A generic similarity test for spatial data

by

René Kirsten

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Department of Statistics

In the Faculty of Natural and Agricultural Sciences

Univeristy of Pretoria

September 2020

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

Signature:

Date:

Summary

Two spatial data sets are considered to be similar if they originate from the same stochastic process in terms of their spatial structure. Many tests have been developed over recent years to test the similarity of certain types of spatial data, such as spatial point patterns, geostatistical data and images. This research develops a similarity test able to handle various types of spatial data, for example images (modelled spatially), point patterns, marked point patterns, geostatistical data and lattice patterns. The test consists of three steps. The first step creates a pixel image representation of each spatial data set considered. In the second step a local similarity map is created from the two pixel image representations from step one. The local similarity map is obtained by either using the well-known similarity measure for images called the Structural SIMilarity Index (SSIM) when having continuous pixel values or a direct comparison in the case of discrete pixel values. The calculation of the final similarity measure is done in the third step of the test. This calculation is based on the S-index of Andresen's spatial point pattern test. The S-index is calculated as the proportion of similar spatial units in the domain where s_i is used as a binary indicator of similarity. In the case of discrete pixel values, s_i are still used as a binary input whereas in the case of continuous pixel values the resulting SSIM values are used as a non-binary s_i input. The proposed spatial similarity test is tested with a simulation study where the simulations are designed to have comparisons that are either 80% or 90% identical. With the simulation study it is concluded that the test is not sensitive to the resolution of the pixel image. The application is done on property valuations in Johannesburg and Cape Town. The test is applied to the similarity of property prices in the same area over different years as well as testing the similarity of property prices between the different areas of properties.

Acknowledgements

Firstly, I want to thank every single family member and friend for the support during this time of writing and working on this mini-dissertation. I could not have done it without your love, support and motivational words.

Thank you to my supervisor, Dr Fabris-Rotelli, for your support in the writing of my mini-dissertation. Thank you for the virtual help during the pandemic and always making time to help and answer emails!

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

I also would like to thank Lightstone for providing me with the data for my application section.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In a forest, the way in which certain species of trees grow may reflect some information about the species' ability to grow in the specific location. The different patterns which results from the different tree species, may also reflect the specific species' ability to survive the competion to grow [22]. Furthermore, the comparison of criminal activity maps can assist researchers to identify the factors that increase the likelihood of those activities occuring [3]. The comparisons mentioned here are examples where tests for spatial similarity are needed which focus on whether the two spatial data sets originate from the same stochastic process in terms of their spatial structure [9]. In this mini-dissertation, we develop a generic similarity test for spatial data.

In traditional statistics, a random variable is defined as the desirable quantity to measure. An experiment is then conducted to obtain observations from this random variable. Each time an experiment is performed under identical conditions, the measured observations differ [6]. In spatial statistics, the random variable is a spatial process from which spatial data are observed [6]. As in traditional statistics, we are concerned about the random variable (spatial process) rather than the specific observations (spatial data) [6]. However, we still use the spatial data as a representation of the process as the process itself is not tangible.

Spatial data can take on three main forms, namely geostatistical data, lattice data and point patterns [13]. Geostatistical data are measured at fixed locations and is a partial realisation of the spatial process. Then an interpolation method, usually Kriging, is used to predict values where measurements are not taken [13]. In Figure 1.1 a continuous map, created by Kriging, is shown of Swiss rainfall which is an example of geostatistical data[1].

[1] Seen on: http://www.gitta.info/ContiSpatVar/en/html/Interpolatio_learningObject3.xhtml. Assessed on: 26 Januray 2021

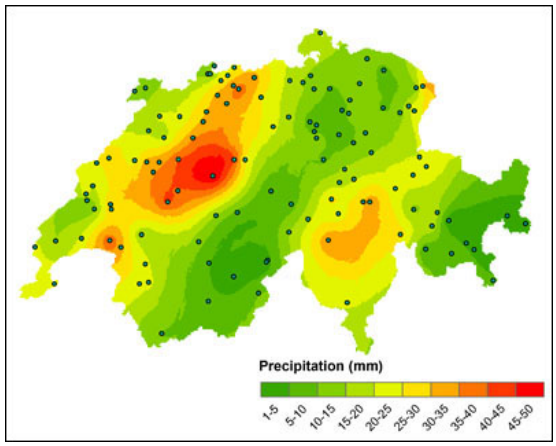


Figure 1.1: Example of geostatistical data. This is an example of the continuous map when the interpolation method, Kriging, is applied on the sample measurements taken.

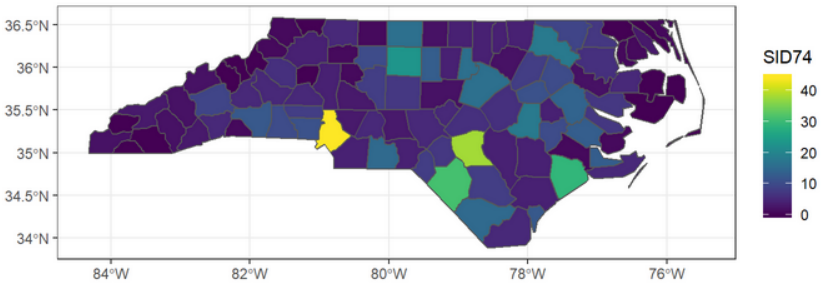


Figure 1.2: Example of lattice data.

Lattice data occur when the observational region is divided into predefined subregions (either a regular grid or an irregular grid) [40]. The spatial data can be observed at the individual subregions and can either be continuous or discrete. Figure 1.2 shows an example of a lattice data set on an irregular grid [33], where the sudden infant deaths of 1974 are shown in North Carolina. The state of North Carolina is divided into the counties. The number of sudden infant deaths for the county in that year are the value observed at each region.

An image is a lattice pattern with a regular grid. The subregions of an image are called pixels. Figure 1.3 is an example of a simple image with 100 pixels. The number of observed values at each pixel depends on the type of image. RGB images have three values for each pixel. The three values range from 0 to 255 for the red, green and blue indices respectively. The three values together form the colour displayed in the pixel. In the case of a greyscale image, each pixel has only one value that can range from 0 to 255.

Spatial point patterns consist of the locations of certain events [6]. In the case where only the locations of one event type is present, we call it an unmarked spatial point pattern (Figure 1.4(a)). Extra information

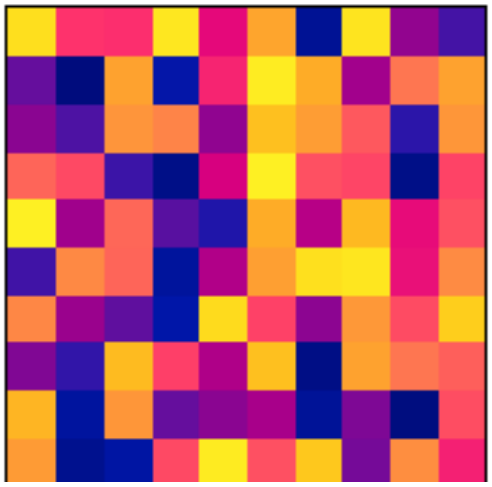
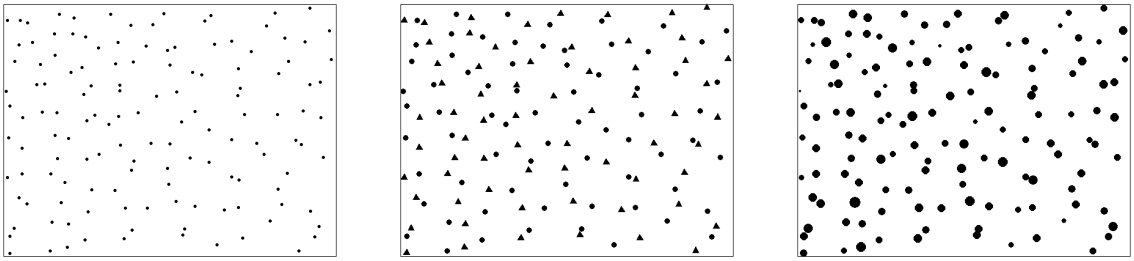


Figure 1.3: Example of a simple image data.

can be presented within the spatial point pattern by associating a value (mark) to each point. This is then called a marked spatial point pattern. This mark can be discrete (Figure 1.4(b)) or continuous (Figure 1.4(c)).



(a) (b) (c)

Figure 1.4: Examples of (a) unmarked, (b) multitype and (c) multivariate spatial point patterns.

Figure 1.4 shows three representations of a data set containing the locations of beta ganglion cells in a cat's retina [42, 46]. All three representations come from the same data set, therefore each of the cells and the locations are identical in all three patterns. The unmarked spatial point pattern in Figure 1.4(a) only shows the locations of the cells. In Figure 1.4(b), there are discrete marks attached to each observation indicating whether the cell is on or off. This is also called a multitype spatial point pattern [6]. The continuous marks in Figure 1.4(c) indicate the area of each cell. This is also called a multivariate spatial point pattern [6].

As far as the authors can determine, there are no spatial similarity tests that are able to handle more than one type of spatial data, that is, are generalisable. The currently available tests, only cover the more popular spatial data which is images, unmarked spatial point patterns and geostatistical data but

in different manners. We discuss these next.

In the literature, the tests of similarity between unmarked spatial point patterns can be divided into two main groups [2]. The first group is distance-based methods. These methods focus on the placement of the points relative to the other points. An example of a test is the comparison of the K-functions of the patterns [7, 20]. If both of the spatial patterns have similar structures, their K-functions are equal [20]. In [25], a more formal method for comparing K-functions is developed by using a studentized permutation test. The test statistic is based on the Behrens-Fisher-Welch t-statistic. The comparison of the K-functions was extended to be used with a Monte Carlo simulation in [16].

In 2012, a more formal test for the similarity of spatial patterns was developed [17]. This test involves the construction of a test statistic through the use of kernel smoothing. The construction of the kernel estimator involves the distances between the points. This test may not produce a nominal significance level and in [20] it was extended by making use of a bootstrap calibration to be able to compare smaller patterns as well. The biggest disadvantage of this method is its computational complexity, especially when comparing bigger patterns. This method is used in [9] on the covariate measurements at the locations within the point pattern.

The second type of similarity test is area-based methods. These methods work by aggregating the points into smaller regions, such as neighbourhood boundaries (irregular) or regular grids [2]. The tests are then based on the number of events in each region [7, 13]. Andresen's spatial point pattern test is the most commonly used test amongst geographers and criminologists [2]. This is a simple test that uses bootstrap sampling to create a non-parameteric 95% confidence interval on one of the patterns for each region. This is compared to the number of points in the corresponding region from the other pattern and then a global similarity index, the S-index, is calculated. This S-index can also be seen as the proportion of regions where the number of points in the one pattern is contained in the confidence interval.

In a follow-up paper, it was mentioned that the two patterns can be identified as similar if $S > 0.8$ [3]. Kirsten et al developed more sensible similarity bounds based on the appearance of the point patterns by making use of a simulation study [28, 29]. In 2018, this test was improved by the proposal of two alternative methods [47]. Our suggestion is that the calculation of the S-index can be improved by changing the calculation so that the S-index is the mean value of the local similarity throughout the observational region instead of the proportion of similar spatial units.

Another area-based test for unmarked spatial point patterns was proposed by [1] in 2016. They suggested testing the similarity of spatial patterns by making use of space-filling curves to order the space. This can be used in any n dimensional space. In recent years, the development of similarity tests for unmarked spatial point patterns using kernel density estimates have become more and more popular [14, 24, 35]. In [24], three two-sample density tests were developed. All three of these tests are based on the Maximum

Mean Discrepancy (MMD) that is also known as tests where the test statistic is the difference between the means of two distributions. The first two tests make no assumption regarding the two distributions, whereas the third test is based on the asymptotic distribution of the previous test statistic.

For the testing of spatial similarity between images, the Structural SIMilarity index (SSIM) is used [10]. This method works with a sliding window approach that measures the mean, variance and covariance of the pixel values from the two images for the sliding window to calculate a luminance, constrast and structure component that is multiplied together to equal the SSIM value for the pixel in the centre of the sliding window. This results in an SSIM value for each pixel. A global SSIM value can be obtained by taking the mean of the individual SSIM values for the pixels.

Before the development of the SSIM, [23] tested the similarity of images by computing an error matrix through direct comparison. The overall similarity statistic is based on the $\hat{K}$ statistic developed in [12]. In [31], image retrieval is done by using spatial similarity testing between the different images. The algorithm used for the image retrieval takes the number of objects that are common between the two images.

The testing of similarity between geostatistical data is not that well used. A measure that can be useful for such a test was used in 2016 by [18]. In this paper, hierarchical clustering was done on geostatistical data. To perform hierarchical clustering, a dissimilarity matrix is constructed before the actual clustering is done. This dissimilarity measure can possibly be used as a measure of similarity between geostatistical data sets. This measure uses non-parametric kernel estimation on the spatial dependence. A more formal method for the testing of similarity between geostatistical data sets was developed in 2010 [38]. This method uses geostatistical entropy to measure the similarity. This method works by calculating the Kriging distortions of the prediction vectors. The average of the two calculated distortions is the similarity measure.

In this mini-dissertation we propose a test for spatial similarity that is generalised to be able to handle any type of spatial data, namely geostatistical data, lattice data, point patterns, marked point patterns as well as images. The test consists of three steps where the first step involves creating a pixel image representation of both the spatial data sets considered. The pixel image representation is obtained differently for each spatial data type, which is discussed in detail in the remaining chapters. In the second step, the SSIM is used to create a local similarity map when the pixel values are continuous. An SSIM value is calculated for each pixel. In the case of discrete pixel values, the local similarity map is created by direct comparison of the pixel values. The calculation of the final similarity measure is done in the third step of the test. This calculation is based on the S-index of Andresen's spatial point pattern test [2]. The S-index is originally calculated as the proportion of similar spatial units in the domain, $S = \frac{\sum_{i=1}^{n} s_i}{n}$, where s_i is a binary value. In the case of discrete pixel values, s_i is still binary whereas in the case of continuous pixel values, the resulting SSIM values are used as a non-binary s_i input.

In this mini-dissertation, we aim to:

- Propose a generalisable spatial similarity test
- Develop a method to represent each spatial data type as a pixel image representation
- Compare the pixel image representations to form a local similarity map
- Develop a global similarity index based on Andresen's S-index with a non-binary input
- Apply the test on property prices in the same area over different years
- Apply the test on property prices of the same year over different areas

In Chapter 2 we describe different types of spatial data and different methods used for the new test. In Chapter 3 the new similarity test is discussed in depth with a simulation study in Chapter 4. Chapter 5 includes an application of the method on property prices in Johannesburg and Cape Town for the years 2017, 2018 and 2019. A summary and discussion of the work is outlined in Chapter 6.

Chapter 2

Theoretical background

In this chapter, the notation for the different types of spatial data is discussed followed by theory about density estimation, SSIM index, Andresen's spatial point pattern test, Kriging and the k-NN estimation. The types of spatial data that we consider are geostatistical data, lattice data and point patterns [13].

We consider the spatial process [13]

$$\begin{aligned} \boldsymbol{Z(s)} &= \{\boldsymbol{Z(s_1)}, \boldsymbol{Z(s_2)}, \ldots, \boldsymbol{Z(s_n)}\} \\ &= \bigcup_{i=1}^{n} \{\boldsymbol{Z(s_i)}, \boldsymbol{s_i} \in D \subset \mathbb{R}^p\}, \end{aligned} \tag{2.1}$$

where $\boldsymbol{s_i} \in \mathbb{R}^p$ is a spatial location in the p-dimensional space, and $\boldsymbol{D}$ is a subset of the p-dimensional space also known as the spatial domain where the spatial random variable is defined [41]. Each spatial location, $\boldsymbol{s_i}$, is defined by the specific coordinates $\boldsymbol{s_i} = (s_{i1}, s_{i2}, \ldots, s_{ip})$ which can either be discrete or continuous.

The spatial domain is the area where spatial data are observed. We also call this a window. The boundary of the spatial domain consists of multiple connected curves or straight lines that do not cross each other [6]. These windows are either rectangular in shape or any other polygonal shape. A spatial domain frequently used is the boundaries of a country, province, city etc.

2.1 Geostatistical data

As mentioned previously, the spatial location $\boldsymbol{s_i}$ may be either discrete or continuous. In the case where it is continuous over the spatial domain, it is called geostatistical data [13]. For this, the spatial process

in Equation (2.1) is defined $\forall\ \boldsymbol{s_i} \in D$ [41].

Geostatistical data are measured at sampled locations, therefore it is a partial realisation of $\boldsymbol{Z(s)}$ [13, 41]. Through an interpolation method such as Kriging, predictions are made regarding the unobserved values of the spatial process [41]. This allows each possible spatial location in the spatial domain to vary continuously [13].

2.2 Lattice data

For lattice data, as in Figure 1.2, the spatial process in Equation (2.1) is defined for a fixed set of spatial locations [13]. We then say that $\boldsymbol{s_i}$ is discrete in the case of lattice data. The realisation of $\boldsymbol{Z(s_i)}$ is the vector of the values observed at spatial location $\boldsymbol{s_i}$.

Each spatial location is a region which makes up part of the spatial domain. The collection of all the spatial locations is equal to the spatial domain [41]. Therefore, the spatial locations can be seen as a partition as defined in Definition 1. We denote the regions of the lattice data as s_i and call it spatial locations for the sake of being consistent. However, other sources may denote the separate regions as A_i to make it clear that the data is observed over regions.

Definition 1. *P is a **partition** [26] of a set X if and only if:*
(1) P does not contain an empty subset, $\emptyset \notin P$
(2) The elements of P are disjoint, if $A \in P$ and $B \in P$, then $A \cap B = \emptyset$
(3) The union of all subsets of P is the set X,

$$\bigcup_{i=1,\ldots,n \forall A_i \in P} A_i = X$$

Each spatial region can be represented by a representative point. The chosen point can either be the centroid of the region or any other appropriate point, for example the capital city of a country or province. As with the window, these regions can also be either regular or irregularly shaped. When the spatial domain over which the spatial process is observed is for example South Africa, the spatial regions can for instance be each of the provinces. In this case, the regions are irregularly shaped and we are then dealing with an irregular lattice. Regular shaped regions are obtained when D is divided into a number of grid cells which forms a regular lattice.

A special case of a regular lattice pattern is an image [41]. With an image, the spatial domain is divided into equally shaped regular regions. Each region is better known as a pixel. The realisation of the spatial process for each pixel is either a single value (in the case of greyscale images) or a vector of multiple values.

When each pixel value has between three and ten values, we are dealing with multispectral images [43]. The well known RGB images has three values (red, green and blue) and is a special case of a multispectral image. In the case of more than 10 values, we are dealing with hyperspectral images [43]. Hyperspectral images can have many values per pixel.

2.3 Point patterns

As mentioned before, a point pattern consists of the locations of a certain event. This type of spatial data need a special case of the spatial process from Equation (2.1) because with geostatistical and lattice data, the spatial locations are fixed and the realisations of the spatial process are the measurements observed at the specific spatial location. With the point pattern, the type of event is known beforehand and the desired variable is the location at which this event occured.

The spatial point process is a random variable specifically for point patterns and is the stochastic mechanism that generated each point in the point pattern. A point process is a random variable whose outcome is a point pattern [6].

Definition 2. *A **finite point process**, X [6], in a p-dimensional space is defined to be any stochastic mechanism for which:*
(1) every possible outcome is a finite point pattern, and
(2) for every spatial domain $D \subset \mathbb{R}^p$, the number of points falling in D is a well-defined random variable.

A spatial point pattern is a realisation of a point process and can be defined as

$$\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}.$$

Each $\boldsymbol{x_i}$ denotes a point in the spatial point pattern that consists of n data points. A point pattern can either be unmarked or marked. In the case of an unmarked spatial point pattern, the only variable for the event represents the specific location in $\mathbb{R}^p$

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}).$$

In the case of a marked spatial point pattern, apart from the variables representing the location of the event, there is also one or more variables representing a value or a mark. The mark can be defined as extra information about that specific event

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}, m_i = m(x_{i1}, x_{i2}, \ldots, x_{ip})),$$

where $m(x_{i1}, x_{i2}, \ldots, x_{ip})$ is a function that includes the mark(s) [5]. This function can either be discrete, continuous or both.

2.4 Estimation of the intensity of a spatial point process

The intensity of a spatial point process is based on the average number of events per area [6]. This is the most basic and important form of descriptive analysis for spatial data and is equivalent to the calculation of the mean of standard data. The calculation of the intensity is used to study the arrangement of the individual events.

The intensity function of the spatial point process is denoted as

$$\lambda(\boldsymbol{s}) = \lim_{|ds|\to 0} \frac{E[N(ds)]}{|ds|}, \tag{2.2}$$

where ds is an infinitesimal area containing the point $\boldsymbol{s}$, $|ds|$ is the area of ds and $N(ds)$ is the number of points in ds [19, 21]. In the case where $\lambda(\boldsymbol{s})$ is constant, the spatial point process is homogeneous and does not vary across the spatial domain. If this is not the case, the spatial point process is inhomogeneous.

As mentioned before, the spatial point process is the random variable. Therefore, the intensity of the spatial point process is estimated by using the spatial point pattern. This is done by using kernel density estimation [6] which is a nonparametric estimator of the intensity and can be calculated as

$$\tilde{\lambda}(u) = \sum_{i=1}^{n} \kappa(u - x_i), \tag{2.3}$$

where u is a spatial location in the spatial domain and $\kappa(\cdot)$ is the kernel function. According to [15], the choice of the kernel function is not that important.

It is always better to have an unbiased estimator, therefore we want to determine whether the estimator is unbiased or not. For this calculation, the following result will come in handy and is called Campbell's formula [5]

$$E\left(\sum_{i=1}^{n} f(x_i)\right) = \int_{\mathbb{R}^2} f(u)\lambda(u)du. \tag{2.4}$$

Following this, we define a function at a fixed spatial location v [6]

$$f(u) = \begin{cases} \kappa(v-u) & \text{if} \quad u \in D \\ 0 & \text{if} \quad u \notin D \end{cases}. \tag{2.5}$$

So, then from applying Equation (2.5)

$$\tilde{\lambda}(v) = \sum_{i=1}^{n} \kappa(v - x_i) = \sum_{i=1}^{n} f(x_i). \tag{2.6}$$

By applying Equation (2.4) to the expected value of the function in Equation (2.3), we obtain

$$\begin{aligned} E\left(\tilde{\lambda}(v)\right) &= E\left(\sum_{i=1}^{n} f(x_i)\right) \\ &= \int f(u)\lambda(u)du \\ &= \int_D \kappa(v-u)\lambda(u)du. \end{aligned} \tag{2.7}$$

From Equation (2.7), we see that $E\left(\tilde{\lambda}(v)\right) \neq \lambda(u)$ and therefore Equation (2.3) is not an unbiased estimator of the true intensity, $\lambda(u)$. If we assume homogeneity, $\lambda(u) = \lambda$, we still have a biased estimator

$$E\left(\tilde{\lambda}(v)\right) = \lambda \int_D \kappa(v-u)du. \tag{2.8}$$

From this, we can define an unbiased estimator for Equation (2.3) under the assumption of homogeneity as [6]

$$\tilde{\lambda}^U(u) = \frac{1}{\int_D \kappa(v-u)du} \sum_{i=1}^{n} \kappa(u - x_i). \tag{2.9}$$

Seeing that the assumption of homogeneity in a spatial point process cannot be taken lightly, we aim for the best possible estimate of the intensity of a spatial point process. Although homogeneity does exist in spatial data, the data will be more often than not be non-homogeneous and thus should be tested for as it affects your modelling approach. This estimator is called Diggle's kernel estimator [6] which outperforms the other estimators in terms of the mean squared error. This means that among the other estimators, Diggle's kernel estimator has the lowest mean squared error. It is given as

$$\tilde{\lambda}^D(u) = \sum_{i=1}^{n} \frac{1}{\int_D \kappa(x_i - v)dv} \kappa(u - x_i). \tag{2.10}$$

The estimator in Equation (2.10) takes care of the edge effects present. Edge effects are present in spatial data as we only observe the data in a certain window. For example, when the spatial point pattern shows the locations of crimes in a neighbourhood, the true crime locations are not limited to the window where our spatial point pattern is observed. There are also criminal activities occurring outside the window that may have an influence on the events in the spatial point pattern which are not observed [6]. Therefore we define Diggle's edge correction factor as

$$e(x_i) = \frac{1}{\int_D \kappa(x_i - v)dv}. \tag{2.11}$$

In the case of a marked spatial point pattern, an estimate for the intensity function can also be calculated. The estimator for the intensity function of a marked spatial point pattern is called the Nadaraya-Watson smoother [6, 36]

$$\tilde{m}(u) = \frac{\sum_{i=1}^{n} m_i \kappa(u - x_i)}{\sum_{i=1}^{n} \kappa(u - x_i)}, \tag{2.12}$$

where m_i is the real-valued mark associated with point x_i. Equation (2.12) is calculated for each spatial location u. Diggle's edge correction can also be applied to this estimator [6] as

$$\tilde{m}^D(u) = \frac{\sum_{i=1}^{n} m_i \kappa(u - x_i)/e(x_i)}{\sum_{i=1}^{n} \kappa(u - x_i)/e(x_i)}, \tag{2.13}$$

where $e(x_i)$ is defined as in Equation (2.11).

2.5 Structural similarity index

The structural similarity index (SSIM) was first developed as a quality index for images and later on used as a similarity index between images [44, 45]. The algorithm works by using a sliding window to move pixel-by-pixel across the two images. In each sliding window, the SSIM is calculated.

The calculation for the SSIM consists of three terms: contrast, structure and luminance [45]. Let $\boldsymbol{x_1}$ be the non-negative real number pixel values from the sliding window in the first image and let $\boldsymbol{x_2}$ be the non-negative real number pixel values from the sliding window in the second image, then we can calculate the separate components as

$$\text{Luminance:} \quad \ell(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\mu_{x_1}\mu_{x_2} + C_1}{\mu_{x_1}^2 \mu_{x_2}^2 + C_1} \tag{2.14}$$

$$\text{Contrast:} \quad c(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\sigma_{x_1}\sigma_{x_2} + C_2}{\sigma_{x_1}^2 \sigma_{x_2}^2 + C_2} \tag{2.15}$$

$$\text{Structure:} \quad s(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\sigma_{x_1,x_2} + C_3}{\sigma_{x_1}\sigma_{x_2} + C_3} \tag{2.16}$$

where

$$\begin{aligned} \mu_{x_1} &= \tfrac{1}{N}\sum_{i=1}^{N} x_{1i} & \mu_{x_2} &= \tfrac{1}{N}\sum_{i=1}^{N} x_{2i} \\ \sigma^2_{x_1} &= \tfrac{1}{N-1}\sum_{i=1}^{N} (x_{1i}-\mu_{x_1})^2 & \sigma^2_{x_2} &= \tfrac{1}{N-1}\sum_{i=1}^{N} (x_{2i}-\mu_{x_2})^2 \end{aligned} \tag{2.17}$$

and

$$\sigma_{x_1,x_2} = \frac{1}{N-1}\sum_{i=1}^{N} (x_{1i}-\mu_{x_1})(x_{2i}-\mu_{x_2})$$

and where $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$ and $C_3 = \frac{C_2}{2}$ are constants to avoid unstable results in the case where the sum of the squared means as well as the sum of the variances in the sliding window approaches zero [45], N is the number of pixels within the sliding window, $K_1, K_2 \ll 1$ are small constants and L is the range of the values for the pixels in the image. It is suggested in [45] that the constant values can be used as, $K_1 = 0.01$ and $K_2 = 0.03$. The SSIM can then be calculated as

$$SSIM(\boldsymbol{x_1}, \boldsymbol{x_2}) = [\ell(\boldsymbol{x_1}, \boldsymbol{x_2})]^{\alpha} [c(\boldsymbol{x_1}, \boldsymbol{x_2})]^{\beta} [s(\boldsymbol{x_1}, \boldsymbol{x_2})]^{\gamma}, \tag{2.18}$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. Usually in literature $\alpha = \beta = \gamma = 1$ which assigns an equal importance to each term [45]. The SSIM is bounded between -1 and 1.

As the SSIM is calculated for each pixel, we map the values for each pixel. This creates an image visualising where the two images being compared are more similar and where not. An overall index for the two images is calculated as the mean value of the SSIM values for each pixel.

2.6 Andresen's spatial point pattern test and the S-index

In 2009, Andresen developed a nonparametric test to calculate the similarity between two spatial point patterns [2]. This test uses an area-based approach for the similarity testing and is not concerned with the statistical distribution of the points in the spatial point pattern but rather if the points in two different spatial point patterns are similarity located or not. The outcome of this test is a proportion, called the S-index, that indicates the degree of similarity between the two spatial point patterns.

Considering two spatial point patterns, $X_1 = \{p_{11}, p_{21}, \ldots, p_{n_1 1}\}$ and $X_2 = \{p_{12}, p_{22}, \ldots, p_{n_2 2}\}$, observed over the same spatial domain. The pattern X_1 is called the base pattern and X_2 the test pattern. The spatial domain is divided into m predefined number of regions $A_i, i = 1, 2, \ldots, m$. These regions can be either regular with a grid or irregular by using the neighbourhood boundaries, for example. Then, the percentage of points in X_1 is calculated for each A_i

$$t_i = \frac{\sum_{k=1}^{n_1} I(p_{k1} \in A_i)}{n_1}. \tag{2.19}$$

For $r = 1, 2, \ldots, 200$, a simple random sample is taken from X_2, $B_r = \{q_{1r}, q_{2r}, \ldots, q_{n_r r}\}$ where $n_r = 0.85 \times n_2$. The pattern B_r is divided into the same regions as above, namely A_i. Then the percentage of points from pattern B_r in each A_i is calculated

$$b_{ir} = \frac{\sum_{k=1}^{n_r} I(q_{kr} \in A_i)}{n_r}. \tag{2.20}$$

This step is repeated 200 times for the sake of being conservative [2]. A sample of 85% is taken following the research done by [39] that if a random sample is taken on a spatial point pattern, the spatial pattern is lost if less than 85% of the points in the spatial point pattern are used.

Let $\boldsymbol{b_i} = (b_{i1}, b_{i2}, \ldots, b_{i200})'$ be the vector of all the percentages of points in region A_i for $r = 1, 2, \ldots, 200$ and c_i be the confidence interval for each region A_i. Then the 2.5^{th} percentile of $\boldsymbol{b_i}$ is the lower limit of c_i and the 97.5^{th} percentile is the upper limit of c_i. The S-index is then calculated as

$$S = \frac{\sum_{i=1}^{m} I(t_i \in c_i)}{m}. \tag{2.21}$$

A threshold of 0.8 for the S-index has been used to indicate that the two spatial point patterns being compared are similar [3, 4].

2.7 Kriging

Kriging [30] is a popular interpolation method used to estimate the unobserved spatial locations within the spatial domain. The general formula to use when we are interested in the value at spatial location $\boldsymbol{u}$ is

$$\hat{Z}(\boldsymbol{u}) = \sum_i w_i Z(\boldsymbol{s_i}) + \epsilon, \tag{2.22}$$

where w_i is the weight depending on how far $\boldsymbol{s_i}$ is from $\boldsymbol{u}$[1]. The restriction on the weights are that they should sum to 1 [13].

[1]Seen on: https://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm. Assessed on: 17 August 2020

The estimation is done based on the semi-variogram that estimates the dependency within the spatial data. The semi-variogram can be obtained from the following formula[2]

$$\gamma(\boldsymbol{s_i}, \boldsymbol{s_j}) = \frac{1}{2} \frac{\sum_i \sum_j (Z(\boldsymbol{s_i}) - Z(\boldsymbol{s_j}))^2}{N}. \tag{2.23}$$

The weights from Equation (2.22) can be obtained by using

$$\hat{\boldsymbol{w}} = A^{-1}\boldsymbol{b}, \tag{2.24}$$

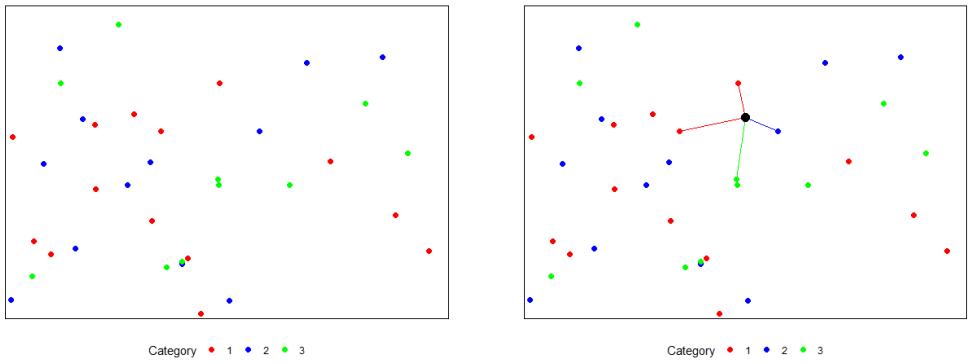
where A is a matrix containing $\gamma(\boldsymbol{s_i}, \boldsymbol{s_j})$ and $\boldsymbol{b}$ is a vector containing $\gamma(\boldsymbol{s_{new}}, \boldsymbol{s_i})$ where $\boldsymbol{s_{new}}$ is the spatial location where the prediction is made.

2.8 k nearest neighbour classification

With k nearest neighbour classification, we consider the distance (for simplicity, Euclidean distance) between each spatial data point, $\boldsymbol{x_i}$, and the spatial location where the prediction should be made, $\boldsymbol{u}$ [27]. In the case of the use of Euclidean distance, we use,

$$d_i = \|\boldsymbol{x_i} - \boldsymbol{u}\|. \tag{2.25}$$

The prediction at $\boldsymbol{u}$ is the modal value of the k nearest spatial data points considered.



(a) (b)

Figure 2.1: Example of the principle of k nearest neighbours classification. (a) The discrete marked point pattern used in the example. This point pattern has discrete marks indicting whether each of the points fall within one of three categories. (b) The principle when estimating the value of the black dot. When $k = 4$, the four closest points to the black dot is considered. Two of these four points fall within category one and one in category two and three. Therefore, the black dot is estimated to be in the first category.

[2]Seen on: https://www.youtube.com/watch?v=J-IB4_QL7Oc. Assessed on: 25 August 2020

Figure 2.1 is an example illustrating the principle of k nearest neighbours classification. Figure 2.1(a) is the marked point pattern with discrete marks used in the explanation. This pattern has marks indicating whether the points fall in one of the three categories. In Figure 2.1(b), we want to estimate in which category the black dot should be. The black dot is the spatial location at which the prediction should be made, $\boldsymbol{u}$. We consider $k = 4$, which means that we are looking at the four closest points to the black dot. The black dot is estimated to fall in the modal category of the four closest points. In this case, it would be the first category, as there are two red points within the four closest points.

2.9 Conclusion

In this chapter, the notation of the spatial data types considered are discussed. This is followed by the theory of the methods considered in the proposed spatial similarity test. The estimation of the intensity of a spatial point pattern, Kriging as well as k nearest neighbour classification are used to create the pixel image representations in the proposed spatial similarity test. The Structural SIMilarity index (SSIM) is used to compare the pixel image representations to form a local similarity map. The calculation of the global similarity index is based on the S-index from Andresen's spatial point pattern test.

In the next chapter the proposed spatial similarity test and these principles will be discussed in detail.

Chapter 3

Proposed spatial similarity test

We propose a generalised method for the testing of similarity between two spatial data sets. The aim of this test is to be able to handle any type of spatial data, namely point patterns, geostatistical and lattice data and calculate a percentage of similarity between the two data sets. The new test consists of three steps which is outlined in detail throughout this chapter.

The two data sets to be compared are called X_1 and X_2. The goal of the first step is to represent each of the spatial data types in the same way. This is what makes the test generic. We create a pixel image representation of X_1 and X_2 and denote this as Y_1 and Y_2. In the second step, we create a local similarity map indicating a local similarity value for each pixel from Y_1 and Y_2. The final step involves the calculation of a similarity percentage from the pixel values in the local similarity map.

The new spatial similarity test is outlined step-by-step in Figure 3.1. The test starts by creating a pixel image representation of both the data sets considered. The pixel image representation is obtained differently for each spatial data type which is indicated by Figure 3.1. When dealing with point patterns, a kernel density estimation is used to obtain the pixel image representation. For unmarked point patterns, Diggle's edge corrected estimator is used as in Equation (2.10) explained in Section 2.4. For marked point patterns with continuous marks, the Nadaraya-Watson smoother is used as in Equation (2.12) from Section 2.4. In the case of discrete marks in a marked point pattern, kNN estimation is used to create the pixel image representation as explained in Section 2.8. The pixel image representation for irregular lattice data is slightly less involved as the pixel takes on the value of the spatial location (region) the centroid falls in. This is the same for both continuous and discrete values. For regular lattice data, such as images, are already in a pixel format and the first step is skipped in this case. For the geostatistical data, the pixel image is obtained by using the well known interpolation method of Kriging from Section 2.7. The second step of creating a local similarity map is done with direct comparison for pixel images containing

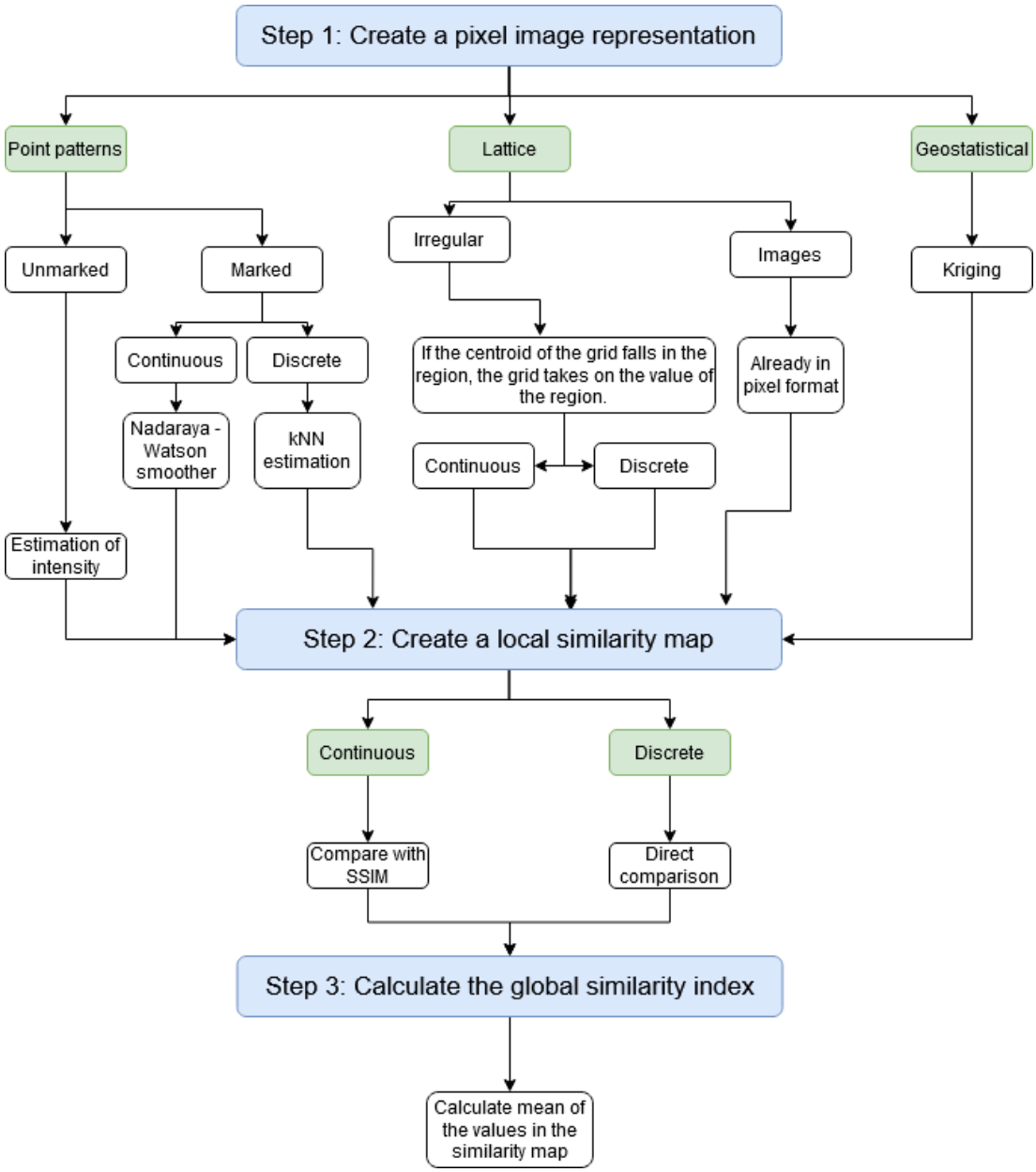


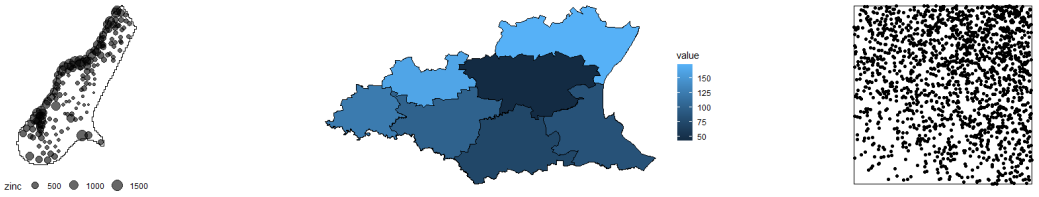
Figure 3.1: Diagram explaining the structure of the proposed spatial similarity test.

discrete values and the SSIM index applied to each pixel for the continuous valued pixel images which is explained in Section 2.5. The third step which involves the calculation of the final similarity measure is done using the values in the local similarity map. This calculation is based on the final calculation of the S-index from Andresen's spatial point pattern test as outlined in Section 2.6.

3.1 Step 1: Create a pixel image representation

In this step of the proposed spatial similarity test, we represent the two spatial data sets, X_1 and X_2 as pixel images, Y_1 and Y_2. The resolution (that is, the number of pixels) is decided by the user before-hand. For the purpose of this chapter, we consider only spatial data in two dimensions. Hence we work in $\mathbb{R}^2$.

Consider the spatial data sets in Figure 3.2. Figure 3.2(a) shows a geostatistical data set, Figure 3.2(b) a lattice data set with irregular regions and Figure 3.2(c) an unmarked spatial point pattern. These data sets are used throughout this section to explain the method of creating a pixel image representation. Take note that although the window for the point pattern in Figure 3.2 is rectangular; this serves only an example and the window may be any polygonal shape.

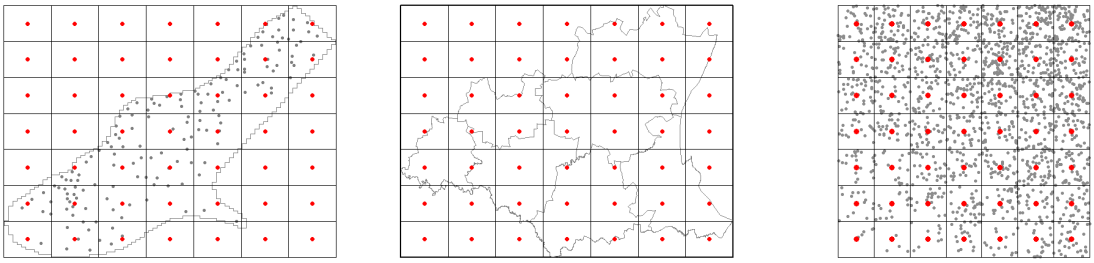


(a) (b) (c)

Figure 3.2: Three different spatial data sets that are used as the example throughout this section. The three data sets are each of a different data type namely (a) geostatistical, (b) lattice with irregular regions and (c) an unmarked spatial point pattern.

To obtain the pixel image representation, the first step divides the spatial domain, D, into an $m \times m$ grid. Each grid cell represents a pixel. We then need to define spatial locations at the centroids of each of the $M = m^2$ pixels as $\boldsymbol{u} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_M})$. For illustration purposes, let us consider $m = 7$. Figure 3.3 shows the way in which the spatial domain, D, is divided into pixels for each data type. The most intuitive way to obtain the pixels and the locations of the centres is to enclose the spatial domain with the smallest rectangular window. The enclosed rectangular window is then divided into pixels. If the centre of the pixel falls outside of the domain, the pixel has an empty value (or an NA value) for the pixel image representation.

At each spatial location, $\boldsymbol{u_j}$ for $j = 1, 2, \ldots, M$, a value for the corresponding pixel needs to be determined.



(a) (b) (c)

Figure 3.3: Illustration of how the spatial domain for each data type is divided into pixels when $m = 7$. The red dots represent the $\boldsymbol{u_j}$'s. (a) is a geostatistical data set, (b) is a lattice data set and (c) is a point pattern data set. With the point patterns, the grey dots represent the $\boldsymbol{x_i}$'s and with the geostatistical data, the grey dots represent the $\boldsymbol{s_i}$'s. Recall that the centres of the grid cells are denoted as $\boldsymbol{u_j}$ and the spatial data points in a point pattern are denoted with $\boldsymbol{x_i}$. Also, the measurements of a geostatistical data set are taken at the spatial locations denoted by $\boldsymbol{s_i}$.

For each type of spatial data, this is done in a different manner which is explained in the following subsections.

3.1.1 Spatial point patterns

With a spatial point pattern, the pixel image representation is obtained by estimating the intensity of the spatial point process, using a kernel density estimation. This is done by calculating Diggle's corrected density estimate at the representative point ($\boldsymbol{u_j}$'s) of each pixel [6].

Diggle's corrected density estimate is used for the calculation as it outperforms the other estimators in terms of the mean squared error [6]

$$\tilde{\lambda}^D(\boldsymbol{u_j}) = \sum_{i=1}^{n} \frac{1}{e(\boldsymbol{x_i})} \kappa(\boldsymbol{u_j} - \boldsymbol{x_i}), \tag{3.1}$$

where the kernel, $\kappa(\cdot)$, we use a bivariate Gaussian density $f(\boldsymbol{d}) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\boldsymbol{d}\Sigma^{-1}\boldsymbol{d}'\right\}$ with $\Sigma =$ bandwidth $\times I_2$. The bandwidth is the standard deviation of the kernel density and can be seen as the smoothing parameter of the kernel. The larger the bandwith, the smoother the estimation. There are different methods of calculating the bandwidth. A popular bandwidth method is Diggle's bandwidth [6]. Although the calculation of Diggle's bandwidth assumes a Cox process, this is the bandwidth used for the purpose of this mini-dissertation as choosing the optimal bandwidth is beyond the scope of the work.

Another advantage of Diggle's corrected density estimate, is the edge correction factor [6]. The edge

correction factor in Equation (3.1) is

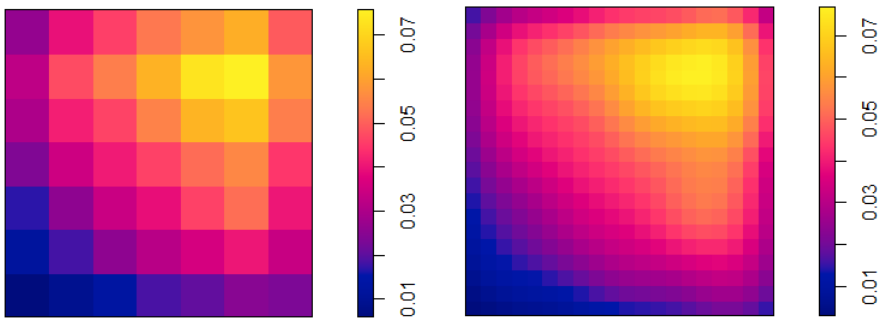
$$e(\boldsymbol{x_i}) = \int_D \kappa(\boldsymbol{x_i} - \boldsymbol{v_k}) d\boldsymbol{v_k}, \tag{3.2}$$

which is estimated using numerical integration. This is done by dividing the spatial domain, D, into a finer $g \times g$ grid. It is important to note that this is a separate calculation as the calculation of the kernel density estimate. The approaches are similar but should be treated separately.

Again, the centroids of the $Q = g^2$ grid cells is used as the spatial locations, $\boldsymbol{v} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_Q})$. Then, the calculation of Equation (3.2) through numerical integration involves that for each observation in the spatial point pattern, $\boldsymbol{x_i}, i = 1, \ldots, n$, we calculate the differences, $\boldsymbol{d_e} = (\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots, \boldsymbol{d_Q})$, between the coordinates of the point $\boldsymbol{x_i}$ and the spatial locations $\boldsymbol{v_k}$, $k = 1, 2, \ldots, Q$. The edge correction factor is then calculated as

$$e(\boldsymbol{x_i}) = \frac{\text{area}(D)}{Q} \sum_{k=1}^{Q} f(\boldsymbol{d_k}), \tag{3.3}$$

where $f(\boldsymbol{d_k})$ is again the bivariate Gaussian density.



(a) (b)

Figure 3.4: Resulting pixel image representations of the unmarked point pattern in Figure 3.2(c) with two different resolutions. (a) $m = 7$ and (b) $m = 20$.

Figure 3.4 is the resulting pixel image representation when $m = 7$ in Figure 3.4(a) and $m = 20$ in Figure 3.4(b) from applying Equation (3.1). It can be seen that the pixel image representations have similar appearances with Figure 3.4(b) with much more detail. From this we can conclude that as m increases, more smaller detail in the patterns is visible.

However, not all spatial point patterns are unmarked. In the case of a marked spatial point pattern we need a slightly different approach as the spatial data points have a mark that needs to be taken into account. As mentioned is Section 2.3, these marks can either be continuous or discrete.

When the marked spatial point pattern has continuous marks, we estimate the intensity of the marked spatial point process using the Nadaraya-Watson smoother with Diggle's edge correction factor [6]

$$\tilde{m}^D(\boldsymbol{u_j}) = \frac{\sum_{i=1}^{n} m_i \kappa(\boldsymbol{u_j} - \boldsymbol{x_i})/e(\boldsymbol{x_i})}{\sum_{i=1}^{n} \kappa(\boldsymbol{u_j} - \boldsymbol{x_i})/e(\boldsymbol{x_i})}, \tag{3.4}$$

where the kernel, $\kappa(\cdot)$ is again the bivariate Gaussian density, m_i denotes the real-valued mark of point $\boldsymbol{x_i}$ and $e(\boldsymbol{x_i})$ is the edge-effect factor defined Equation (2.11).

With a marked spatial point pattern that has discrete marks, the Nadaraya-Watson smoother in Equation (3.4) is not valid as the marks are now categorical instead of real-valued. The approach to obtain a pixel image representation for a marked spatial point pattern with discrete marks involves a k nearest neighbour classification as discussed in Section 2.8.

A prediction is made at each grid centre, $\boldsymbol{u_j}$. Therefore, the distance between each spatial data point and the grid centres are calculated

$$d_i(\boldsymbol{u_j}) = \|\boldsymbol{x_i} - \boldsymbol{u_j}\|. \tag{3.5}$$

The predicted value is the modal value of the k closest points to the spatial location $\boldsymbol{u_j}$. The choice of k is completely up to the user. Care should be taken that the value for k should be strictly less than the number of spatial data points. For the purpose of this mini-dissertation, the value of k is chosen as 10% of the number of points in the pattern.

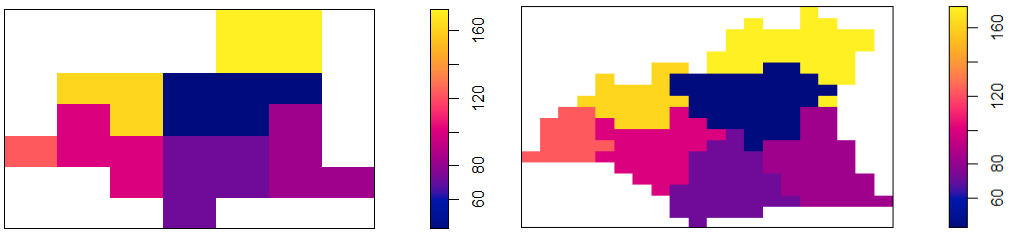
3.1.2 Lattice data

Compared to spatial point patterns, there is no intensity to be estimated with lattice data. To obtain a pixel image representation of lattice data, we again divide the spatial domain into a grid. Each grid cell takes on the value of the region in which its centroid falls.

Before dividing the spatial domain into a grid, we find the smallest rectangular window that encloses the entire spatial domain. Then this window is divided into a grid and representative points for each pixel is obtained. This is illustrated in Figure 3.3.

The spatial location (area) of the lattice pattern in which $\boldsymbol{u_j}$ is contained should be determined. If $\boldsymbol{u_j} \in \boldsymbol{s_i}$ then $\boldsymbol{u_j} = Z(\boldsymbol{s_i})$. Also, if $\boldsymbol{u_j} \notin D$, then the grid block can be omitted. Whether the $Z(\boldsymbol{s_i})$ values are discrete or continuous, the method stays the same.

Figure 3.5 is the resulting pixel image representations with $m = 7$ in Figure 3.5(a) and $m = 20$ in Figure



(a) (b)

Figure 3.5: Resulting pixel image representation of the lattice data in Figure 3.2(b) with two different resolutions. (a) $m = 7$ and (b) $m = 20$.

3.5(b). The same detail can be seen in both the pixel image representations, however in Figure 3.5(b), the different spatial locations are more visible than in Figure 3.5(a).

3.1.3 Geostatistical data

With geostatistical data, it is observed at sampled locations, however the location of measurement is considered fixed and the value observed a random variable [13]. With geostatistical data in general, we are interested in estimating a continuous map throughout the entire spatial domain, which is obtained with an interpolation method called Kriging, for example by predicting the unobserved values [13]. For the pixel image representation of a geostatistical data set, we divide the spatial domain into pixels and then Krige at each u_j as outlined in Section 2.7.

3.2 Step 2: Create a similarity map

The SSIM [44] was first designed as a quality index for images and then later on used to test the similarity between images. In this step of our proposed similarity test, we use the SSIM index to obtain a similarity map between the two spatial data sets. For this, we use the two pixel images constructed for the spatial data sets in the previous step as the input images for this algorithm. For illustration purposes, let us consider the two pixel images in Figure 3.6. Each of the pixel images has 49 pixels with one value for each pixel. For the calculation of the SSIM, the values of the pixels should be real-valued and not discrete. The images are ultimately considered in the calculations as an array rather than an image.

The SSIM algorithm [44] uses a sliding window approach to move across the image pixel-by-pixel simultaneously for the two images. For each sliding window, an SSIM value is calculated for the centre pixel. In our approach, we always use an odd number of pixels as the length and width. This is so that the pixel considered is right at the centre of the sliding window. For this example, we consider a sliding window of

0.014	0.729	0.25	0.161	0.017	0.486	0.103
0.748	0.823	0.955	0.685	0.501	0.275	0.229
0.771	0.356	0.536	0.093	0.17	0.9	0.423
0.615	0.775	0.356	0.406	0.707	0.838	0.24
0.358	0.429	0.052	0.264	0.399	0.836	0.865
0.272	0.616	0.43	0.652	0.568	0.114	0.596
0.507	0.307	0.427	0.693	0.085	0.225	0.275

(a)

0.494	0.653	0.329	0.864	0.638	0.014	0.529
0.642	0.277	0.104	0.256	0.058	0.247	0.215
0.125	0.398	0.505	0.328	0.412	0.202	0.813
0.68	0.364	0.35	0.062	0.483	0.399	0.016
0.734	0.574	0.482	0.331	0.158	0.48	0.204
0.29	0.881	0.123	0.175	0.441	0.907	0.851
0.277	0.001	0.511	0.014	0.065	0.955	0.086

(b)

Figure 3.6: Two sample pixel images for illustration.

size 3×3.

Figure 3.7 shows the two pixel images from Figure 3.6 each with two different sliding windows. The red sliding windows represent the scenario when the centre pixel is close to the border causing the sliding window to have empty pixels. In this case, we only consider the pixels of the sliding window overlapping with the pixel image, the rest of the pixels are omitted. The second scenario, shown by the blue sliding window, is when the centre pixel is closer to the middle of the image.

	0.014	0.729	0.25	0.161	0.017	0.486	0.103
	0.748	0.823	0.955	0.685	0.501	0.275	0.229
	0.771	0.356	0.536	0.093	0.17	0.9	0.423
	0.615	0.775	0.356	0.406	0.707	0.838	0.24
	0.358	0.429	0.052	0.264	0.399	0.836	0.865
	0.272	0.616	0.43	0.652	0.568	0.114	0.596
	0.507	0.307	0.427	0.693	0.085	0.225	0.275

(a)

	0.494	0.653	0.329	0.864	0.638	0.014	0.529
	0.642	0.277	0.104	0.256	0.058	0.247	0.215
	0.125	0.398	0.505	0.328	0.412	0.202	0.813
	0.68	0.364	0.35	0.062	0.483	0.399	0.016
	0.734	0.574	0.482	0.331	0.158	0.48	0.204
	0.29	0.881	0.123	0.175	0.441	0.907	0.851
	0.277	0.001	0.511	0.014	0.065	0.955	0.086

(b)

Figure 3.7: Two examples of where the sliding window may occur. The red window is an example of what happens when the centre pixel of the sliding window occurs on the border of the image while the blue window is an example of a centre pixel in the centre of the image.

The next step in the SSIM algorithm is to calculate an SSIM value for the centre pixel in the sliding

window. As mentioned before in Section 2.5, the SSIM consists of three terms, namely contrast, structure and luminance, which are calculated separately.

The values from the sliding window are used as a vector of values for each image. Consider the example and the sliding windows indicated in Figure 3.7.

To calculate the three terms for the SSIM, we need the mean, variance and covariance of the vectors of pixel values. The means for the red sliding window are calculated as: $\mu_{\boldsymbol{x_1}} = 0.5785$ and $\mu_{\boldsymbol{x_2}} = 0.5165$. The means for the blue sliding window are calculated as: $\mu_{\boldsymbol{x_1}} = 0.426$ and $\mu_{\boldsymbol{x_2}} = 0.2894444$. The variance and covariance values for the red sliding window are as follows: $\sigma^2_{\boldsymbol{x_1}} = 0.1433$, $\sigma^2_{\boldsymbol{x_2}} = 0.0307$ and $\sigma_{\boldsymbol{x_1},\boldsymbol{x_2}} = -0.0013$. The variance and the covariance values for the blue sliding window are as follows: $\sigma^2_{\boldsymbol{x_1}} = 0.0403$, $\sigma^2_{\boldsymbol{x_2}} = 0.0266$ and $\sigma_{\boldsymbol{x_1},\boldsymbol{x_2}} = -0.0032$.

The last values needed to calculate the luminance, contrast and structure terms are the constants. As mentioned in Section 2.5, the constants are used in order to avoid inconsistency [45]. The constants are $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$ and $C_3 = \frac{C_2}{2}$ where we choose $K_1 = 0.01$ and $K_2 = 0.03$ [45]. Also, L is the range of pixel values in the image. It is the difference between the maximum pixel value from the two images and the minimum pixel value. The three terms are calculated separately and multiplied together for the SSIM value.

$$\text{Luminance:} \quad \ell(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\mu_{x_1}\mu_{x_2} + C_1}{\mu^2_{x_1}\mu^2_{x_2} + C_1} \tag{3.6}$$

$$\text{Contrast:} \quad c(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\sigma_{x_1}\sigma_{x_2} + C_2}{\sigma^2_{x_1}\sigma^2_{x_2} + C_2} \tag{3.7}$$

$$\text{Structure:} \quad s(\boldsymbol{x_1}, \boldsymbol{x_2}) = \frac{2\sigma_{x_1,x_2} + C_3}{\sigma_{x_1}\sigma_{x_2} + C_3}. \tag{3.8}$$

Figure 3.8 shows the SSIM values for each of the pixels in the image. The result in Figure 3.8 forms the local similarity map between the two pixel image representations obtained from Step 1. In the case of a non-rectangular pixel image, that is when some of the pixel values are omitted, the calculation is the same as for the rectangular pixel image. The difference being that if a pixel values are omitted from the pixel image representations, the corresponding pixel in the local similarity map is also omitted.

In the case of discrete, specifically categorical, pixel values, the SSIM is not sensible to compare the images. In such a case, we compare the pixel values directly. This means that if the pixel in position (i, j) from the first image is the same as the corresponding pixel from the second image, then the pixel in the same position in the similarity map has a value of 1. If the two pixels are not the same, the pixel in the similarity map has a value of -1.

-0.011	-0.186	-0.583	-0.75	-0.696	-0.92	-0.887
-0.14	-0.247	-0.406	-0.606	-0.595	-0.45	-0.262
-0.209	-0.324	-0.179	-0.173	0.029	0.119	0.145
-0.58	-0.331	0.017	0.037	0.14	0.042	-0.003
0.151	0.224	0.056	-0.076	-0.191	-0.11	-0.223
0.412	0.251	-0.056	-0.212	-0.322	-0.088	-0.411
0.586	0.501	0.121	0.069	-0.465	0.127	0.008

Figure 3.8: The resulting SSIM values for each pixel.

3.3 Step 3: Calculate global similarity index

In this final step of the proposed spatial similarity test, we calculate the percentage of similarity between the two spatial data sets. Up to now in the test, we have represented the different spatial data sets as a pixel image. This is done differently for each of the spatial data types. The reason for the pixel image representation is so that different spatial data types is transformed to a general type that eases the following steps. In the next step, the pixel images created in the first step are used to create a local similarity map.

From the similarity map in the second step, we calculate a global similarity index that is the final result of the test. In the case of continuous pixel values in the local similarity map, the global similarity is calculated similarly as Andresen's S-Index in Equation (2.21),

$$GS = \frac{\sum_{j=1}^{M} SSIM(u_j)}{M}, \tag{3.9}$$

where $SSIM(u_j)$ is the SSIM value for the pixel with centroid $\boldsymbol{u_j}$ and M the number of pixels in the pixel image. $SSIM(u_j)$ is a non-binary input for Andresen's S-Index. This is expected to improve the accuracy of the test by providing a mean similarity value instead of a proportion of similar areas within the domain.

In the case of the similarity map containing discrete values, the global similarity is calculated as a proportion of similar values as indicated by the local similarity map.

3.4 Conclusion

In this chapter, the proposed spatial similarity test was presented. The test consists of three steps and is generic by design to handle any type of spatial data. The first step in this test is to create a pixel image representation from each of the two spatial data sets being compared. This step is the most involved of the three steps as it is done differently for each type of spatial data. In the second step, a local similarity map is created from the two pixel images and the third step involves the calculation of a similarity index using the similarity map. The calculation of the similarity index in step three is based on the calculation of the S-index from Andresen's spatial point pattern test.

Chapter 4

Simulation study

In this chapter, a simulation study is conducted to test this method on the various spatial data types. This is a popular method to test a statistical method [34]. It involves the creation of data with the main reason that the user knows what the outcome of the method should be. In our simulation study, we simulate several data sets for each of the spatial data types considered. Each data type is handled separately to see how this method reacts in each case.

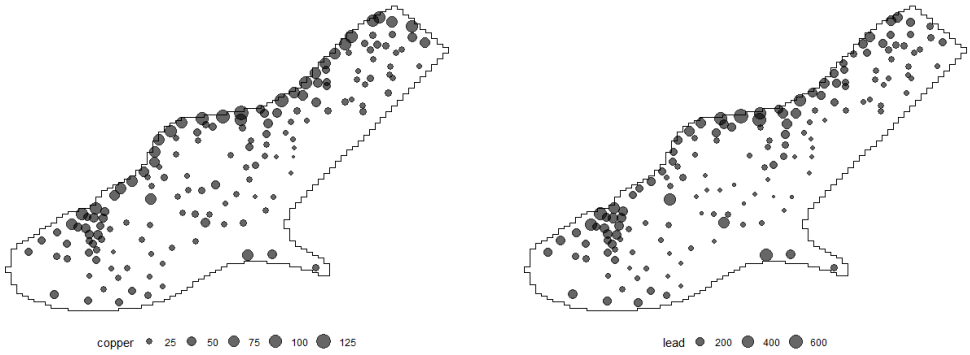
Seeing that we developed a method to test the similarity of spatial data, we want to simulate spatial data sets to compare that are known to be either 80% or 90% identical. To do this, we simulate several spatial data sets to be used as X_1. For X_2, a certain percentage of the data points are replaced with some other data points. After this, we expect the comparison between each pair of data sets, should yield an answer of about 80% or 90%.

Recall that the first step of the test is to obtain a pixel image representation of the spatial data sets. As mentioned in Chapter 3, the resolution of the pixel image representation should be decided by the user before-hand. In our simulation, we are also interested to explore the influence of the resolution of the pixel image on the outcome of the test, as it is user-defined. For this reason we repeat the test for each comparison for three different resolutions. We use a 10×10 image, 20×20 image and a 50×50 image.

We start with the geostatistical simulations, followed by the lattice data and then the different point patterns. A detailed discussion is included at the end of the chapter.

4.1 Geostatistical simulations

For the geostatistical simulations, a built-in `R` data set is used. This data set is contained within the `sp` package [8, 37] and is called `meuse`. An illustration of the data set is given in Figure 4.1. The data set consists of 155 spatial locations with six different measurements taken at each point. Measurements were taken of metals in the topsoil alongside the Meuse river flowing through France, Belgium and Netherlands.



(a) (b)

Figure 4.1: Examples of two X_1 geostatistical data sets of two metals observed in the top soil alongside the Meuse river. (a) Measurements of the copper and (b) Measurements of the lead.

The two data sets to compare are obtained by taking the spatial locations and each of the measurement (separately) as the data sets used as X_1 in the test. Then, the X_2 data set is obtained by randomly removing and replacing either 10% or 20% of the locations, attributes or both. In the case where the spatial locations are replaced, either 10% or 20% of the locations within the data set is replaced with other simulated spatial points. The attributes remain unchanged. When the attributes are changed, the spatial locations remain the same but new measurements are simulated as random uniform numbers. These values are simulated to be between the minimum and maximum of the original values. When both the locations and the attributes are changed, the above mentioned is done simultaneously.

Figure 4.2 is a visual representation of the results from applying the proposed spatial similarity test to the geostatistical simulations for different pixel image resolutions. Figure 4.2(a), (c) and (e) shows the results from all the simulations where X_1 and X_2 are 80% identical and Figure 4.2(b), (d) and (f) shows the results from all the simulations where X_1 and X_2 are 90% identical. Figure 4.2(a) and (b) are the visual representation of the results where the spatial locations are changed in X_2. Figure 4.2(c) and (d) are the visual representation of the results where the attributes are changed and Figure 4.2(e) and (f) are the visual representation of the results where both the spatial locations and the attributes are changed.

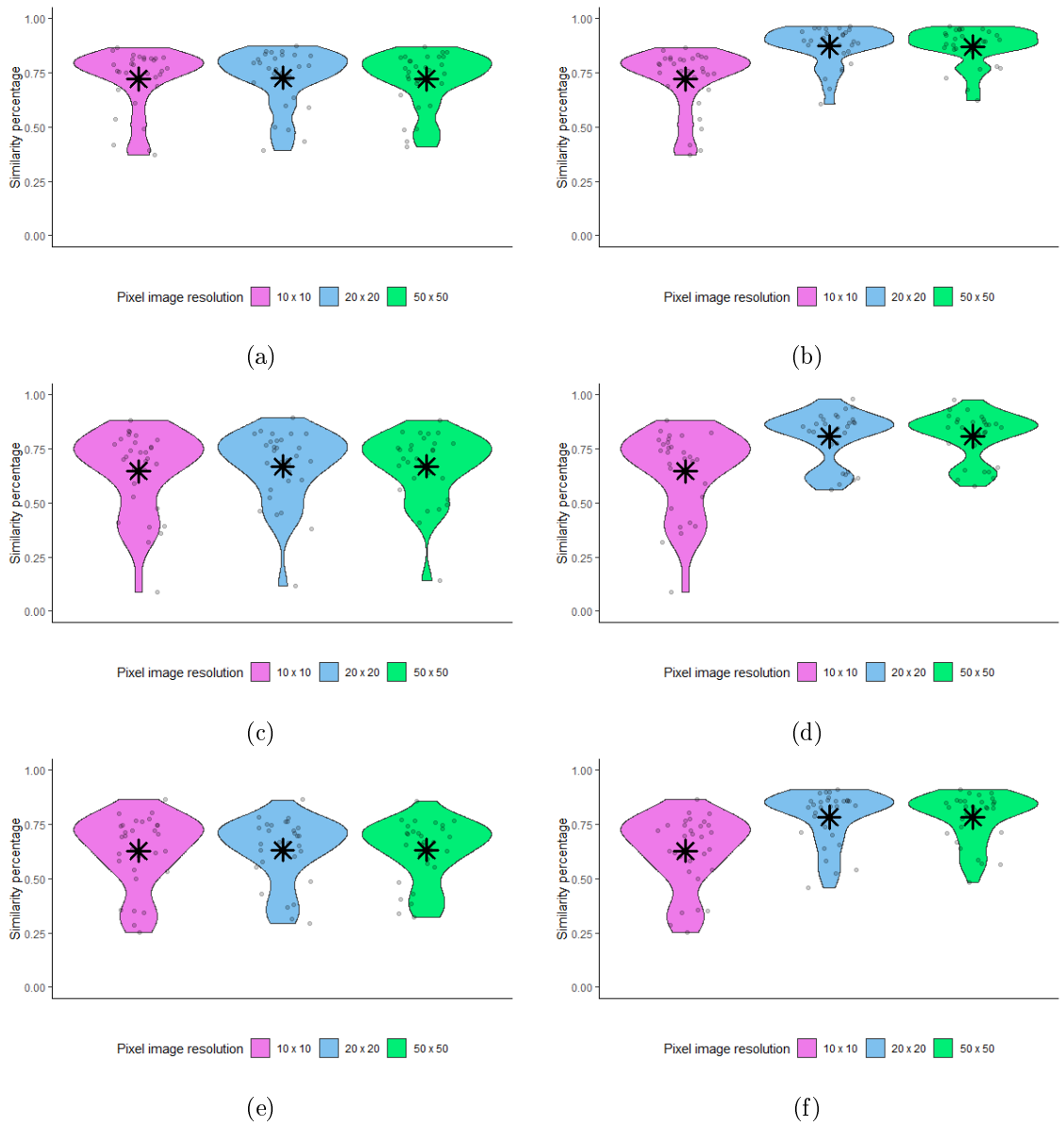
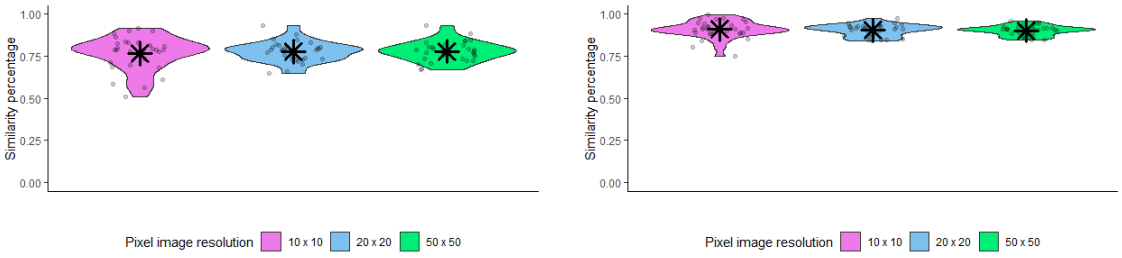


Figure 4.2: Visual representation of the results from applying the proposed spatial similarity test to the geostatistical simulations to different pixel image resolutions. (a), (c) and (e) represent the results where the geostatistical data sets are 80% identical and (b), (d) and (f) represent the results where the geostatistical data sets are 90% identical. (a) and (b) represent the results of the geostatistical simulations where the spatial locations are changed while all the attributes remained the same. (c) and (d) represent the results of the geostatistical simulations where the attributes are changed while all the spatial locations remained the same. (e) and (f) represent the results of the simulations where both the spatial locations and the attributes are changed. The mean for each pixel image resolution group are indicated with a star.

4.2 Lattice data simulations

To simulate lattice data sets, we use the South African borders as the spatial domain and the spatial locations as the municipalities in South Africa. This is shown in Figure 4.3. The values for each spatial location is simulated as random uniform numbers. For these values, there are three groups where the range of values differ. The first group of data sets has simulated values between 0 and 50, the second between 0 and 100 and the third between 0 and 1000. To obtain the testing data sets, either 10% or 20% of the values are removed and replaced with other random uniform numbers within the same range. Figure 4.4(a) is a visual representation of the results from applying the proposed spatial similarity test to the lattice data sets which are 80% identical for different pixel image resolutions. Figure 4.4(b) is the representation of results for the lattice data sets which are 90% identical.

Figure 4.3: The spatial domain which is used for the simulation of the lattice data sets. The South African borders are used as the spatial domain and the separate municipalities as the spatial locations.



(a) (b)

Figure 4.4: Visual representation of the results from applying the proposed spatial similarity test to the lattice simulations to different pixel image resolutions. (a) represents the results from the data sets being compared that are 80% identical and (b) represents the results from the data sets being compared that are 90% identical. The mean for each pixel image resolution group are indicated with a star.

4.3 Point pattern simulations

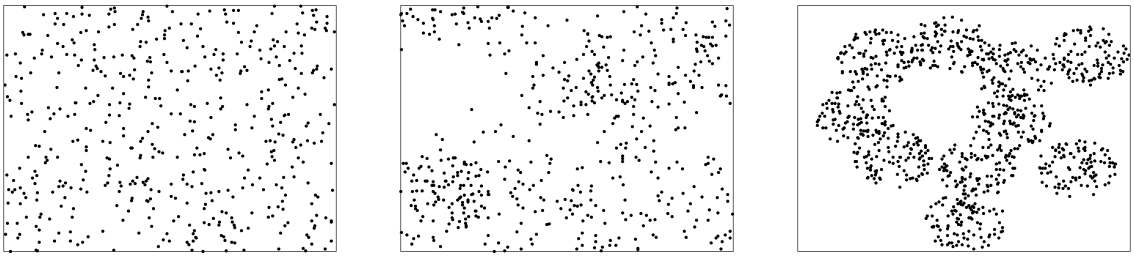
When simulating spatial point patterns, it is important to cover many possible scenarios. Therefore, we simulate regular as well as clustered spatial point patterns on both a rectangular and polygonal window. The spatial point patterns are simulated with different intensities (constant and non-constant). Also, for three different pattern sizes: Small (±100 points), Medium (±500 points) and Large (±1000 points).

The simulations of the spatial point patterns are done by using built-in `R` functions. The function that we use to simulate the regular spatial point patterns is the `rSSI` function [6] while the clustered spatial point patterns are simulated with the `rMatClust` function [6].

To add more variety to the simulation study, we use three approaches for the simulations. The first approach being to create noisy patterns. In this approach, the regular and clustered point patterns is simulated with the above functions. When we replace some of the data points to create X_2, the spatial data points is replaced with any other simulated points. In the case of clustered spatial point patterns, it creates visible noise within the pattern.

The goal of the second simulation approach is to create spatial point patterns with strong clusters. For this approach, the centres are simulated as a regular spatial point pattern with a large inhibition distance. The clusters are then simulated as discs around these points. The replaced data points are then simulated to be contained within these strong clusters. With the third approach, we create a comparison with uneven patterns. This happens by only removing either 10% or 20% of the spatial data points from X_1 to X_2.

4.3.1 Unmarked point patterns



(a) (b) (c)

Figure 4.5: Examples of some of the X_1 data sets for simulated unmarked point patterns. (a) and (b) are simulations using the first method of simulations with (a) being the regular pattern and (b) the noisy clustered pattern. (c) is a simulation from the second method of simulations where the aim is to have strict clusters in the pattern.

For the unmarked spatial point patterns, the above simulations are used as is. The method is applied to each of the X_1 and X_2 pairs, Figure 4.5 shows three examples of X_1 data sets. Figure 4.5(a) and (b) are from the first method of simulations with Figure 4.5(a) a regular point pattern and (b) a clustered point pattern which contains some noise in between the clusters. Figure 4.5(c) is an example of the second method of simulations with the strict clusters. Hence, there is no noise between the clusters. The replaced points in X_2 are simulated to be again within the strict clusters. The data sets from the third method of simulations are similar to the simulations from the third method.

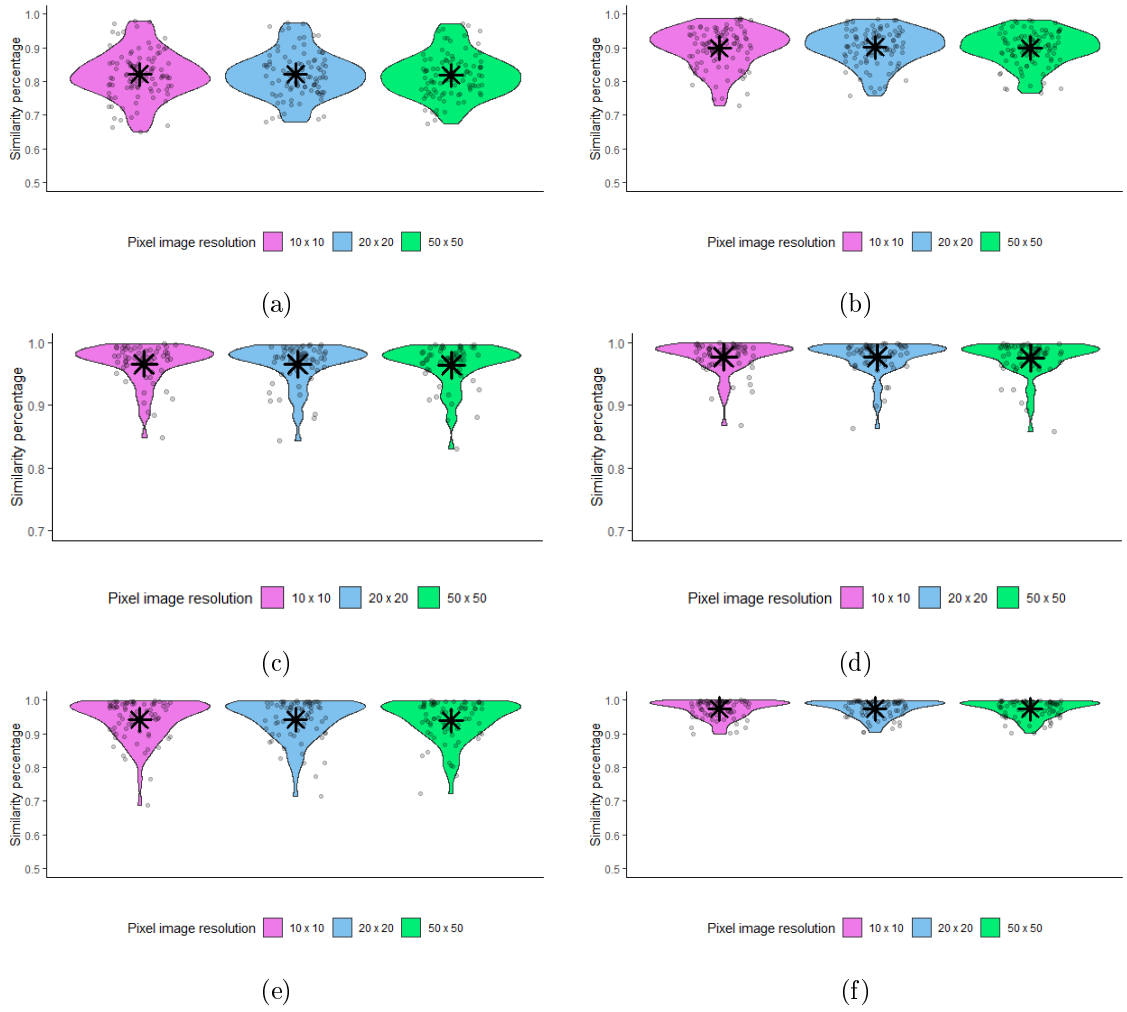


Figure 4.6: Visual representation of the results from applying the proposed spatial similarity test to the point pattern simulations to different pixel image resolutions. (a), (c) and (e) represent the results where the unmarked point pattern data sets are 80% identical and (b), (d) and (f) represent the results where the unmarked point pattern data sets are 90% identical. (a) and (b) represent the results of the first method of simulations. (c) and (d) represent the results of the second method of simulations. (e) and (f) represent the results of the third method of simulations. The mean for each pixel image resolution group are indicated with a star.

Figure 4.6(a) and (b) are visual representations of the results from the first method of simulations. Figure 4.6(c) and (d) are visual representations of the results from the second method of simulations. Figure 4.6(e) and (f) are visual representations of the results from the third method of simulations.

4.3.2 Continuous marked point patterns

The simulation of the marked point patterns is done by taking the unmarked point patterns from the first method of simulations and simply adding a continuous value for the mark. This continuous value is simulated as random uniform numbers. For these values, there are three groups where each group have a different range of values. For the first group, the random uniform numbers range from 0 to 20. For the second group, they range from 0 to 50. And for the last group, they range from 0 to 100. Figure 4.7 is a continuous marked pattern with marks between 0 and 50.

Figure 4.8 is a visual representation of the results from applying the proposed similarity test to the continuous marked point pattern simulations for different pixel image resolutions. Figure 4.8(a) and (b) are visual representations of the results where the locations were changed.Figure 4.8(c) and (d) are a visual representation of the results where the attributes are changed.

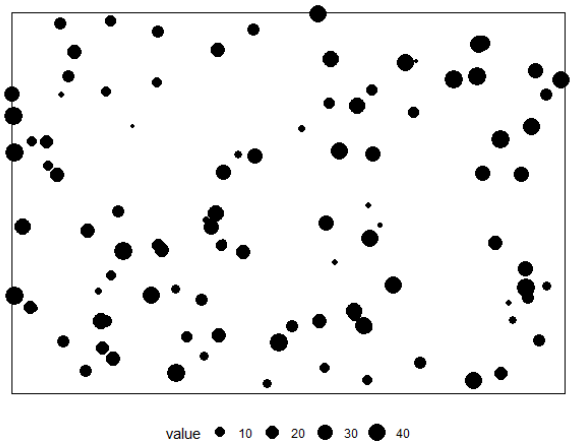
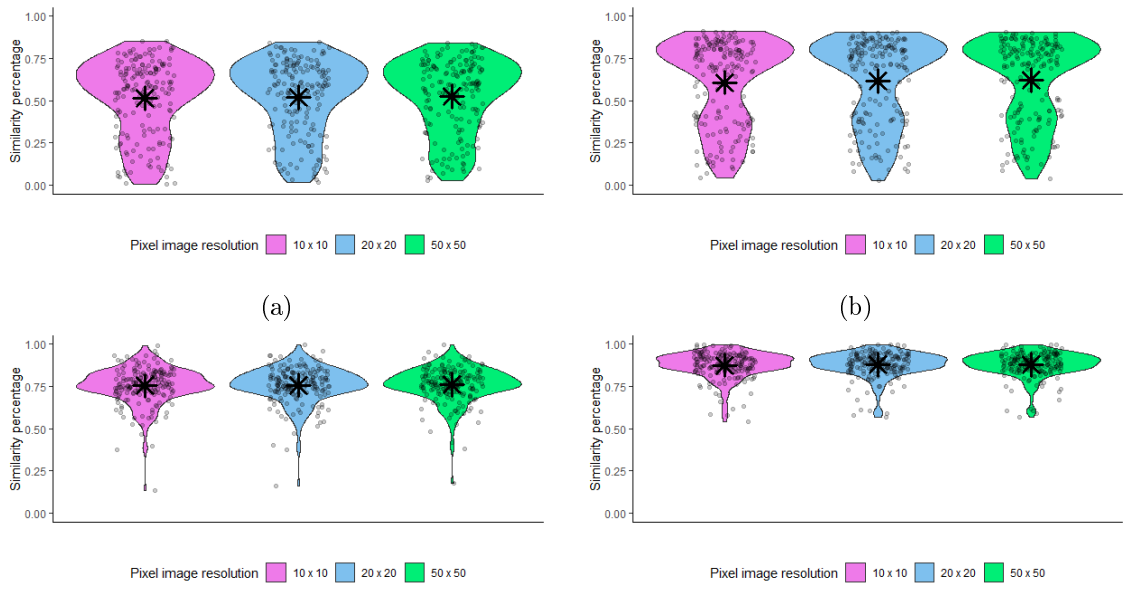


Figure 4.7: Example of one of the point patterns with continuous marks.



(a) (b)

(c) (d)

Figure 4.8: Visual representation of the results from applying the proposed spatial similarity test to the continuous marked point pattern simulations to different pixel image resolutions. (a) and (c) represent the results where the marked point pattern data sets are 80% identical and (b) and (d) represent the results where the marked point pattern data sets are 90% identical. (a) and (b) represent the results of the marked point pattern simulations where the spatial locations are changed while all the attributes remained the same. (c) and (d) represent the results of the marked point pattern simulations where the attributes are changed while all the spatial locations remained the same. The mean for each pixel image resolution group are indicated with a star.

4.3.3 Discrete marked point patterns

The simulation of the marked point patterns are done by taking the unmarked point patterns from the first method of simulations and simply adding a discrete value for the mark. This value is simulated so that the pattern has either two, three or four different categories. Figure 4.9 is one of the simulated point patterns with discrete marks used. This pattern has three categories.

Figure 4.10 is a visual representation of the results from applying the proposed spatial similarity test to the simulations with discrete marks. This is done for different pixel image resolutions. Figure 4.10(a) and (b) are visual representations of the results where the locations were changed.Figure 4.10(c) and (d) are a visual representation of the results where the attributes are changed.

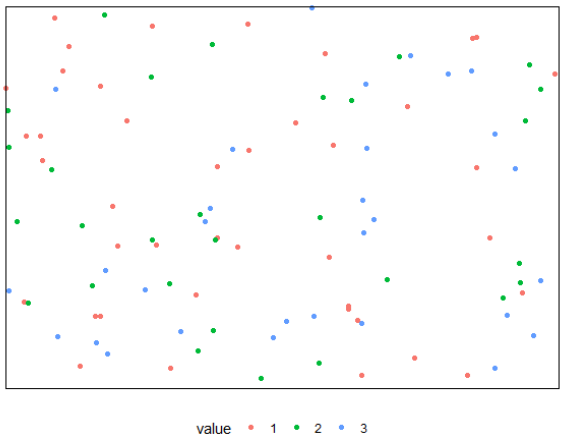


Figure 4.9: Example of one of the point patterns with discrete marks.

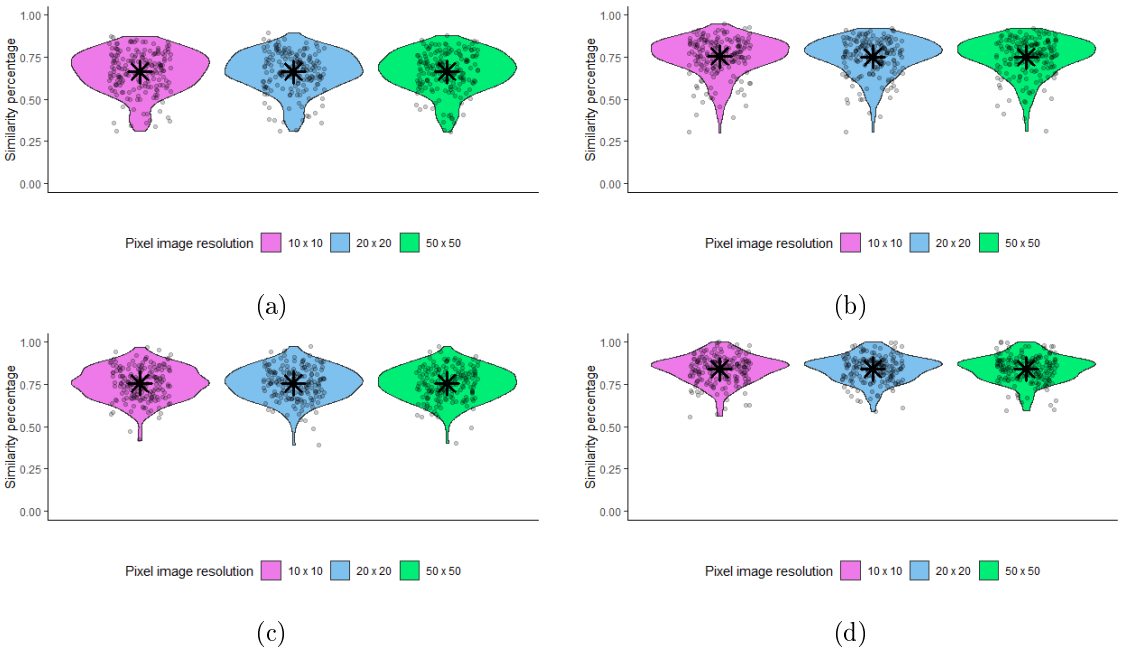


Figure 4.10: Visual representation of the results from applying the proposed spatial similarity test to the discrete marked point pattern simulations to different pixel image resolutions. (a) and (c) represent the results where the marked point pattern data sets are 80% identical and (b) and (d) represent the results where the marked point pattern data sets are 90% identical. (a) and (b) represent the results of the marked point pattern simulations where the spatial locations are changed while all the attributes remained the same. (c) and (d) represent the results of the marked point pattern simulations where the attributes are changed while all the spatial locations remained the same. The mean for each pixel image resolution group are indicated with a star.

4.4 Comparison of resolution choice

As can be seen from Figure 4.2, 4.4, 4.6, 4.8 and 4.10, the means of the different pixel image resolutions seem to be close to each other. If it is the case that the means can be classified as equal, we can consider this proposed spatial similarity test as not sensitive to the resolution of the pixel image representation. We test the hypothesis of equal means across the three groups to the alternative hypothesis of at least one of the means being unequal to the rest of them.

The Kruskal-Wallis test was applied to the results of the newly proposed similarity test to test whether the means of the different pixel image resolutions are equal. As the assumption of normality is rejected in all the cases at a 5% level of significance, the Kruskal-Wallis test was applied instead of an ANOVA test. Table 4.1 shows the p-values of the Kruskal-Wallis test.

Table 4.1: P-values of the Kruskal-Wallis test.

	P-value (80%)	P-value (90%)
Geostatistical data		
Locations changed	0.9076	< 0.0001
Attributes changed	0.8991	< 0.0001
Both changed	0.971	< 0.0001
Lattice data	0.9805	0.6028
Unmarked point patterns		
Method one	0.9047	0.832
Method two	0.923	0.9153
Method three	0.935	0.9456
Continuous marked		
Location changed	0.8263	0.7601
Attributes changed	0.8005	0.7765
Discrete marked		
Location changed	0.9997	0.9457
Attributes changed	0.9789	0.9688

From Table 4.1, it can be clearly seen that the hypothesis of equal means cannot be rejected in all the cases except for the geostatistical simulations that are 90% identical. From Figure 4.2(b), (d) and (f), it is visually clear that the mean of the results for the 10×10 pixel image resolution is lower than the means for the other two groups. From doing a pairwise Wilcoxon test it is concluded that the mean for

the 10×10 pixel image resolution differs from the other two means at a 5% level of significance, while the other two groups (20×20 and 50×50) does not differ significantly from each other at a 5% level of significance.

Table 4.2: Summary statistics of the results from the proposed spatial similarity test.

	Identical	Mean	Median	Standard Deviation	Coefficient of variation
Geostatistical data					
Locations changed	80%	0.7220	0.7645	0.1310	0.1814
	90%	0.8202	0.8577	0.1281	01562
Attributes changed	80%	0.6610	0.7047	0.1712	0.2590
	90%	0.7523	0.8144	0.1614	0.2145
Both changed	80%	0.6302	0.6691	0.1495	0.2372
	90%	0.7302	0.7740	0.1531	0.2097
Lattice data	80%	0.7740	0.7846	0.0751	0.0970
	90%	0.9043	0.9079	0.0394	0.0436
Unmarked point patterns					
Method one	80%	0.8195	0.8166	0.0667	0.0814
	90%	0.8992	0.9060	0.0536	0.0596
Method two	80%	0.9654	0.9755	0.0329	0.0341
	90%	0.9763	0.9847	0.0266	0.0272
Method three	80%	0.9409	0.9598	0.0591	0.0628
	90%	0.9732	0.9815	0.0252	0.0259
Continuous marked					
Locations changed	80%	0.5066	0.5736	0.2344	0.4627
	90%	0.6057	0.7178	0.2596	0.4286
Attributes changed	80%	0.7571	0.7656	0.1074	0.1419
	90%	0.8771	0.8860	0.0760	0.0866
Discrete marked					
Locations changed	80%	0.6624	0.6750	0.1187	0.1792
	90%	0.7514	0.7700	0.1104	0.1469
Attributes changed	80%	0.7550	0.7600	0.0942	0.1248
	90%	0.8399	0.8500	0.0785	0.0935

From the above, it can be concluded that the proposed spatial similarity test is not sensitive to the user-defined choice of the resolution of the pixel image representation. Seeing that the mean from smaller resolution for the geostatistical data differs from the rest of the means, while the finer resolutions did not differ from each other, it is advisable to rather use a finer pixel image resolution when working with geostatistical data.

4.5 Discussion

For a deeper look into the results from the proposed spatial similarity test for the different spatial data sets, we consider some summary statistics such as the mean, median, standard deviation and coefficient of variation. These values are given in Table 4.2. The method can be classified as accurate if the mean or the median is close to the known similarity of the data sets with a rather small standard deviation and coefficient of variation.

When looking at the geostatistical summary statistics in Table 4.2, it can be seen that the means and medians of the results do tend to the true similarity of the data sets with the values for the 80% similar spatial data pairs lower than for the 90% similar data. However, the standard deviations and the coefficients of variation are still large. The inaccuracy of this data type can be accounted for due to the method of Kriging that may be too general for the type of data used. A more optimal model for the Kriging may yield to better results [32].

From Table 4.2, the proposed spatial similarity test seems to compare the similarity between two lattice data sets quite accurately with the means and medians of the results close to the theoretical values. The standard deviations and the coefficients of variation are also small which is also an indication that this method performs well in the case of lattice data.

For the unmarked point patterns, it can be seen in Table 4.2 that the method does perform well on the simulations from the first method. This can be said since the means and the medians of the results are close to the theoretical values and the standard deviations and the coefficients of variation are small. However, for the strong clustered patterns (second method of simulations) and the unequal patterns (third method of simulations), this test yields large similarity values. For the second method of simulations, it may be the case that the way in which the pixel image representations are obtained may not be designed to pick up such small differences in the pattern. Recall that the second method of simulations is designed to keep the two patterns visually as similar as possible by simulating the original points as well as the replaced points within the same clusters. This case is highly theoretical and will possibly not occur in real life. In the third method of simulations, some of the points are removed to obtain the X_2. The reason the method may yield such high similarity values may be in the way in which the pixel image representations

are obtained.

This proposed spatial similarity test can still be improved to perform better on marked spatial point patterns with continuous values as the summary statistics in Table 4.2 show that the mean and the median is far from the theoretical values with large standard deviations and coefficients of variation. In the case where the locations of the points are changed, this method does not perform ultimately. This may be again due to the way in which the pixel image representation are obtained. When the attributes of some of the points are changed, this method performs better again. However, in reality it will happen more often that we have a scenario that the attributes are changed than the locations.

In the case of marked spatial point patterns with discrete values, the method performs better when the attributes are changed than when we change the locations. The summary statistics in Table 4.2 indicated closer means and medians to the theoretical values when the attributes are changed. The standard deviation and coefficient of variation is also smaller.

4.6 Conclusion

In this chapter, the proposed spatial similarity test was tested using a simulation study. Each type of spatial data, considered in this mini-dissertation, was simulated. For each comparison, an X_1 and X_2 data set were simulated. As we are interested in testing a similarity test, the two data sets being compared are simulated to be either 80% or 90% identical. Therefore we expect the test to result in either an 80% or 90% result.

The resolution for the pixel images are user-defined and therefore we were interested in providing a guideline for the user as to how to pick an appropriate resolution. The test for the simulations was performed on different pixel image resolutions and evaluated with a Kruskal-Wallis hypothesis test. It was concluded that the proposed spatial similarity test is not sensitive for the choice of the resolution of the pixel image. The only exception is in the case of the geostatistical simulations where the mean for the 10×10 pixel image resolution was significantly different from the means of the other two pixel image resolutions. It is then suggested that a finer pixel image resolution should be considered in the case of geostatistical data.

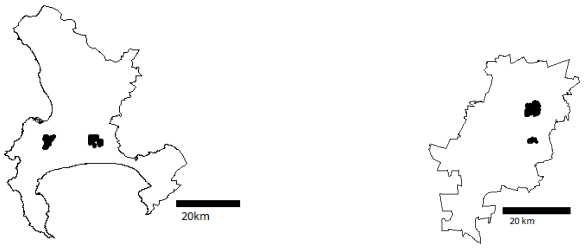
The proposed spatial similarity test works well on lattice data and in some unmarked point pattern cases. Some work can still be done on to improve the test on geostatistical data and marked point patterns.

Chapter 5

Application

5.1 Study area and data

A data set provided by Lightstone[1,2] consists of the evaluation prices of 1018 properties in the City of Cape Town and City of Johannesburg metros. In both these metros, there are two blocks of properties and each property has three evaluation prices, one for each 2017, 2018 and 2019. We apply the proposed spatial similarity test on each block within the two metros.



(a) (b)

Figure 5.1: Locations of the properties in the provided data set within the two metros. Each metro consists of two blocks of properties. (a) The City of Cape Town and (b) The City of Johannesburg.

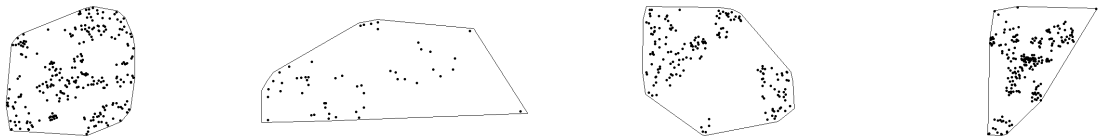
Figure 5.1(a) shows the property locations of the two blocks in the City of Cape Town metro and Figure 5.1(b) is the property locations of the two blocks in the City of Johannesburg. It is evident that property locations for the entire metro is not included in both the metros. Therefore, the four blocks of properties are handled separately. The window of each block of properties is obtained by taking the enclosing convex

[1] https://lightstone.co.za/

[2] Data was provided by Lightstone. The right to use this data was approved by the NAS ethics committee NAS078/2020.

hull around the points.

Figure 5.2(a)-(d) are separate spatial point patterns for the four blocks of properties. The prices of these properties are of interest over the three years. The price of each property is considered as the continuous mark in a marked spatial point pattern. Figure 5.2(a) and (b) are the two blocks in the City of Johannesburg metro and Figure 5.2(c) and (d) are the two blocks in the City of Cape Town metro.



(a) (b) (c) (d)

Figure 5.2: The four separate blocks of property locations that is considered in this application section. (a) and (b): Two blocks in the City of Johannesburg metro and (c) and (d): Two blocks in the City of Cape Town metro.

5.2 Analysis

Figure 5.3(a)-(d) is density plots of the property prices over the different years within the four blocks of properties. Figure 5.3(a) is the density plot of the property prices in the first block of properties in the City of Johannesburg metro and Figure 5.3(b) is the second block of properties. These blocks consist of 423 and 120 properties respectively. Figure 5.3(c) is the density of the property prices in the first block of properties in the City of Cape Town metro. This block consists of 168 properties. Figure 5.3(d) is the property prices in the second block of 307 properties.

Three comparisons are made for each of the four blocks. The first comparison is between the property prices of 2017 and the property prices of 2018. While the second comparison is between the prices of 2018 and 2019 and the third comparison is between the prices of 2017 and 2019. After the first step of obtaining the pixel image representations is done for the three comparisons is each of the four blocks, we create a local similarity map in the second step of the proposed spatial similarity test, see Figure 5.4. Figure 5.4 consists of the local similarity maps obtained by the proposed spatial similarity test.

These local similarity maps are useful in the sense that they allow the user to see where the potential differences lie between the two spatial data sets considered. They can also be used to identify the areas in the spatial data sets that have a high similarity between them. The global similarity index, from the third step, is calculated by simply taking the mean of the values from the local similarity map.

For the purpose of this chapter, the pixel image representations used have a resolution of 30×30. The

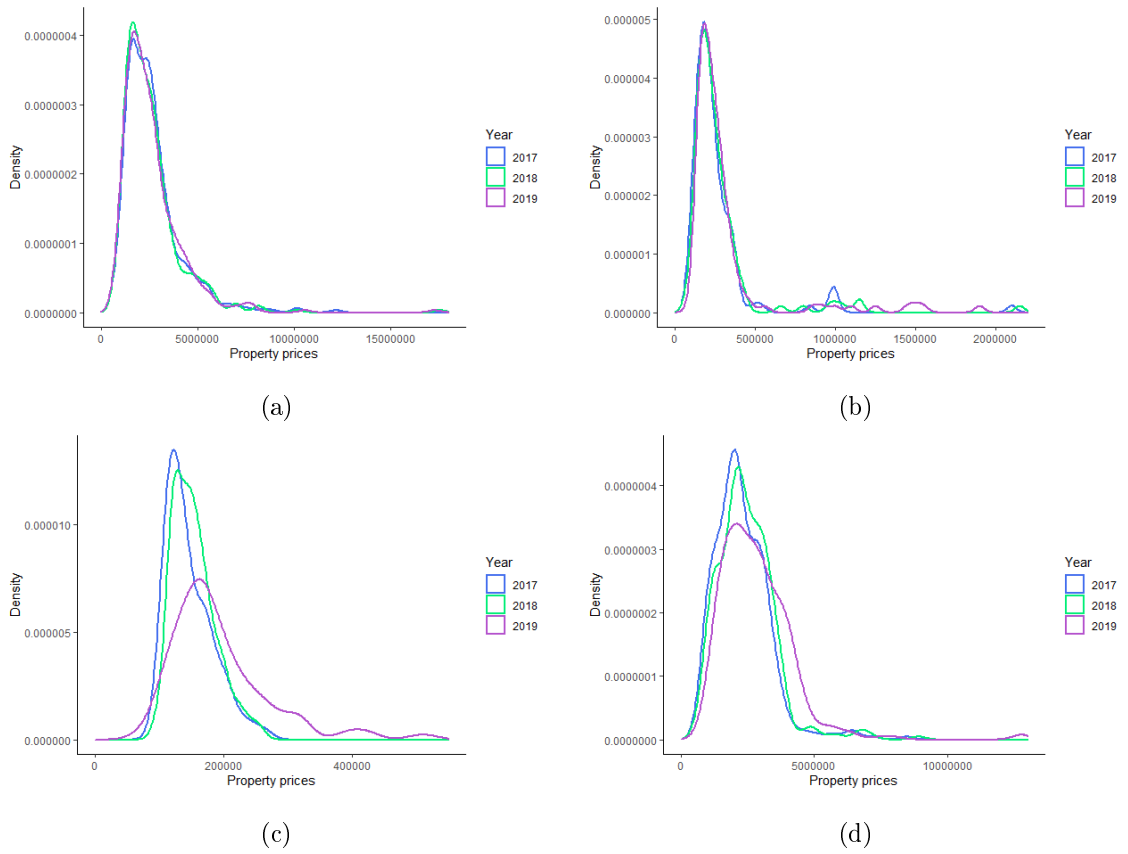


Figure 5.3: Density plots for the property prices for the different years within the four blocks of properties. (a) The property prices for the first block in Johannesburg and (b) the second block. (c) The property prices in the first block of properties in Cape Town and (d) the second block.

bandwidth used is Diggle's bandwidth [6]. The windows of the patterns are the same as displayed in Figure 5.2(a)-(d) which is obtained by taking the enclosed convex hull around the points. The sliding window in the SSIM calculation is chosen to be of size 11×11. This choice is made with reference to [10, 45].

The similarity maps show a high similarity between the property prices across years of three of the blocks of properties (Johannesburg Block 1 and 2, Cape Town Block 2). Lower similarity is observed for all three of the comparisons of the first block of properties in the City of Cape Town metro.

The global similarity indices can be seen in Table 5.1. These values support the observations from the similarity maps in Figure 5.4. It also indicated that something significant happened with the property prices from the first block of the City of Cape Town metro.

When comparing the similarity of the property prices between the different areas, we need to take a slightly different approach. The first step of the test where the data sets are represented as pixel images

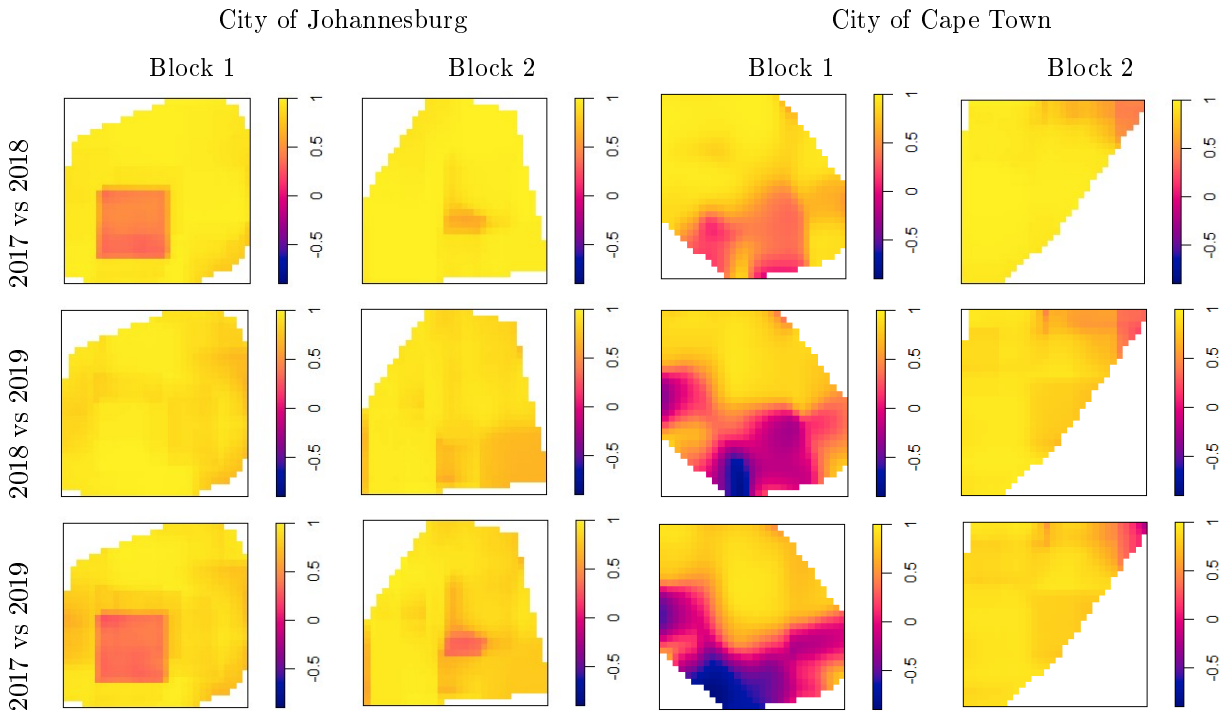


Figure 5.4: Local similarity maps for each comparison done on the four blocks of property prices by year.

stays the same. When calculating the local similarity map, we first determine which pixels of the pixel images contain values in both representations. In the case where either none of them or only one of them contains values, the pixel are omitted from the local similarity map. The calculation of the final similarity measure stays the same again.

With the comparison of spatial data with differing windows, it should be taken into account that the areas cannot always be compared as is. Therefore, we consider numerous rotations of one of the data sets before applying the proposed spatial similarity test. This gives us a better idea of the spatial similarity between the two data sets observed over different areas as compared to only looking at one of the rotations. The first data set in the comparisons stays as is whereas the second data set is rotated at 0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°.

We consider the following comparisons between the areas:

1. Johannesburg Block 1 vs Johannesburg Block 2
2. Johannesburg Block 1 vs Cape Town Block 1
3. Johannesburg Block 1 vs Cape Town Block 2
4. Johannesburg Block 2 vs Cape Town Block 1

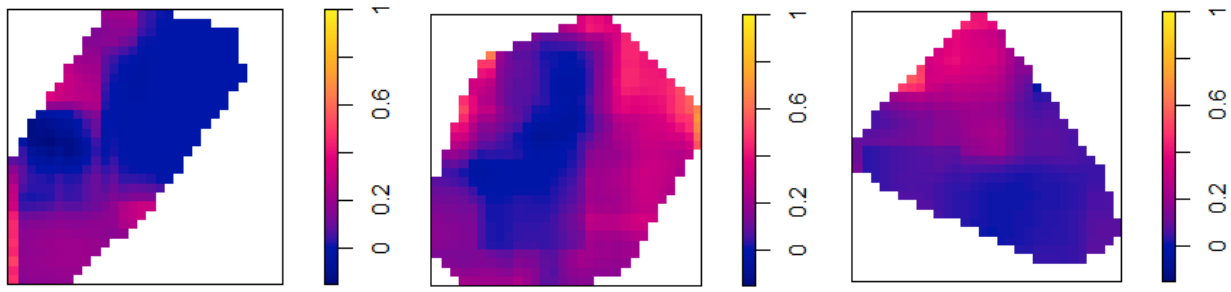
	City of Johannesburg		City of Cape Town	
Comparison	Block 1	Block 2	Block 1	Block 2
2017 vs 2018	0.8696	0.9636	0.7371	0.9216
2018 vs 2019	0.9105	0.8773	0.4182	0.8407
2017 vs 2019	0.7969	0.8527	0.3183	0.8342

Table 5.1: Similarity indices from the newly proposed similarity test

5. Johannesburg Block 2 vs Cape Town Block 2

6. Cape Town Block 1 vs Cape Town Block 2

Table 5.2 contains the results from applying the proposed spatial similarity test to the property prices between the four different blocks on all the rotations. From this, it can be concluded that there exists differences between the property prices among the blocks. It can also be seen that although there are slight differences in the results between the rotations, the values are close to each other. It will be beneficial to do a more in depth study on the optimal rotations when comparing spatial data on different domains.



(a) (b) (c)

Figure 5.5: Local similarity maps of three comparisons whose results are in Table 5.2. (a) Comparison of the second block in Johannesburg and the second block in Cape Town for the year 2017 where none of the data sets are rotated. (b) Comparison of the first block in Johannesburg and the first block in Cape Town for the year 2018 where the Cape Town block is rotated 180°. (c) Comparison of the first block in Cape Town and the second block in Cape Town for the year 2019 where the second block in Cape Town is rotated 45°.

Three local similarity maps from applying the proposed spatial similarity test on the six comparisons and the different rotations are shown in Figure 5.5. Figure 5.5(a) is the comparison of the second block in Johannesburg and the second block in Cape Town for the year 2017 where none of the data sets are rotated. Figure 5.5(b) is the comparison of the first block in Johannesburg and the first block in Cape

Town for the year 2018 where the Cape Town block is rotated 180°. Figure 5.5(c) is the comparison of the first block in Cape Town and the second block in Cape Town for the year 2019 where the second block in Cape Town is rotated 45°.

Table 5.2: Results from applying the proposed similarity test on the property prices between four different blocks in the data set. The largest values for each rotation and year is shown in italics.

		Comparison					
Rotation	Year	1	2	3	4	5	6
0°	2017	0.0565	*0.1204*	0.0729	0.0718	0.0719	0.0333
	2018	0.0571	0.1662	0.0449	0.0695	0.0678	0.0431
	2019	0.0498	*0.1725*	0.0463	0.0629	0.0549	0.1105
45°	2017	0.0511	0.0655	0.0472	0.0843	0.0795	*0.1104*
	2018	0.0430	0.0741	0.0176	0.0776	0.0707	*0.1062*
	2019	0.0357	0.1258	0.0308	0.0667	0.0631	*0.1216*
90°	2017	0.0343	0.0000	0.0533	0.0807	0.0665	0.0522
	2018	0.0323	0.0503	0.0501	0.0768	0.0754	0.0684
	2019	0.0291	0.0690	*0.0905*	0.0707	0.0647	0.0473
135°	2017	*0.0743*	0.0381	0.0000	0.0663	0.0928	0.0988
	2018	*0.0735*	0.1300	0.0491	0.0662	0.0855	0.0532
	2019	*0.0548*	0.1347	0.0165	0.0520	0.0667	0.0433
180°	2017	0.0398	0.1200	*0.0795*	*0.0934*	*0.1186*	0.0777
	2018	0.0374	*0.1678*	0.0668	*0.0834*	*0.1291*	0.0868
	2019	0.0328	0.1607	0.0647	0.0712	0.0969	0.1036
225°	2017	0.0255	0.0398	0.0314	0.0791	0.1183	0.0640
	2018	0.0257	0.0796	*0.0844*	0.0786	0.1202	0.0939
	2019	0.0229	0.0450	0.0672	*0.0755*	*0.0974*	0.0845
270°	2017	0.0550	0.1041	0.0458	0.0731	0.0232	0.0159
	2018	0.0476	0.1538	0.0448	0.0713	0.0227	0.0161
	2019	0.0370	0.1670	0.0326	0.0581	0.0238	0.0484
315°	2017	0.0518	0.0753	0.0000	0.0831	0.0275	0.0851
	2018	0.0440	0.0808	0.0000	0.0768	0.0263	0.0851
	2019	0.0370	0.1335	0.000	0.0665	0.0196	0.0896

5.3 Conclusion

In this chapter, the proposed spatial similarity test was applied to property prices in Cape Town and Johannesburg. Two blocks of properties were considered separately in each city. The proposed spatial similarity test was used to test the similarity of property prices within each block between 2017 and 2018, 2018 and 2019 and lastly 2017 and 2019. High similarity is observed between the property prices across the years for the two block of properties in Johannesburg and one block of properties in Cape Town. Low similarity is observed between the property prices of different years for one block of properties in Cape Town.

This test was also applied to test the spatial similarity of the property prices between the blocks of properties. Different rotations of X_2 were considered. Low similarity was observed for these comparisons.

From Figure 5.3(a) and (b), the property prices are relatively similar between the different years. This can also be seen in the similarity maps in Figure 5.4(a) and (b) which indicates a high similarity. Figure 5.3(c) and (d) shows that the property prices for these properties differ between the years. This relates to Figure 5.4(c) that also indicates low similarity between the property prices. However, Figure 5.4(d) indicates a higher similarity than expected after seeing the density plot. This indicates that our proposed method works well to identify spatial similarity between property prices.

Chapter 6

Conclusion

Up to now in literature, only a few spatial similarity tests have been developed. These test the similarity between two spatial data sets for only a certain type of data. In this mini-dissertation, a new spatial similarity test is proposed. This test determines the spatial similarity between two spatial data sets of any type, namely geostatistical data, lattice patterns, unmarked point patterns and marked point patterns.

The proposed spatial similarity test consists of three steps. The first being where the spatial data set is represented as a pixel image. This is obtained differently for each type of spatial data. In the second step, a local similarity map is created that shows where the two data sets are locally similar and where they differ. The final global similarity measure is calculated in the third step of the test by using the values from the local similarity map. Future work that can be considered is to extend the final global similarity measure to take the variation of the pixel values in the local similarity map into account.

In Chapter 4, a simulation study was done to test the accuracy of the proposed test. For a future study, a larger simulation study is suggested. A larger simulation study will allow more variation to be covered. The simulation can also be done by using real data and changing some of the data points to mimic the similarity aspect. The simulations will then be less theoretical and more realistic.

In the first step of the spatial similarity test for geostatistical data, investigation on the influence of the specific Kriging method on the outcome of the test should be done [32]. This will bring insight in choosing the optimal Kriging model when applying the test. For the lattice data, the pixel image representation can be obtained by using a more refined method. Instead of only assigning the value of the spatial location in which the centroid of the pixel falls to the specific pixel, a weighted average across the spatial locations falling within the pixel to calculate the value of that pixel could be more representative.

In the case of unmarked point patterns and marked point patterns with continuous marks, a suggestion

for a future study can be to optimise the bandwidth selection [6]. A study can also be conducted to investigate the influence of the bandwidth on the outcome of the test. For marked point patterns with discrete marks, the choice of k can be investigated. Recall that k is the number of closest points considered in the estimation of k nearest neighbour classification. Guidelines can also be put in place on how to choose the value of k such that the test gives the most accurate result.

It is also possible to vary the α, β and γ parameters within the SSIM calculation [11]. This adjusts the importance of each component in the calculation. A study can be done on the influence of the change in these parameters.

The proposed spatial similarity test was applied to property prices in Chapter 5. We considered four blocks of properties with prices over three years. We first used the test to compare the property prices over different years within the same block of properties. A future study can investigate the possibility of using the proposed spatial similarity test to test for a trend in longitudinal data. The test indicating similarity can then be an indication of a lack in trend. With this, is will be valuable to extend the proposed methodology so that more than one spatial data set can be compared simultaneously. Then the test was applied between the property prices of the different blocks. For this, different rotations of one of the data sets are considered.

In this mini-dissertation, we:

- Proposed a new generalised spatial similarity test
- Developed a method to create a pixel image representation for each type of spatial data
- Obtained local similarity maps by comparing the pixel image representations
- Developed a global similarity index that is calculated from the local similarity map
- Extended the S-index from Andresen's spatial point pattern test to have non-binary input.
- Applied this test on the property prices in Johannesburg and Cape Town over different years
- Applied this test on the property prices of the same year over different areas

Bibliography

[1] MV Alba-Fernández, FJ Ariza-López, MD Jiménez-Gamero, and J Rodríguez-Avi. On the similarity analysis of spatial patterns. Spatial Statistics, 18:352–362, 2016.

[2] MA Andresen. Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. Applied Geography, 29(3):333–345, 2009.

[3] MA Andresen. An area-based nonparametric spatial point pattern test: The test, its applications and the future. Methodological Innovations, 9, 2016.

[4] MA Andresen and SJ Linning. The (in) appropriateness of aggregating across crime types. Applied Geography, 35(1-2):275–282, 2012.

[5] A Baddeley, I Bárány, and R Schneider. Spatial point processes and their applications. Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004, pages 1–75, 2007.

[6] A Baddeley, E Rubak, and R Turner. Spatial Point Patterns: Methodology and Applications with R. CRC Press, 2015.

[7] TC Bailey and AC Gatrell. Interactive Spatial Data Analysis. Longman Scientific & Technical, 1995.

[8] RS Bivand, E Pebesma, and V Gomez-Rubio. Applied Spatial Data Analysis with R, Second edition. Springer, NY, 2013.

[9] MI Borrajo, W González-Manteiga, and MD Martínez-Miranda. Testing for significant differences between two spatial patterns using covariates. Spatial Statistics, 2019.

[10] D Brunet, ER Vrscay, and Z Wang. On the mathematical properties of the structural similarity index. IEEE Transactions on Image Processing, 21(4):1488–1499, 2012.

[11] C Charrier, K Knoblauch, LT Maloney, AC Bovik, and AK Moorthy. Optimizing multiscale SSIM for compression via MLDS. IEEE Transactions on Image Processing, 21(12):4682–4694, 2012.

[12] RG Congalton, RG Oderwald, and RA Mead. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. Photogrammetric Engineering and Remote Sensing, 49(12):1671–1678, 1983.

[13] NAC Cressie. Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics, 1993.

[14] TM Davies and ML Hazelton. Adaptive Kernel estimation of spatial relative risk. Statistics in Medicine, 29(23):2423–2437, 2010.

[15] P Diggle. A kernel method for smoothing point process data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 34(2):138–147, 1985.

[16] PJ Diggle, N Lange, and FM Benes. Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. Journal of the American Statistical Association, 86(415), 1991.

[17] T Duong, B Goud, and Kristine K Schauer. Closed-form density-based framework for automatic detection of cellular morphology changes. Proceedings of the National Academy of Sciences, 109(22):8382–8387, 2012.

[18] F Fouedjio. A hierarchical clustering method for multivariate geostatistical data. Spatial Statistics, 18:333–351, 2016.

[19] I Fuentes-Santos, W González-Manteiga, and J Mateu. Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. Scandinavian Journal of Statistics, 43(2):416–435, 2016.

[20] I Fuentes-Santos, W González-Manteiga, and J Mateu. A nonparametric test for the comparison of first-order structures of spatial point processes. Spatial Statistics, 22:240–260, 2017.

[21] AC Gatrell, TC Bailey, PJ Diggle, and BS Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. Royal Geographical Society, 1996.

[22] S Getzin, T Wiegand, K Wiegand, and F He. Heterogeneity influences spatial patterns and demographics in forest stands. Journal of Ecology, 96(4):807–820, 2008.

[23] PT Gilruth, SE Marsh, and R Itami. A dynamic spatial model of shifting cultivation in the highlands of Guinea, West Africa. Ecological Modelling, 79(1-3):179–197, 1995.

[24] A Gretton, KM Borgwardt, MJ Rasch, B Schölkopf, and A Smola. A kernel two-sample test. Journal of Machine Learning Research, 13(Mar):723–773, 2012.

[25] U Hahn. A studentized permutation test for the comparison of spatial point patterns. Journal of the American Statistical Association, 107(498):754–764, 2012.

[26] PR Halmos. Naive Set Theory. Courier Dover Publications, 2017.

[27] T Hastie, R Tibshirani, and J Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2009.

[28] R Kirsten, IN Fabris-Rotelli, and G Breetzke. Re-examining the similarity threshold of Andresen's S-index. Draft paper for journal submission, 2020.

[29] R Kirsten, IN Fabris-Rotelli, and C Kraamwinkel. On the similarity of spatial point patterns. Honours research, University of Pretoria, 2018.

[30] DG Krige. A statistical approach to some mine variations and allied problems at the witwatersrand. Master's thesis, University of Witwatersrand, 1951.

[31] AM Kulkarn and RC Joshi. Content-based image retrieval by spatial similarity. Defence Science Journal, 52(3):285, 2002.

[32] J Li and AD Heap. A review of spatial interpolation methods for environmental scientists. 2008.

[33] P Moraga. Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny. CRC Press, 2019.

[34] TP Morris, IR White, and MJ Crowther. Using simulation studies to evaluate statistical methods. Statistics in medicine, 38(11):2074–2102, 2019.

[35] K Muandet, K Fukumizu, B Sriperumbudur, and B Schölkopf. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning, 10(1-2):1–141, 2017.

[36] EA Nadaraya. On estimating regression. Theory of Probability & Its Applications, 9(1):141–142, 1964.

[37] EJ Pebesma and RS Bivand. Classes and methods for spatial data in R. R News, 5(2):9–13, November 2005.

[38] TD Pham. Geoentropy: A measure of complexity and similarity. Pattern Recognition, 43(3):887–896, 2010.

[39] JH Ratcliffe. Geocoding crime and a first estimate of a minimum acceptable hit rate. International Journal of Geographical Information Science, 18(1):61–72, 2004.

[40] SR Sain and N Cressie. A spatial model for multivariate lattice data. Journal of Econometrics, 140(1):226–259, 2007.

[41] C Stasch, S Scheider, E Pebesma, and W Kuhn. Meaningful spatial prediction and aggregation. Environmental Modelling & Software, 51:149–165, 2014.

[42] MNM Van Lieshout and AJ Baddeley. Indices of dependence between types in multivariate point patterns. Scandinavian Journal of Statistics, 26(4):511–532, 1999.

[43] F Vasefi, N MacKinnon, and DL Farkas. Hyperspectral and multispectral imaging in dermatology. In Imaging in Dermatology, pages 187–201. Elsevier, 2016.

[44] Z Wang and AC Bovik. A universal image quality index. IEEE Signal Processing Letters, 9(3):81–84, 2002.

[45] Z Wang, AC Bovik, HR Sheikh, and EP Simoncell. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.

[46] H Wässle, BB Boycott, and R-B Illing. Morphology and mosaic of on-and off-beta cells in the cat retina and some functional considerations. Proceedings of the Royal Society of London. Series B. Biological Sciences, 212(1187):177–195, 1981.

[47] AP Wheeler, W Steenbeek, and MA Andresen. Testing for similarity in area-based spatial point patterns: Alternative methods to Andresen's spatial point pattern test. Transactions in GIS, 22(3):760–774, 2018.