggesting less dependence of household locations on terrain elevation implying that elevation by itself is not an appropriate covariate to employ in the window selection process.

Using terrain slope as a covariate, we implement the algorithm to the selected villages as follows:

1. The minimum Euclidean distance[9] between all pairs of points is calculated.

2. A square moving window with side length equal to a multiple of the minimum Euclidean distance between the observed points is defined.

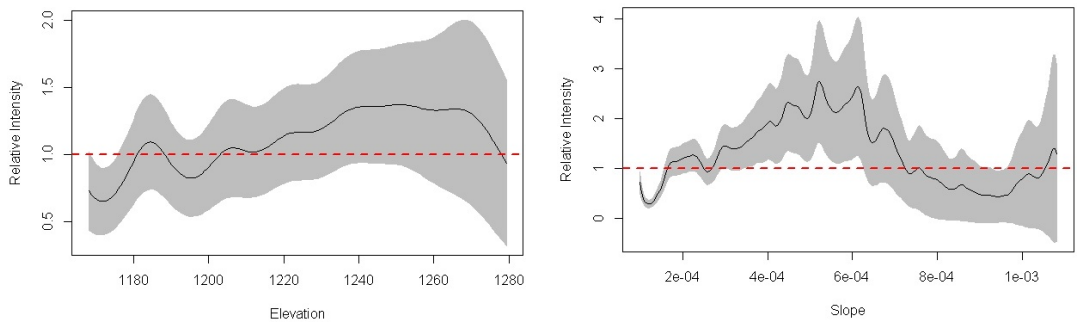
3. The moving window scans the neighbourhood of each point in the pattern and calculates the average slope.

4. A range for these values is used as the selection criterion to determine which quadrats should be included for the window W: a quadrat should be deleted from the window if the average slope is larger than the average slope of values considered viable, as given by the range derived from the observed points.

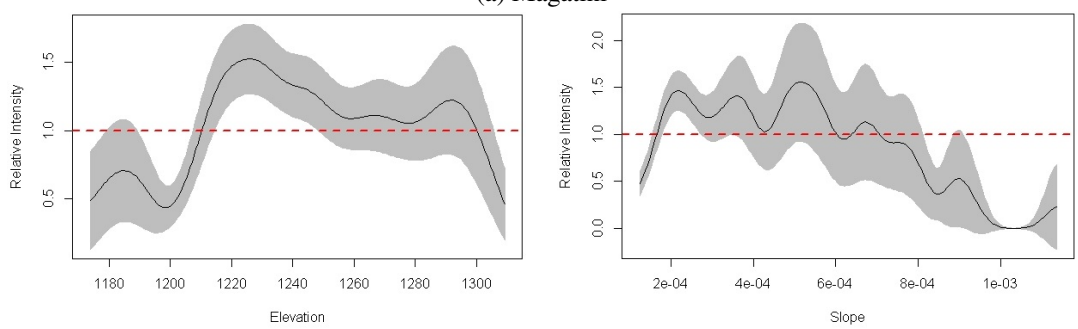
The results of the algorithm for the three villages are shown in the left panel of Figure 12. In each of the villages, the algorithm filters out areas that are outside the range of values determined by the average slope in the neighbourhood of observed locations in the pattern. We observe that the areas identified and excluded by the algorithm in the window construction process are regions with high terrain slope. Regions with terrain slope values that are in the range of covariate values at observed

[8]This is not the final estimation of intensity and is only done here to investigate whether the occurrence of a point is dependent on covariates. The result of this analysis is used to decide which covariate should be used in the window construction process. The literature based on this analysis is well developed and can be found in Baddeley et al. (2015, 2012).

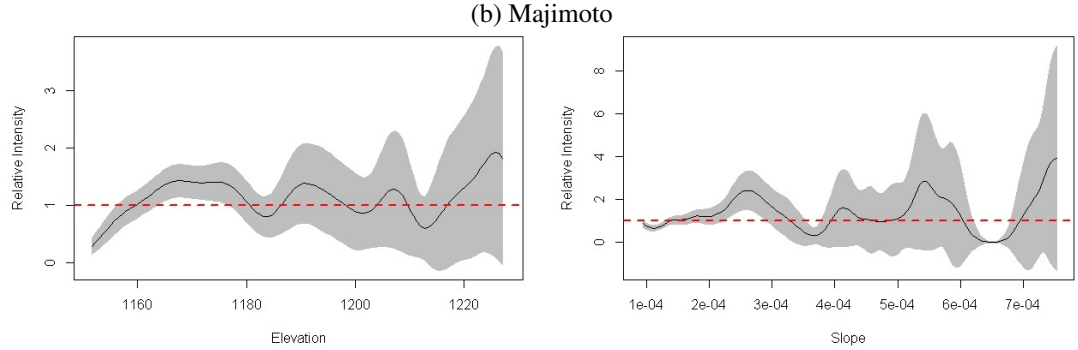
[9]The algorithm uses the Euclidean distance to define of the moving window since the use of the shortest path distance may result in a larger sized moving window.



(a) Magatini

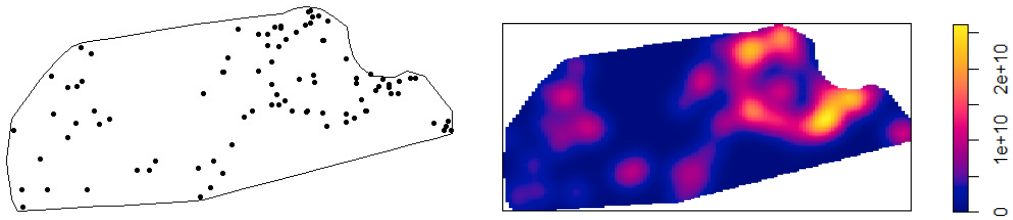


(b) Majimoto

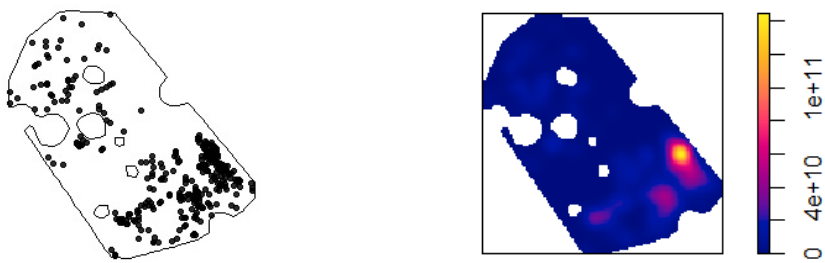


(c) Hekwe

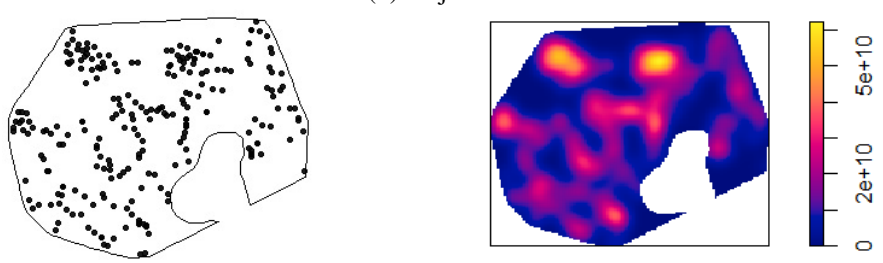
Figure 11. Estimated relative intensity for household locations against terrain elevation (left) and terrain slope (right), relative to the fitted kernel smoothed intensity estimate model with no covariate effects. The dashed red line corresponds to $\rho = 1$ and indicates agreement with the fitted model. The grey shaded area shows a 95% confidence band assuming an inhomogeneous Poisson point process.



(a) Magatini



(b) Majimoto



(c) Hekwe

Figure 12. 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

point locations have been identified by the algorithm as low terrain slope areas. Areas where point locations have not been observed, but which satisfy the condition for viable land, are included in the final window construction. 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.

The right panel of Figure 12 shows the modified kernel smoothed intensity estimate for validation. A Gaussian kernel is used and the Jones-Diggle bias correction term applied. The bandwidth is chosen using the criterion defined by Diggle (1985).

4. Discussion

In this paper we considered spatial window selection techniques for point pattern data. The selection of a spatial window precedes analysis of a point pattern data set and is important since estimation and prediction are dependent on it. Some methods typically rely on the assumption of a homogeneous

Poisson process (Moore, 1984; Ripley and Rasson, 1977) and the assumption of a convex domain, which may not necessarily be true in practice. In real world applications, the distribution of points may be influenced by some underlying process, expressed as a covariate, resulting in more complex spatial windows. When covariate information is available, the dependence of the point pattern on the covariate should be investigated. Parametric models that incorporate this dependence 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 effects (Baddeley et al., 2012, 2015).

We presented an algorithm for the selection of the spatial point pattern domain without the restriction of convexity. The algorithm is applied to village household locations in a rural setting, and elevation data from a DEM is used. In another case 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.

The moving window defined by the algorithm has a side length proportional to the minimum distance of the point observations; chosen such that the number of points in the area centred at the ith observed point is one. A drawback of defining the moving window in this way is that if the observed pattern has points which are very close, it will increase the number of cells that are moved through and thus the computation time.

During the implementation phase of this paper, we found that using Gaussian blurring on the computed terrain slope allowed the algorithm to detect all mountain pixels including mountain peaks. It should however be noted that this does not work if the mountain peak pixels form large wide ridges. In such a case, the mountain peak pixels can be extracted and eliminated from the DEM using the algorithm discussed in Sathymoorthy et al. (2007).

Intensity estimation through the process of kernel smoothing is used as a validation technique to illustrate the effect of the spatial window on intensity estimates. The selected spatial window directly affects the intensity estimate. If a window is chosen too large, estimation will occur over areas for which data has not been observed and where it has not been confirmed that a point can occur. The result is spurious estimation of intensity in void areas where the point occurrence of an object or event cannot happen. We have shown that, when the movements of points is constrained to nonconvex window domains, the kernel smoothed intensity estimate that applies kernel weights along the path formed by the Euclidean distance disregards the window boundaries and holes in the window domain. The problem of finding a representative distance measure that respects the window boundaries and avoids holes in the window domain was solved by finding the length of the Euclidean shortest path. The Euclidean shortest path distance gives a more representative measure of proximity on a connected path in a nonconvex domain than the Euclidean distance. One drawback of using the Euclidean shortest path distance is the computation time of the Euclidean shortest path finding algorithm which increases as the number of vertices of the polygon increases. Optimising the computation time for this algorithm warrants further investigation.

5. Conclusion

Window selection for spatial data is a complex process, most often requiring expert knowledge if not obtained using a data-driven approach, such as herein. The common generic approaches used

are the smallest rectangular bounding window and convex windows due to the use of the Euclidean distance. A chosen window must however cover the true domain of the sampled spatial data in order to facilitate modelling. Here we presented a new algorithm for selecting the spatial point pattern domain without the restriction of convexity. The algorithm works by using a moving window to search over a larger domain than that of the true window. Using a function or feature of the spatial covariate in regions at observed points in the pattern, the proposed algorithm constructs a nonconvex window.

We applied the algorithm in the setting of rural villages in Tanzania's Mara province and used remotely sensed data from a DEM as a covariate. 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 improbable. When the movement between points is constrained to such a nonconvex window, as depicted in Figure 4, 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 the calculation of the kernel smoothed intensity estimate, and in any further spatial analysis. This paper has validated the algorithm using intensity estimation, however the changes in other first- and second-order methods should be investigated in future research as well. This needs to take into account an appropriate metric as shown in this paper.

In future work the algorithm could be extended to allow for an ensemble of spatial covariate effects. One could also investigate ways to refine the selection criterion (function of the covariate), used as a filter to remove regions, and make the identification of this feature more robust and automatic. The definition of the moving window using a non-Euclidean metric or other window definitions such as Voronoi cells also warrants further investigation.

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

References

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., Shah, R., and Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, **4**, 1151–1201.

Baddeley, A., Chang, Y.-M., Song, Y., and Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and its Interface*, **5**, 221–236.

Baddeley, A., Davies, T. M., Rakshit, S., Nair, G., and McSwiggan, G. (2022). Diffusion smoothing for spatial point patterns. *Statistical Science*, **37**, 123–142.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*, volume 413. Longman Scientific & Technical Essex.

Bivand, R. and Lewin-Koh, N. (2022). *maptools: Tools for Handling Spatial Objects*. R package version 1.1-4.
URL: *https://CRAN.R-project.org/package=maptools*

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**, 1695.

Dare, P. and Barry, J. (1990). Population size, density and regularity in nest spacing of Buzzards Buteo Buteo in two upland regions of North Wales. *Bird Study*, **37**, 23–29.

Dattorro, J. (2010). *Convex Optimization & Euclidean Distance Geometry*. Lulu.

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C*, **34**, 138–147.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press.

Efron, B. (1965). The convex hull of a random set of points. *Biometrika*, **52**, 331–343.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, 256–274.

Ginrich, S. F. (1967). Measuring and evaluating stocking and stand density in upland hardwood forests in the central states. *Forest Science*, **13**, 38–53.

Hijmans, R. J. (2022). *raster: Geographic Data Analysis and Modeling*. R package version 3.5-15. URL: *https://CRAN.R-project.org/package=raster*

Horn, B. K. (1981). Hill shading and the reflectance map. *Proceedings of the IEEE*, **69**, 14–47.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*, volume 70. John Wiley & Sons.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, **3**, 135–146.

Kutoyants, Y. A. (2012). *Statistical Inference for Spatial Poisson Processes*, volume 134. Springer Science & Business Media.

Li, F. and Klette, R. (2011). Euclidean shortest paths. *In Euclidean Shortest Paths*. Springer, 3–29.

Longley, P. and Batty, M. (2003). *Advanced Spatial Analysis: The CASA book of GIS*. ESRI, Inc.

Moller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC.

Moore, M. (1984). On the estimation of a convex set. *Annals of Statistics*, 1090–1099.

Myllymäki, M., Kuronen, M., and Mrkvička, T. (2020). Testing global and local dependence of point patterns on covariates in parametric models. *Spatial Statistics*, 100436.

Newton, I., Marquiss, M., Weir, D., and Moss, D. (1977). Spacing of Sparrowhawk nesting territories. *Journal of Animal Ecology*, 425–441.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, **10**, 439–446.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, **5**,

9–13.

Phelps, R. (1957). Convex sets and nearest points. *Proceedings of the American Mathematical Society*, **8**, 790–797.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasson, J., Rémon, M., and Henry, F. (1996). Finding the edge of a Poisson forest with inside and outside observations: The discriminant analysis point of view. *In From Data to Knowledge*. Springer, 94–101.

Rasson, J.-P., Remon, M., Kubushishi, T., and Henry, F. (1994). 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.

Ravibabu, M. V. and Jain, K. (2008). Digital elevation model accuracy aspects. *Journal of Applied Sciences*, **8**, 134–139.

Remon, M. (1994). The estimation of a convex domain when inside and outside observations are available. *Supplemento ai Rendiconti del Circolo Matematico di Palermo*, **35**, 227–235.

Ripley, B. and Rasson, J.-P. (1977). Finding the edge of a Poisson forest. *Journal of Applied Probability*, **14**, 483–491.

Sathymoorthy, D., Palanikumar, R., and Sagar, B. (2007). Morphological segmentation of physiographic features from DEM. *International Journal of Remote Sensing*, **28**, 3379–3394.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press.

Manuscript received 2023-03-02, revised 2023-07-22, accepted 2023-07-26.