South African Statistical Journal
Vol. 55, No. 1, 55–71
https://doi.org/10.37920/sasj.2021.55.1.5

A GENERIC TEST FOR THE SIMILARITY OF SPATIAL DATA

R. Kirsten
Department of Statistics, University of Pretoria

I. N. Fabris-Rotelli[1]
Department of Statistics, University of Pretoria
e-mail: *inger.fabris-rotelli@up.ac.za*

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 proposes a generic spatial 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. A simulation study is done in order to test the method for each spatial data set. After the simulation study, it was concluded that the proposed spatial similarity test is not sensitive to the user-defined resolution of the pixel image representation. From the simulation study, the proposed spatial similarity test performs well on lattice data, some of the unmarked point patterns and the marked point patterns with discrete marks. We illustrate this test on property prices in the City of Cape Town and the City of Johannesburg, South Africa.

Key words: Generic, Similarity, S-index, Spatial similarity test, SSIM.

1. Introduction

Spatial data can take on three main forms, namely geostatistical data, lattice data and point patterns (Cressie, 1993). Geostatistical data is 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 (Cressie, 1993). Lattice data is when the observational region is divided into predefined subregions (either a regular grid or an irregular grid) (Sain and Cressie, 2007). The spatial data can be observed at the individual subregions and can either be continuous or discrete. Spatial point patterns consist of the locations of certain events (Baddeley et al., 2015). In the case where only the locations of one event type is present, we call it an unmarked point pattern. Extra information can be presented within the point pattern by associating a value (mark) to each point. This is then called a marked point pattern. This mark can either be discrete or continuous. Spatial data thus takes many forms.

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. Tests for the spatial similarity focus on whether the two spatial data sets originate from the same stochastic process in terms of their spatial

[1] Corresponding author.
MSC2020 subject classifications. 62G10, 62M40.

structure (Borrajo et al., 2019). The currently available tests only cover the more popular spatial data which is images (Brunet et al., 2012; Congalton et al., 1983; Gilruth et al., 1995; Kulkarn and Joshi, 2002; Wang and Bovik, 2002; Wang et al., 2004), unmarked spatial point patterns (Andresen, 2009; Bailey and Gatrell, 1995; Borrajo et al., 2019; Diggle, 1985; Diggle et al., 1991; Fuentes-Santos et al., 2017; Hahn, 2012; Duong et al., 2012; Alba-Fernández et al., 2016; Wheeler et al., 2018) and geostatistical data (Fouedjio, 2016; Pham, 2010) but in different manners.

In this paper we propose a test for spatial similarity that is generalised 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. 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 as either the mean or the proportion of the values in the local similarity map.

The methodology of the proposed spatial similarity test is discussed in Section 2. The method is tested on a simulation study in Section 3 with an application of the method on property prices in Section 4. We conclude in Section 5.

2. Methodology

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. The resolution (that is, the number of pixels) of the pixel image should be decided by the user before-hand. 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 test starts off by creating a pixel image representation of X_1 and X_2. The pixel image representation is obtained differently for each spatial data type. The spatial domain should be divided into an $m \times m$ grid as in Figure 1. 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 $\boldsymbol{u} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u}_M)$. 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 and is excluded in the calculations.

Figure 2 is a diagram showing the logic of the proposed spatial similarity test. In the rest of this section, we explain the methods behind the proposed spatial similarity test as it is outlined in the diagram.

2.1 Step 1: Create a pixel image representation

2.1.1 Point patterns

When dealing with point patterns, a kernel density estimation is used to obtain the pixel image representation. For unmarked point patterns, Diggle's corrected density estimate is used for the

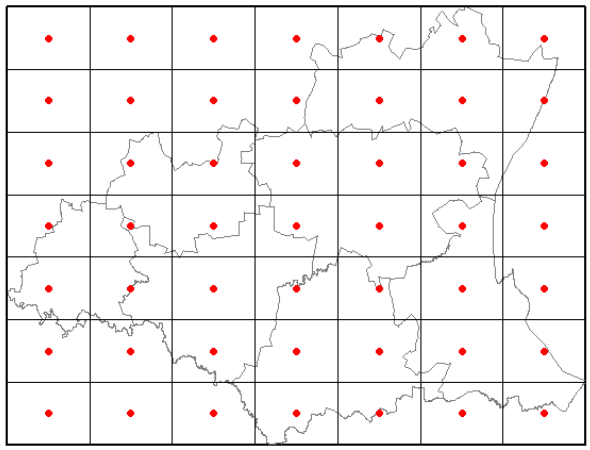


Figure 1. Illustration of how the spatial domain for lattice data is divided into pixels with $m = 7$. The dots represent the $\boldsymbol{u_j}$.

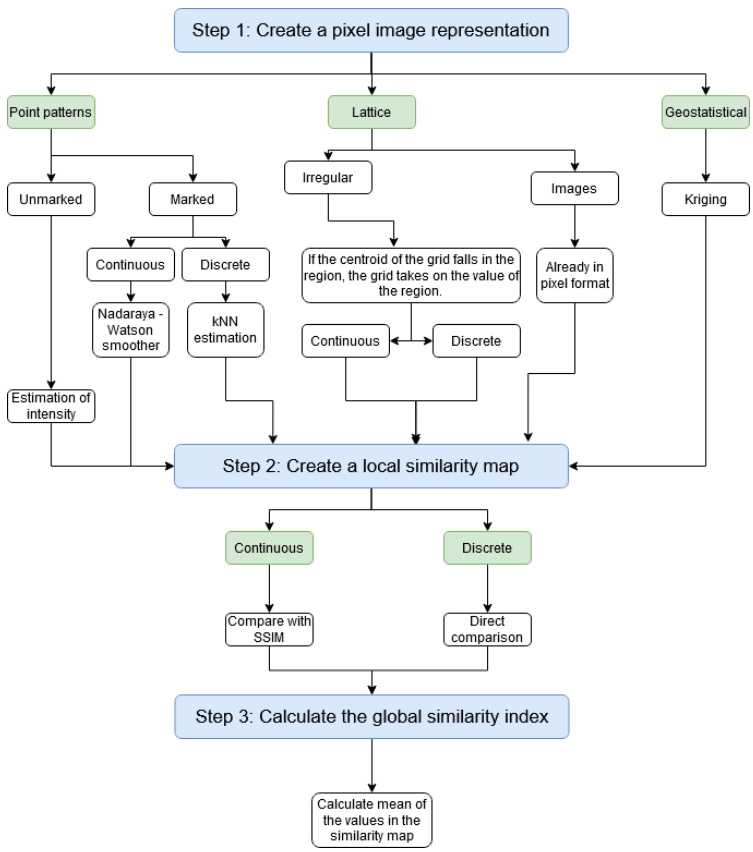


Figure 2. Diagram explaining the logic of the proposed spatial similarity test.

calculation as it results in a lower mean squared error compared to similar estimators (Baddeley et al., 2015)

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

where the kernel $\kappa(\cdot)$ is taken to be a bivariate Gaussian density $f(\boldsymbol{d}) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{d}\Sigma^{-1}\boldsymbol{d}'\right\}$ with Σ = bandwidth $\times I_2$. The bandwidth is the standard deviation of the kernel and can be seen as the smoothing parameter of the kernel. There are different methods of calculating the bandwidth. A popular method is Diggle's bandwith (Baddeley et al., 2015). Although the calculation of Diggle's bandwidth assumes a Cox process, this is the bandwidth used for the purpose of this paper 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 (Baddeley et al., 2015). The edge correction factor from (1) is

$$e(\boldsymbol{x_i}) = \int_D \kappa(\boldsymbol{x_i} - \boldsymbol{v_k}) d\boldsymbol{v_k}, \tag{2}$$

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 from 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 are used as the spatial locations $\boldsymbol{v} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_Q})$. Then, the calculation of (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

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

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

However, not all spatial point patterns will be 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 before, 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 (Baddeley et al., 2015)

$$\tilde{m}^D(\boldsymbol{u_j}) = \frac{\sum_{i=1}^{n} m_i \kappa(\boldsymbol{u_j} - \boldsymbol{x_i})/e(\boldsymbol{x_i})}{\sum_{i=1}^{n} \kappa(\boldsymbol{u_j} - \boldsymbol{x_i})/e(\boldsymbol{x_i})}, \tag{3}$$

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 same edge-effect factor from Equation (2).

With a marked spatial point pattern that has discrete marks, the Nadaraya–Watson smoother in Equation (3) will not be 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 will involve a 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 locations $\boldsymbol{u_j}$ (Hastie et al., 2009). In the case of the use of Euclidean distance, we use

$$d_{ij} = \|\boldsymbol{x_i} - \boldsymbol{u_j}\|.$$

The k nearest spatial data points are considered at each spatial location. Each $\boldsymbol{u_j}$ takes on the modal value of the k nearest spatial data points considered. The choice of k is arbitrary and up to the user. Care should be taken that the value for k should be strictly less than the number of spatial data points. Some work has been done in literature on guidelines on how to choose a value of k using misclassification error curves (Hall et al., 2008). As we are working with spatial data, a data-driven k would be advised. For the purpose of this paper, the value of k will be chosen as 10% of the number of points in the pattern.

2.1.2 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.

2.1.3 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 (Cressie, 1993). With geostatistical data in general, we are interested in creating a continuous map throughout the entire spatial domain, which is obtained with a method called Kriging by predicting the unobserved values (Cressie, 1993).

For the pixel image representation of a geostatistical data set, we can divide the spatial domain into pixels. We then Krige at each $\boldsymbol{u_j}$.

2.2 Step 2: Create a local similarity map

In the second step of our proposed similarity test, we obtain a similarity map between Y_1 and Y_2. In the case of continuous pixel values, the SSIM algorithm is used (Wang and Bovik, 2002). This algorithm was first developed as a quality index for images and later on used as a similarity index between images (Wang and Bovik, 2002; Wang et al., 2004). It 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.

The SSIM index can be calculated as

$$SSIM(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}}) = \left[\ell(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}})\right]^{\alpha} \left[c(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}})\right]^{\beta} \left[s(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}})\right]^{\gamma},$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ and $\boldsymbol{y_{ij}}$ are the values contained in sliding window j of data set i. Usually in literature, $\alpha = \beta = \gamma = 1$, which assigns an equal importance to each term (Wang et al., 2004). Equation (6) is calculated for each sliding window j.

The components are calculated separately as follows and then multiplied together for the SSIM

value:

$$\begin{aligned}
\text{Luminance:} \quad & \ell(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}}) = \frac{2\mu_{y_{1j}}\mu_{y_{2j}} + C_1}{\mu_{y_{1j}}^2\mu_{y_{2j}}^2 + C_1}. \\
\text{Contrast:} \quad & c(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}}) = \frac{2\sigma_{y_{1j}}\sigma_{y_{2j}} + C_2}{\sigma_{y_{1j}}^2\sigma_{y_{2j}}^2 + C_2}. \\
\text{Structure:} \quad & s(\boldsymbol{y_{1j}}, \boldsymbol{y_{2j}}) = \frac{2\sigma_{y_{1j},y_{2j}} + C_3}{\sigma_{y_{1j}}\sigma_{y_{2j}} + C_3}.
\end{aligned}$$

The last values needed to calculate the Luminance, Contrast and Structure terms are the constants. The constants are used in order to avoid inconsistency (Wang et al., 2004). The constants are $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$ and $C_3 = C_2/2$ where we will choose $K_1 = 0.01$ and $K_2 = 0.03$ (Wang et al., 2004). 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.

In the case of discrete, especially categorical, pixel values, the SSIM will not be 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 0.

2.3 Step 3: Calculate global similarity index

From the local similarity map in the second step, we calculate a global similarity index that is the 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 (Andresen, 2009):

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

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. $SSIM(u_j)$ can also be seen as 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. Simulation study

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 (Morris et al., 2019). 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 will be 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%.

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.

3.1 Geostatistical simulations

For the geostatistical simulations, a built-in `R` data set is used. This data set is contained within the `sp` package (Bivand et al., 2013; Pebesma and Bivand, 2005) and is called `meuse`. 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.

The two data sets to compare are obtained by taking the spatial locations and each of the measurements (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 are 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.

3.2 Lattice data

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. The values for each spatial location are 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 will be removed and replaced with other random uniform numbers within the same range.

3.3 Point patterns

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 were 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 (Baddeley et al., 2015) while the clustered spatial point patterns are simulated with the `rMatClust` function (Baddeley et al., 2015).

To add more variety to the simulation study, we use three approaches for the simulations. The

first approach creates noisy patterns. In this approach, the regular and clustered point patterns are simulated with the above functions. When we replace some of the data points to create the test pattern, the spatial data points are 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 create X_2.

3.3.1 Unmarked point patterns

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.

3.3.2 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 has 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. For the last group, they range from 0 to 100.

3.3.3 Discrete marked point patterns

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

3.4 Results

Figure 3 and Figure 4 represent the results from applying the proposed spatial similarity test to the simulated spatial data.

We test the hypothesis of equal means across the three groups for pixel image resolutions 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 1 shows the p-values of the Kruskal–Wallis test.

From Table 1, it can clearly be seen that the hypothesis of equal means cannot be rejected in all the cases except for the geostatistical simulations that are 90% identical. From doing a pairwise Wilcoxon test it can be concluded that the means for the 10×10 pixel image resolution differ from the other two means at a 5% level of significance, while the other two groups (20×20 and 50×50) do not differ significantly from one another at a 5% level of significance.

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

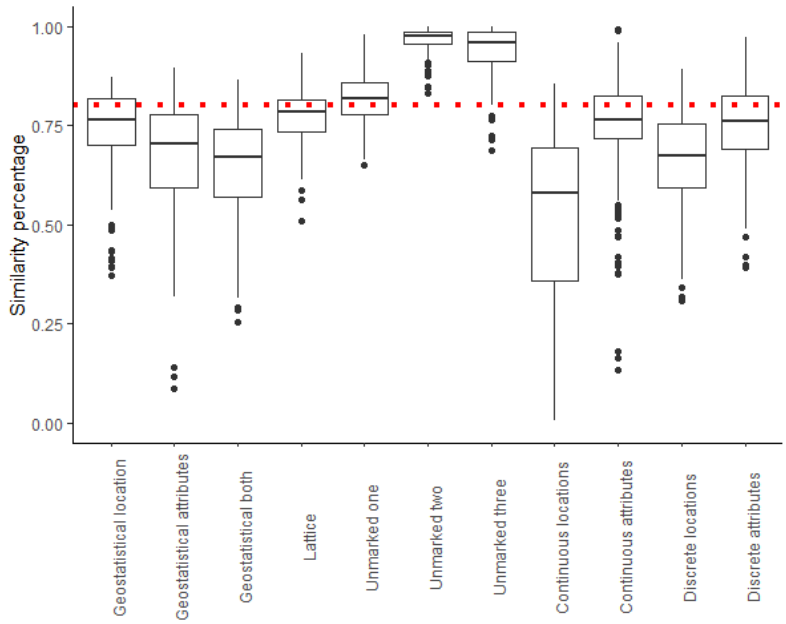


Figure 3. Results from applying the proposed spatial similarity test to the simulated spatial data being 80% identical. The dotted horizontal line indicates 0.8.

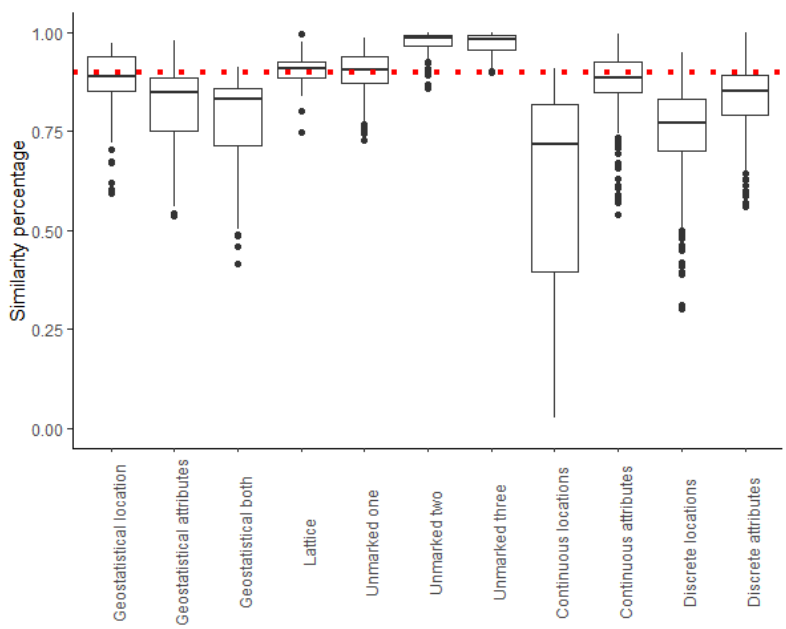


Figure 4. Results from applying the proposed spatial similarity test to the simulated spatial data being 90% identical. The dotted horizontal line indicates 0.9.

Table 1. p-values of the Kruskal–Wallis test to test the hypothesis of equal means across the three groups for the pixel image representations. This shows that the test is not sensitive to the user-defined resolution of the pixel image representation.

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

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.

3.5 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 and standard deviation. These values are given in Table 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.

When looking at the geostatistical summary statistics in Table 2, it can be seen that the means and medians of the results do tend to theoretical 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 are still large with the lowest of the standard deviations equal to 0.1281. For the geostatistical data where the locations are changed, the means are equal to 0.722 and 0.8202 while the medians are equal to 0.7645 and 0.8577. When changing the attributes, the means are 0.661 and 0.7523. The medians are then 0.7047 and 0.8144. When both the locations and attributes are changed, the means are 0.6302 and 0.7302 while the medians are 0.6691 and 0.774. The inaccuracy of this data type can be accounted to the method of Kriging that may be too general for the type of data used. A more optimal variogram model for the Kriging may yield better results (Li and Heap, 2008).

The proposed spatial similarity test seems to compare the similarity between two lattice data sets

Table 2. Summary statistics of the results from the proposed spatial similarity test.

	Identical	**Mean**	**Median**	**Standard deviation**
Geostatistical data				
Locations changed	80%	0.7220	0.7645	0.1310
	90%	0.8202	0.8577	0.1281
Attributes changed	80%	0.6610	0.7047	0.1712
	90%	0.7523	0.8143	0.1614
Both changed	80%	0.6302	0.6691	0.1495
	90%	0.7302	0.7740	0.1531
Lattice data	80%	0.7740	0.7846	0.0751
	90%	0.9043	0.9079	0.03941
Unmarked point patterns				
Method one	80%	0.8195	0.8166	0.0667
	90%	0.8992	0.9060	0.0536
Method two	80%	0.9654	0.9755	0.0329
	90%	0.9763	0.9847	0.0266
Method three	80%	0.9408	0.9599	0.0591
	90%	0.9732	0.9815	0.0252
Continuous marked				
Locations changed	80%	0.5066	0.5736	0.2344
	90%	0.6057	0.7178	0.2596
Attributes changed	80%	0.7571	0.7656	0.1074
	90%	0.8771	0.8860	0.0760
Discrete marked				
Locations changed	80%	0.6624	0.675	0.1187
	90%	0.7514	0.77	0.1104
Attributes changed	80%	0.7550	0.76	0.0942
	90%	0.8399	0.85	0.0785

quite accurately with the means (0.774 and 0.904) and medians (0.7846 and 0.9079) of the results close to the theoretical values. The standard deviations (0.0751 and 0.0394) are also small which is also an indication that this method performs well in the case of lattice data.

For the unmarked point patterns, the method does perform well on the simulations from the first method. This can be said since the means (0.8195 and 0.8992) and the medians (0.8166 and 0.906) of the results are close to the theoretical values and the standard deviations (0.0667 and 0.0536) 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 where the means are equal to 0.965 and 0.976 and the medians 0.9755 and 0.9847. 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 was 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 were removed to obtain the test set. The reason the test may yield such high similarity values may be in the way in which the pixel image representations are obtained. The means of the results from the third method of simulations are 0.9409 and 0.9732 while the medians are 0.9598 and 0.9815.

This proposed spatial similarity test can still be improved to perform better on marked spatial point patterns with continuous values. In the case where the locations of the points are changed, this method does not perform well. The means of the results when the locations are changed are equal to 0.5066 and 0.6057 while the medians are equal to 0.5736 and 0.7178. This may be again due to the way in which the pixel image representations were obtained. When the attributes of some of the points are changed, this method performs better again in terms of the means (0.7571 and 0.8771) and medians (0.7656 and 0.886) of the results. However, in reality it will happen more often that we have a scenario in which the attributes are changed rather than the locations.

In the case of marked spatial point patterns with discrete values, the method performs better when the attributes are changed (with means equal to 0.755 and 0.8399 and medians equal to 0.76 and 0.85) than when we change the locations (with means equal to 0.6624 and 0.7514 and medians equal to 0.675 and 0.77).

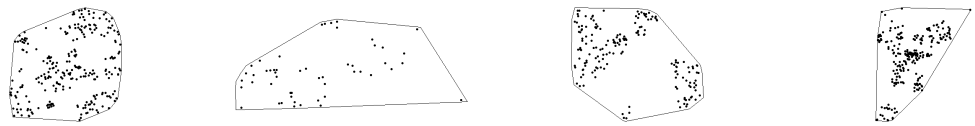
The size of the data sets to be compared should be intuitively dealt with to some extent. Data sets differing significantly in size would not be considered for comparison in practice. The methodology presented here, however, allows for comparison of data sets of different size by aggregating at the pixel image representation step as well as standardising in the kernel density step.

4. Application

A data set provided by Lightstone[2,3] 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 of 2017, 2018 and 2019. We apply the proposed spatial similarity test to each block within the two metros.

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

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



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

Figure 5. The four separate blocks of property locations that will be 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.

Figure 5(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(a) and (b) are the two blocks in the City of Johannesburg metro and Figure 5(c) and (d) are the two blocks in the City of Cape Town metro.

Figure 6(a)–(d) show density plots of the property prices over the different years within the four blocks of properties. Figure 6(a) is the density plot of the property prices in the first block of properties in the City of Johannesburg metro and Figure 6(b) is the second block of properties. These blocks consist of 423 and 120 properties, respectively. Figure 6(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 6(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. 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 of each of the four blocks, we create a local similarity map in the second step of the proposed spatial similarity test; see Figure 7. Figure 7 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 paper, the pixel image representations used have a resolution of 30×30. The bandwidth used is Diggle's bandwidth (Baddeley et al., 2015). The windows of the patterns are the same as displayed in Figure 5(a)–(d) which are 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 Brunet et al. (2012); Wang et al. (2004).

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). Low similarity is observed for the three comparisons of the first block of properties in the City of Cape Town metro.

The global similarity indices can be seen in Table 3. These values support the observations from the similarity maps in Figure 7. It also indicates that something significant happened with the property prices from the first block of the City of Cape Town metro.

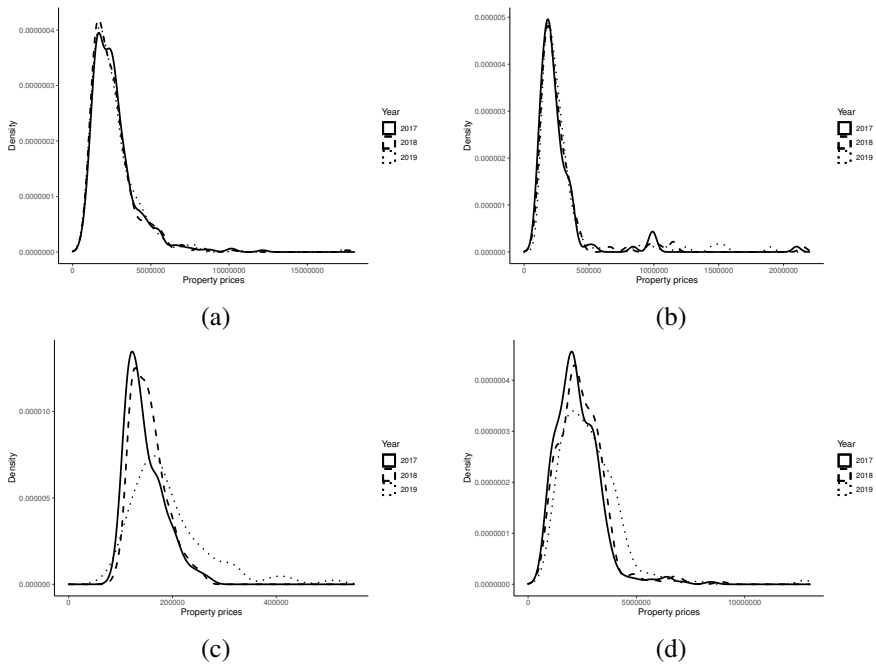


Figure 6. Density plots for the property prices for the different years within the four blocks of properties. (a) The property prices for the first block in Johannesburg and (b) the second block. (c) The property prices in the first block of properties in Cape Town and (d) the second block.

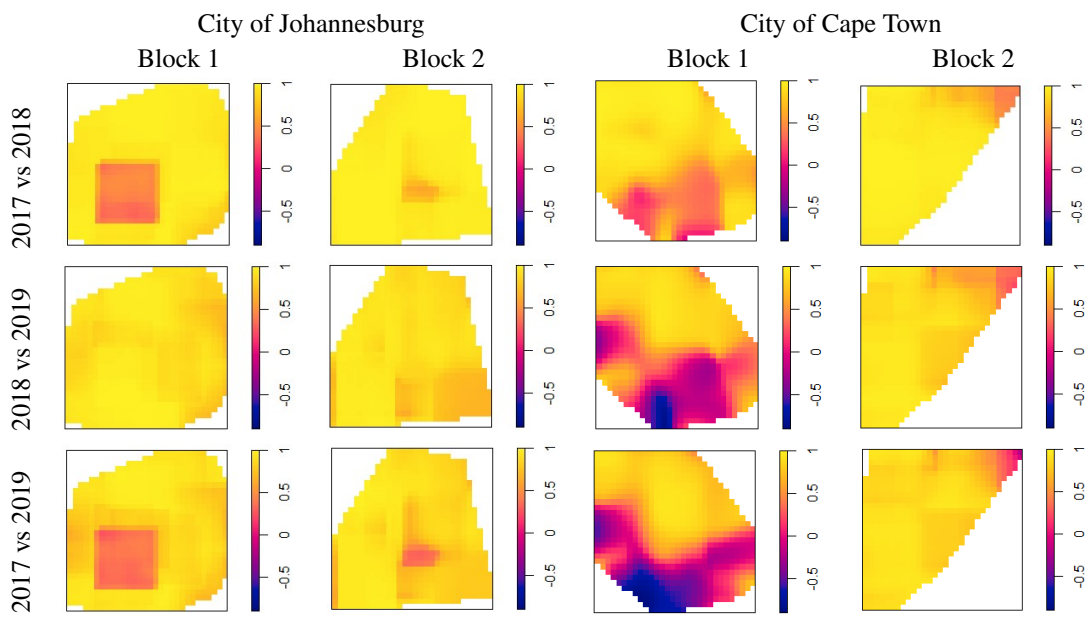


Figure 7. Local similarity maps for each comparison done on the four blocks of property prices by year.

Table 3. Similarity indices from the newly proposed similarity test

	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

5. Conclusion

Up to now in literature, only a few spatial similarity tests have been developed. These tests test the similarity between two spatial data sets for only a certain type of data. In this paper, a generic spatial similarity test is proposed. This test can determine the spatial similarity between two spatial data sets of any type.

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 done 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 more similar and where they differ. The final similarity measure is calculated in the third step of the test by using the values from the local similarity map. The final result of this test should be interpreted as a percentage of similarity between the two spatial data sets.

A simulation study was done to test the accuracy of the proposed test. For a future study, a larger simulation study would be the suggestion. 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.

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 (Li and Heap, 2008). 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 is to optimise the bandwidth selection (Baddeley et al., 2015). 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. Guidelines can also be put in place on how to choose a data-driven value of k. For example, the analysis of the strength of the spatial dependency distance can also be used to choose a data-driven value of k.

It is also possible to vary the α, β and γ parameters within the SSIM calculation (Charrier et al., 2012). 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 Section 5. We considered four blocks of properties with prices over three years. We use the test to compare the property prices over different years within the same block of properties. The price drop for the properties in the first

block in the City of Cape Town metro could possibly be explained by other economic factors. To further explain such differences, some economic factors can be considered for the time period.

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

References

Alba-Fernández, M., Ariza-López, F., Jiménez-Gamero, M., and Rodríguez-Avi, J. (2016). On the similarity analysis of spatial patterns. *Spatial Statistics*, **18**, 352–362.

Andresen, M. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography*, **29**, 333–345.

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

Bailey, T. and Gatrell, A. (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical, London.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. 2nd edition. Springer, New York, NY.

Borrajo, M., González-Manteiga, W., and Martínez-Miranda, M. (2019). Testing for significant differences between two spatial patterns using covariates. *Spatial Statistics*, **40**, 1–20.

Brunet, D., Vrscay, E., and Wang, Z. (2012). On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, **21**, 1488–1499.

Charrier, C., Knoblauch, K., Maloney, L., Bovik, A., and Moorthy, A. (2012). Optimizing multiscale SSIM for compression via MLDS. *IEEE Transactions on Image Processing*, **21**, 4682–4694.

Congalton, R., Oderwald, R., and Mead, R. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, **49**, 1671–1678.

Cressie, N. (1993). *Statistics for Spatial Data*. Revised edition. Wiley & Sons, Hobokem, NJ.

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

Diggle, P. J., Lange, N., and Benes, F. M. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association*, **86**, 618–625.

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

Fouedjio, F. (2016). A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, **18**, 333–351.

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

Gilruth, P., Marsh, S., and Itami, R. (1995). A dynamic spatial model of shifting cultivation in

the highlands of Guinea, West Africa. *Ecological Modelling*, **79**, 179–197.

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

Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Annals of Statistics*, **36**, 2135–2152.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, New York, NY.

Kulkarn, A. and Joshi, R. (2002). Content-based image retrieval by spatial similarity. *Defence Science Journal*, **52**, 285.

Li, J. and Heap, A. D. (2008). *A Review of Spatial Interpolation Methods for Environmental Scientists*. Geoscience Australia, Canberra.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, **5**, 9–13.
URL: *https://CRAN.R-project.org/doc/Rnews/*

Pham, T. D. (2010). Geoentropy: A measure of complexity and similarity. *Pattern Recognition*, **43**, 887–896.

Sain, S. and Cressie, N. (2007). A spatial model for multivariate lattice data. *Journal of Econometrics*, **140**, 226–259.

Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, **9**, 81–84.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncell, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**, 600–612.

Wheeler, A., Steenbeek, W., and Andresen, M. (2018). Testing for similarity in area-based spatial point patterns: Alternative methods to Andresen's spatial point pattern test. *Transactions in GIS*, **22**, 760–774.