entropy

MDPI

Article

A New Technique Based on Voronoi Tessellation to Assess the Space-Dependence of Categorical Variables

Pedro J. Zufiria * and **Miguel Á. Hernández-Medina**

ETS Ingenieros de Telecomunicación, Information Processing and Telecommunications Center (IPTC),
Universidad Politécnica de Madrid, 28040 Madrid, Spain

* Correspondence: pedro.zufiria@upm.es; Tel.: +34-9106-722-86

Received: 2 May 2019; Accepted: 2 August 2019; Published: 8 August 2019

Abstract: Based on a sample of geolocated elements, each of them labeled with a (not necessarily ordered) categorical feature, several indexes for assessing the relationship between the geolocation variables (latitude and longitude) and the categorical variable are evaluated. Among these indexes, a new one based on a Voronoi tessellation presents several advantages since it does not require a variable transformation or a previous discretization; in addition, simulations show that this index is considerably robust when compared with the previously known ones. Finally, the use of the presented indexes is also illustrated by analyzing the geolocation of communities in some communication networks derived from Call Detail Records.

Keywords: spatial correlation; independence indices; Voronoi tessellation; entropy

1. Introduction

The statistical analysis of space-related information is a long-developed research field with applications in biology [1], geology [2,3], sociology [4], etc., where different measures have been defined for assessing different features such as spatial dispersion [5,6], spatial autocorrelation [7,8], or spatial homogeneity [9]. Presently, geolocated information is becoming widely available from many sources of data such as Call Detail Records (CDRs) of telephone operators [10,11], vehicle geolocation systems [12,13], Internet of Things (IOT) architectures [14,15] or population surveys [16]. In general, these sets of geolocated data gather sample vectors which contain at least two continuous variables determining the latitude and longitude of the actor, and some other categorical variables which may take values belonging to a (not necessarily ordered) finite set of labels. In addition, new scenarios where these types of data are encountered are becoming common when modelling social networks. In this context, for instance, secondary variables are frequently computed for labeling network communities [17], whose relationships with geolocation have been recently analyzed for different countries at global and city scopes [18].

The above-mentioned scenarios encourage the study and development of new tools for the statistical analysis of random vectors containing different types of component variables (continuous, discrete, or categorical). Most of the common techniques to assess the relationship between random variables assume that such variables are of the same type: for example, classical Pearson's and G-test [19,20] can be used to assess the independence between discrete variables, whereas the relationship between continuous variables has been tested using binning techniques [21], mutual information estimators [22–25], kernel-based methods [26,27], correlation distance estimators [28] or detectors based on the analysis of subsequences [29].

On the other hand, the relationship between heterogeneous variables can be evaluated, for instance, if a previous binning step is performed on the continuous variables in order to convert all of them into a discrete format; then, some of the above-mentioned procedures can be applied. In addition, note that ANOVA or MANOVA-type tests can also be employed when the dependent variables are continuous and the independent ones are categorical; unfortunately, these tests rely on very strong assumptions on the distribution of the continuous involved variables (i.e., they have to be jointly Gaussian). Recently, some tools have been developed which directly estimate the mutual information between discrete and continuous variables [30]. Most of these techniques perform well for specific settings, whereas their behavior assessment in other frameworks such as the one addressed in this paper remains a challenging problem.

Grounded on the above motivations and challenges, some algorithms to evaluate the relationship between the geolocation variables and a categorical variable are presented in this paper: we elaborate a detailed characterization and a comparative analysis of these algorithms for evaluating such relationship, making special emphasis on a new scheme based on a Voronoi tessellation (a preliminary version of this scheme was just proposed in [31] without a rigorous assessment of its capabilities). In addition, this paper illustrates the application of these algorithms to assess the geographical distribution of the communities detected in communications networks derived from CDRs. Precisely, they will allow us to shed some light on a hypothesis proposed in [18] about the distribution of communities being unrelated to geolocation within three different European cities.

The paper is organized as follows: the problem statement is formalized in Section 2, and the different existing alternatives for testing independence are presented in Section 3, including the index based on a Voronoi tessellation. All the presented alternative indexes are computationally evaluated on different scenarios in Section 4. Finally, a discussion of the results is presented in Section 5 and concluding remarks are outlined in Section 6.

2. Formal Problem Statement

The available data can be formally represented by a set of measurements $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, such that each $\mathbf{x}_i = (x_{i1}, x_{i2}) \in \mathbb{R}^2$ gathers geolocation information of individual i, and each $y_i \in C = \{c_1, \ldots, c_K\}$ is a sample from a (not necessarily ordered) categorical random variable which represents some property or feature associated with such individual i. Accordingly, we can define $\mathbf{X} = (X_1, X_2) \in \mathbb{R}^2$ to be the vector random variable determining geolocation, and $Y \in C$ the categorical random variable which represents the mentioned property or feature. Then, each $(\mathbf{x}_i, y_i)$, which reflects an individual satisfying property y_i at location $\mathbf{x}_i$, can be seen as a sample of the joint random variables $(\mathbf{X}, Y)$.

In general, together with the sample set $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, N$, the proximity between any two elements $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^2$ must be defined. Usually, this is formalized via a matrix W of weights that may generalize the usual notion of the inverse of the distance $d(\mathbf{x}_i, \mathbf{x}_j)$:

$$w_{ij} = \frac{1}{d(\mathbf{x}_i, \mathbf{x}_j)}, \ \forall i, j \in \{1, \ldots, N\},\ i \neq j, \tag{1}$$

$$w_{ii} = 0,\ \forall i = 1, \ldots, N. \tag{2}$$

Please note that although we are focusing on geolocated data (i.e., in $\mathbb{R}^2$) the reasoning and the algorithms can be extended to cases where $\mathbf{x}_i \in \mathcal{X}$ being $(\mathcal{X}, d)$ a metric space with a standard distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$.

Our main objective is to evaluate, based on the available sample, the possible relationship (i.e., dependence) between the geolocation vector variable $\mathbf{X}$ and the categorical variable Y.

In the following Section we present several tools aimed to assess such relationship.

3. Assessing the Relationship between $\mathbf{X}$ and Y

The most natural procedure to assess the relationship between $\mathbf{X}$ and Y is to construct a test for independence between such variables. As advanced in Section 1, usual tests for checking the independence between random variables are defined for either categorical variables [19,20] or continuous random variables or vectors [21–24,26–29]. As mentioned above, ANOVA or MANOVA-type tests [32,33] are available for cases where the dependent variables are continuous and the independent ones are categorical, but they require the distribution of the continuous variables to be jointly Gaussian (which is not usually the case for the geolocation variables).

Ripley's K index [9], as a measure of dispersion homogeneity can be employed by analyzing the dispersion relationship between pairs of categories. This could provide an indirect way to assess independence which is not straightforward since it would involve the analysis of all possible pairs of categorical values. An extension of such index [34], based on multiple co-occurrences, could also been adapted to provide an indirect way to test for independence.

Tests of independence between a dependent categorical variable and independent continuous variables are well known for the two categories (i.e., binary) case. Nevertheless, no tests seem to be established when we have dependent categorical variables taking more than two values and continuous independent vector variables, which is the case considered in this paper.

There are two simple indirect ways to address this problem which go through transforming some variable into a new type one for allowing the application of some known test scheme. The first approach converts Y into a numerical variable by just assigning a different real number to each possible category. From there on, classical correlation schemes can be applied as shown below. Please note that this procedure may be quite sensitive to the arbitrary number assignment to each class. Alternatively, the second approach converts the continuous independent variables into categorical ones via a binning procedure, and then applies existing tests between categorical variables. Note also that this binning procedure may lose much information, especially when the independent variable $\mathbf{X}$ is a vector. A procedure based on a k-nearest neighbors analysis to estimate the mutual information has been also proposed for the scalar X case [30], but its extension to the case when $\mathbf{X}$ is a vector has not been evaluated.

In the following subsection we illustrate the first approach of transforming Y into a real variable, and then we show some known indexes that estimate spatial autocorrelation for real labeled data in $\mathbf{X} \in \mathbb{R}^2$.

3.1. Transforming Y into a Real Variable Z. Spatial Autocorrelation

Let $y \in C = \{c_1, \dots c_K\}$ so that Y can only take one of these K categorical values. Then, random variable Y can be transformed into a real valued random variable Z by defining an injective function

$$\begin{aligned} z : C &\longrightarrow \mathbb{R} \\ c_i &\longrightarrow z_i = z(c_i),\ i = 1, \dots, K \end{aligned}$$

which assigns a real value $z_i \in \mathbb{R}$ to each categorical value $c_i \in C$.

Then, different indexes can be computed for the resulting $Z(Y)$ variable with respect to $\mathbf{X}$. We present now two well-known autocorrelation measures.

Spatial Autocorrelation

The estimation of spatial autocorrelation was addressed in [35] based on the work of [7,8]. Since then, many improvements have been proposed [36], where always $z \in \mathbb{R}$, meaning that it is quantified via a numerical value. Many measurements of spatial correlation that can be defined; among them, the most common are the following ones (where we name $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ and $W = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$).

Moran's I

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{N}(z_i - \bar{z})^2} \tag{3}$$

Geary's C

$$C = \frac{N-1}{2W} \cdot \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{N}(z_i - \bar{z})^2} \tag{4}$$

Please note that these indexes provide an estimator of space autocorrelation; hence, we can formulate a test for the existence of such correlation by estimating the p-value corresponding to such indexes via a randomization procedure based on a random shuffling of the z_i values while preserving the proximity matrix values w_{ij}. This randomization procedure will be generally performed to evaluate the relevance of the index values provided by all the algorithms proposed in this paper.

Coming back to assessing the relationship between $\mathbf{X}$ and Y, two problems arise. First, the resulting index values I and C may strongly depend on the selected assignment function z. Hence, some computationally demanding schemes may be required for checking different assignment functions z to search for the maximum index value attainable with this approach. The second limitation of these procedures is that they check for correlation, not for independence, so that they are quite sensitive to distances between points rather than focusing on their relative geolocations.

We now illustrate the above-mentioned second indirect approach to evaluate the relationship between $\mathbf{X}$ and Y.

3.2. Quantizing **X**. *Mutual Information Based Index*

We now illustrate the alternative procedure where the continuous variable $\mathbf{X} \in \mathbb{R}^2$ can be binned or quantized. Let us consider a partition of the region of interest into a collection of disjoint subsets $A_l \subset \mathbb{R}^2$, $l = 1, \dots, L$, so that each subset A_l is assigned a label category $q_l \in Q = \{q_1, \dots, q_L\}$. This partition can be formulated via the function

$$\begin{aligned} q: \quad & \mathbb{R}^2 \longrightarrow Q \\ & (x_1, x_2) \longrightarrow q(x_1, x_2) = q_l, \ \forall (x_1, x_2) \in A_l \end{aligned}$$

which assigns label q_l to all points belonging to region $A_l \subset \mathbb{R}^2$. Then, standard independence tests between categorical variables $Q(\mathbf{X})$ and Y can be applied. In order to illustrate these procedures, a simple Mutual Information (MI)-based index can be computed as:

$$M = I(Q, Y) = \sum_{q_l \in Q} \sum_{c_k \in C} p(q_l, c_k) \log\left(\frac{p(q_l, c_k)}{p(q_l)p(c_k)}\right)$$

Again, this alternative index can be quite sensitive to the selected quantizing function q. Therefore, new alternative statistics which avoid the transformation or quantization of variables may be valuable for testing the independence between $\mathbf{X}$ and Y. In the following, we present a statistic proposed in [18] which applies directly to the original $\mathbf{X}$ and Y variables.

3.3. Herrera's Index

In [18] Herrera analyzed the geographical distribution of the communities detected in communications networks derived from CDRs for three different countries (France, Portugal, and Spain). The analysis was performed at both country and city scales (Paris, Lisbon and Madrid).

For assessing the relationship between geolocation and communities, a new D index was computed which employs the information gathered in the categories or classes (i.e., communities) $c_k \in C' \subset C$ with more than one element, i.e., such that $N_k = \#\{i \in \{i, \dots, N\} : y_i = c_k\} > 1$. If we define an ordering among the measurements of each class $c_k \in C'$ (e.g., the ordering induced by the labeling of the whole set of measurements), and denoting k_i the (absolute) index in the whole set for the i-th element of class k, Herrera's index calculates:

$$D = \frac{\sum_{c_k \in C'} \sum_{i=2}^{N_k} \sum_{j=1}^{i-1} d(\mathbf{x}_{k_i}, \mathbf{x}_{k_j})}{\sum_{c_k \in C'} \sum_{i=2}^{N_k} (i-1)} = \frac{\sum_{c_k \in C'} \sum_{i=2}^{N_k} \sum_{j=1}^{i-1} d(\mathbf{x}_{k_i}, \mathbf{x}_{k_j})}{\sum_{c_k \in C'} \binom{N_k}{2}} \tag{5}$$

Please note that this index adds up distances between pairs of elements belonging to the same class; the denominator is just a normalizing factor so that the index provides an average distance between pairs of points in the same class.

To have a baseline reference, Herrera proposed in [18] to compute also this index for a random shuffling of labels (preserving the N_k values) on the same $\mathbf{x}_i$ values; the ratio between this random-based index D_r and the one obtained in (5) was then provided as a final indicator (Please note that for obtaining the ratio $\frac{D_r}{D}$ it is not necessary to compute the (same) normalizing denominator (it would cancel when computing the quotient)). This final indicator value was interpreted in [18] by saying that if the ratio was clearly large than 1, it meant that $\mathbf{X}$ and Y were not independent.

Interestingly, Herrera's index is strongly related to a diversity index proposed in [37] where two types of distances are computed for each class. On the one hand, the average distance between the elements within each class k is computed an denoted as the intra-distance d_k^{int}; on the other hand, the average distance between elements of each class k and the elements of other classes is computed and denoted as the extra-distance d_k^{ext}:

$$d_k^{int} = \frac{\sum_{i=1}^{N_k} \sum_{j=1, j \neq i}^{N_k} d(\mathbf{x}_{k_i}, \mathbf{x}_{k_j})}{N_k(N_k - 1)}, \text{ for } N_k > 1; \quad d_k^{ext} = \frac{\sum_{i=1}^{N_k} \sum_{j=1}^{N-N_k} d(\mathbf{x}_{k_i}, \mathbf{x}_{k_j})}{N_k(N - N_k)}, \text{ for } N > N_k. \tag{6}$$

Please note that d_k^{int} values correspond to the terms added up in the numerator of index D, whereas d_k^{ext} values serve as a comparative reference. The comparison of both distances provides similar information as the one obtained by comparing intra-distances of the original data and the randomized data proposed in [18], in the sense that a data set with randomized label distribution should provide similar values for d_k^{int} and d_k^{ext}.

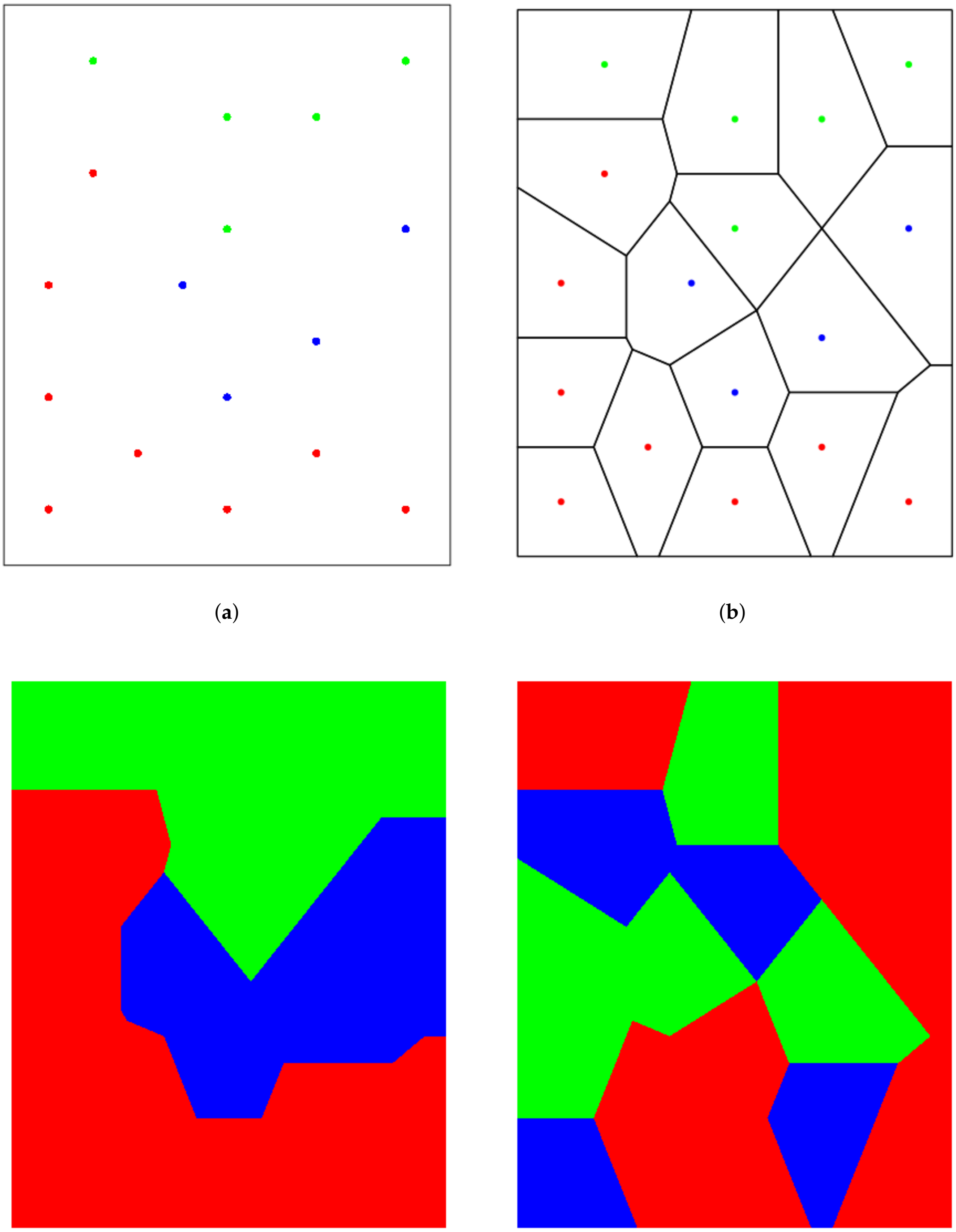
The analysis in [18] using index D concluded that when considering country/region scales, $\frac{D_r}{D}$ would take values from 3.5 to 4.4, suggesting a very strong space correlation among the different detected communities. On the other hand, when the analysis was performed at a city scale (e.g., Paris, Portugal and Madrid) the $\frac{D_r}{D}$ index would take values between 1.08 and 1.39, much lower than the one obtained for the country scale case; based on these numbers, independence between communities and geolocation was hypothesized.

In Section 4 this procedure will be formalized by estimating the p-value corresponding to D via an appropriate randomization procedure. Please note that the same p-value estimation procedure could be developed by using the distances proposed in [37]. The results in the example will show that there still exists a clear dependence between communities and geolocation even at the city scale.

In the following subsection we present another index which can also be directly applied to $\mathbf{X}$ and Y variables.

3.4. Voronoi Tessellation Based Index

This new index is based on the topological properties of the Voronoi tessellation [31,38] associated with the sample set $\{\mathbf{x}_1 \ldots, \mathbf{x}_N\}$ (with $\mathbf{x}_i \in \mathbb{R}^2$). This tessellation defines a partition of the space region under analysis into a collection of disjoint sub-regions (called cells) each one associated with a point $\mathbf{x}_i$ (see Figure 1a,b in the example explained below). We will denote cell V_i the one associated with $\mathbf{x}_i$. Two cells in the tessellation are called adjoining if they share a common side; note that Voronoi cells V_i and V_j corresponding to close points $\mathbf{x}_i$ and $\mathbf{x}_j$ are likely to be adjoining cells. Please note that V_i can be assigned the categorical value y_i corresponding to $\mathbf{x}_i$. Then, if adjoining cells with the same y_i value are assembled into a single piece, the number of pieces associated with each different y_i value in the tessellation gives information about the relationship between $\mathbf{X}$ and Y.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. (**a**) Distribution of labeled points. (**b**) Corresponding Voronoi tessellation. (**c**) Colored groups of Voronoi cells. (**d**) Colored groups after randomly shuffling color labels.

It is worth mentioning that sometimes we may get some Voronoi cells which happen to share a side outside the region under analysis, especially if they correspond to $\mathbf{x}_i$ values located close to the boundary of such region. In order to deal with such cases in a flexible way, an extension of the algorithm proposed in [31] has been developed in this paper by incorporating an adjustable margin in the region of analysis, so that we may allow Voronoi cells to become adjoining out of the initial boundaries of the original region of analysis.

Figure 1 illustrates the information provided by the Voronoi tessellation procedure in a simple example. Figure 1a displays a distribution of points where the label y can take three values represented by three colors (red, blue, and green). The corresponding Voronoi tessellation is shown in Figure 1b whereas the cells are colored according to their corresponding y value in Figure 1c. Please note that only one piece is obtained for each color due to the very strong relationship between $\mathbf{X}$ and Y. Finally, Figure 1d shows the result after shuffling the label values among the points: a larger number of pieces is obtained for independent $\mathbf{X}$ and Y.

To efficiently perform the computation of the number of pieces (groups of adjoining cells) associated with each label y_i, note that the Voronoi tessellation adjacency structure can be modelled via a graph G where each vertex represents a cell V_i (with label y_i), and two vertices are connected if their corresponding cells are adjoining. By selecting only those nodes with a given label value $y_i = c_k$ the corresponding subgraph G_k (composed only by the selected nodes and the corresponding links) gathers all the required information for determining the pieces or groups of cells. Precisely, for each class k we can easily determine both the corresponding number of pieces and the size (number of cells) of each piece by computing the connected components CC_k^l of the corresponding subgraph G_k (note that each CC_k^l, $l = 1, \dots, L_k$ is again a connected subgraph of G_k). We can denote $|CC_k^l|$ the number of nodes of each connected component CC_k^l. Please note that $\sum_{l=1}^{L_k} |CC_k^l| = N_k$, the number of nodes of subgraph G_k.

Once the number of pieces associated with each class are computed, an entropy measure is proposed to comparatively quantify the distribution of such number of pieces among all classes. In the same way as Herrera's index, the Voronoi index only gathers the information in classes $c_k \in C'' \subset C$ which have two or more elements, i.e., satisfying $N_k = \#\{i \in \{i, \dots, N\} : y_i = c_k\} \geq 1$. For each class the index computes the ratio between the entropy of the distribution of the sizes of the corresponding connected components and the maximal entropy associated with the overall size (total number of cells) of such class. Finally, the index is obtained by adding up these entropy ratios:

$$E = -\sum_{c_k \in C''} \frac{\sum_{l=1}^{L_k} \frac{|CC_k^l|}{N_k} \log(\frac{|CC_k^l|}{N_k})}{\log(N_k)} \tag{7}$$

Please note that this index gathers information on the relative distribution of the labels on the points, so that it is not affected by distance scaling transformations. Finally, the use and interpretation of this new index can be formalized, once again, by estimating the p-value associated with the obtained value of E via an appropriate randomization simulation scheme.

4. Comparative Evaluation of Indexes

4.1. Simulation Examples

The following family of examples comparatively illustrates how the performance of the different indexes depends on the size and shape of the geographical regions associated with the different categories. We consider a region $\mathcal{R} = [-5, 5] \times [-1, 1] \subset \mathbb{R}^2$ where we can define three sub-regions whose size, range and shape depend on the value of a real parameter $a \in [0, 5]$:

$$\begin{aligned}
\mathcal{R}_1(a) &= \{(x_1, x_2) \in \mathcal{R} \mid |x_2 - \sin(x_1)| < 0.25 \text{ and } |x_1| < a\}, \\
\mathcal{R}_2(a) &= \{(x_1, x_2) \in \mathcal{R} \mid (x_1, x_2) \in \mathcal{R} \setminus \mathcal{R}_1(a) \text{ and } |x_1 + x_2| < 0.25\}, \\
\mathcal{R}_3(a) &= \mathcal{R} \setminus (\mathcal{R}_1(a) \cup \mathcal{R}_2(a)).
\end{aligned}$$

By selecting the value for $a \in [0, 5]$ we can adjust the size of $\mathcal{R}_1(a)$ which has the shape of a merging (thick) sine graph in $\mathcal{R}$; this variation will allow a characterization of the sensitivity of the indexes to the size and shape of the regions. Now, each point $\mathbf{x}$ is assigned a label value $y(\mathbf{x}) \in \{\text{Red}, \text{Blue}, \text{Green}\}$ according to the following distribution:

$$\begin{array}{ll} P(Y = \mathrm{R}) = 0.9, P(Y = \mathrm{B}) = P(Y = \mathrm{G}) = 0.05, & \text{if } (x_1, x_2) \in \mathcal{R}_1(a), \\ P(Y = \mathrm{B}) = 0.9, P(Y = \mathrm{R}) = P(Y = \mathrm{G}) = 0.05, & \text{if } (x_1, x_2) \in \mathcal{R}_2(a), \\ P(Y = \mathrm{R}) = P(Y = \mathrm{B}) = P(Y = \mathrm{G}) = \frac{1}{3}, & \text{if } (x_1, x_2) \in \mathcal{R}_3(a). \end{array} \qquad (8)$$

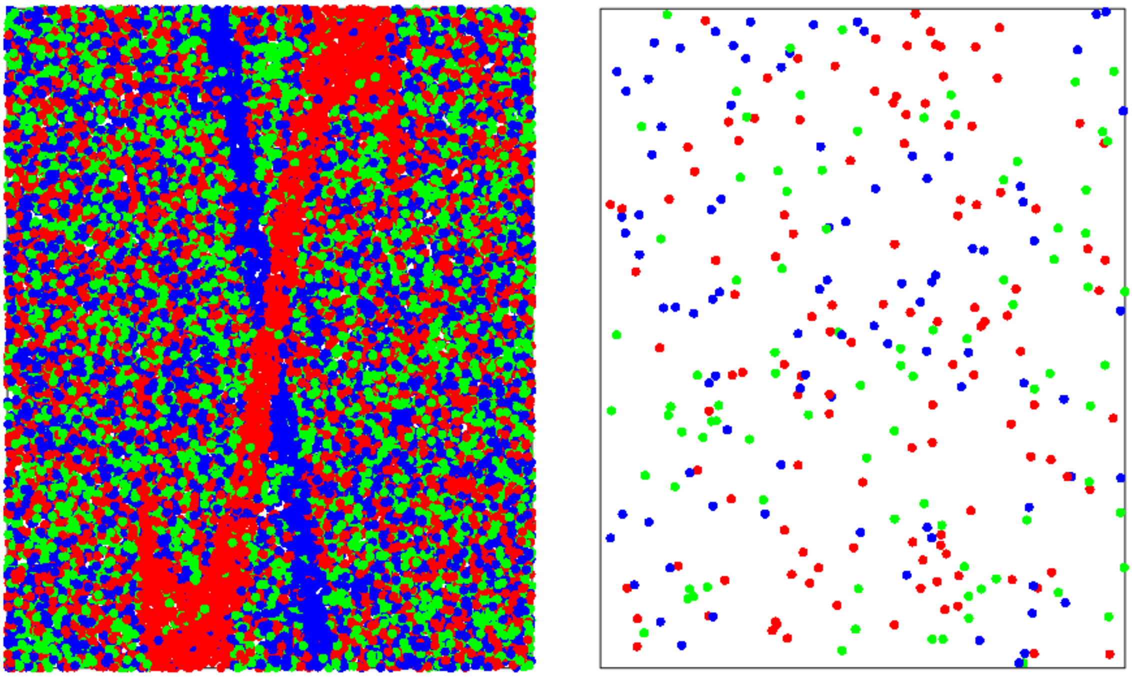
Based on taking uniform samples in $\mathcal{R}$ following such distribution, we now analyze the performance of the different indexes.

4.1.1. Detailed Analysis for $a = 2.5$

For the case of $a = 2.5$, Figure 2a shows the spatial distribution of a sample ($\mathbf{X}$ following a uniform distribution within $\mathcal{R}$) of 25000 points, which have been color labeled according to the $y(\mathbf{x})$ distribution in (8); the only purpose of this large number of points is to clearly delineate the regions $\mathcal{R}_i(2.5)$, $i = 1, 2, 3$. Figure 2b displays the spatial distribution of only the first 300 sampled points and, for this case, Figure 2c shows the colored groups of the corresponding Voronoi cells. Finally, for comparative purposes, Figure 2d displays the Voronoi cells for the same spatially distributed 300 points after having performed a random shuffling of the color labels.

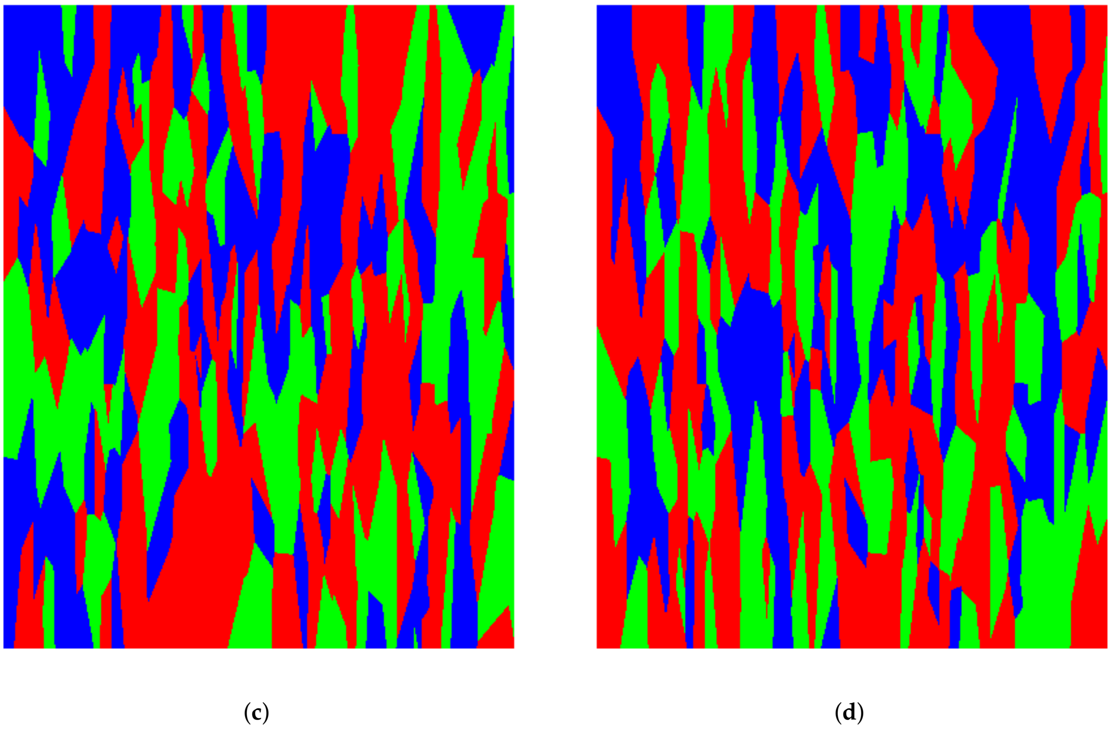
Monte Carlo simulations were performed by taking different samples of size 300 to estimate the p-values corresponding to the Moran's I, Geary's C, Mutual Information M, Herrera's ratio D and the Voronoi-based index E. Since, as mentioned above, Moran's I and Geary's are sensitive to the number assignment function z, we assigned numbers to the labels in two different ways, z_1 and z_2, so that: $z_1(\mathrm{R}) = 0, z_1(\mathrm{B}) = 1, z_1(\mathrm{G}) = 2$ and $z_2(\mathrm{R}) = 0, z_2(\mathrm{B}) = 2, z_1(\mathrm{G}) = 1$.

For each sample of size 300 a randomization technique was applied 200 times, by randomly shuffling the labels among all data geolocations, for estimating the p-value corresponding to each index. 200 Monte Carlo simulations were performed to assess the distribution of the estimated p-values. Table 1 shows a summary of the p-value distribution for the different indexes when applied to the data generated in this example.



(**a**) (**b**)

Figure 2. *Cont.*



(**c**) (**d**)

Figure 2. (**a**) Distribution of 25,000 sampled labeled points. (**b**) Distribution of 500 sampled labeled points. (**c**) Colored groups of Voronoi cells for sample of 500 points. (**d**) Colored groups after randomly shuffling color labels for sample of 300 points.

Table 1. Summary of distribution of estimated p-values corresponding to Moran's I, Geary's C, Mutual Information M, Herrera's ratio D and the Voronoi-based index E. Number of points = 300; randomization iterations = 200; Monte Carlo iterations = 200.

	Min	**1st Q.**	**Median**	**3rd Q.**	**Max**	**Mean**	**Std**
Moran I (z_1)	0	0.03	0.135	0.37	0.98	0.237	0.262242
Moran I (z_2)	0	0.01	0.07	0.23	0.99	0.18735	0.261614
Geary C (z_1)	0	0.02	0.13	0.3925	0.97	0.2405	0.267861
Geary C (z_2)	0	0.03	0.15	0.43	0.99	0.2764	0.292252
Mut. Inf. M	0	0.02	0.135	0.305	0.975	0.207925	0.237545
Herrera D	0	0.05375	0.1835	0.3825	0.935	0.239050	0.224115
Voronoi E	0	0.0150	0.06	0.18	0.815	0.128475	0.172679

Please note that the Voronoi index provides the smallest expected p-value with also smallest standard deviation. The rest of indexes display similar behaviors.

4.1.2. Influence of the Value of Parameter a

In this subsection, the performance of the indexes is evaluated for different sizes and forms of region $\mathcal{R}_1(a)$ whose scope can be regulated as a function of the parameter a. In Figure 3 the expected p-value associated with each one of the indexes is displayed as a function of such parameter a. Note again that the Voronoi index consistently provides the smallest expected p-value. The rest of indexes perform well except for Herrera's index D which seems to have trouble for capturing the relationship when region $\mathcal{R}_1(a)$ has a large range (i.e., the parameter a takes values close to 5).

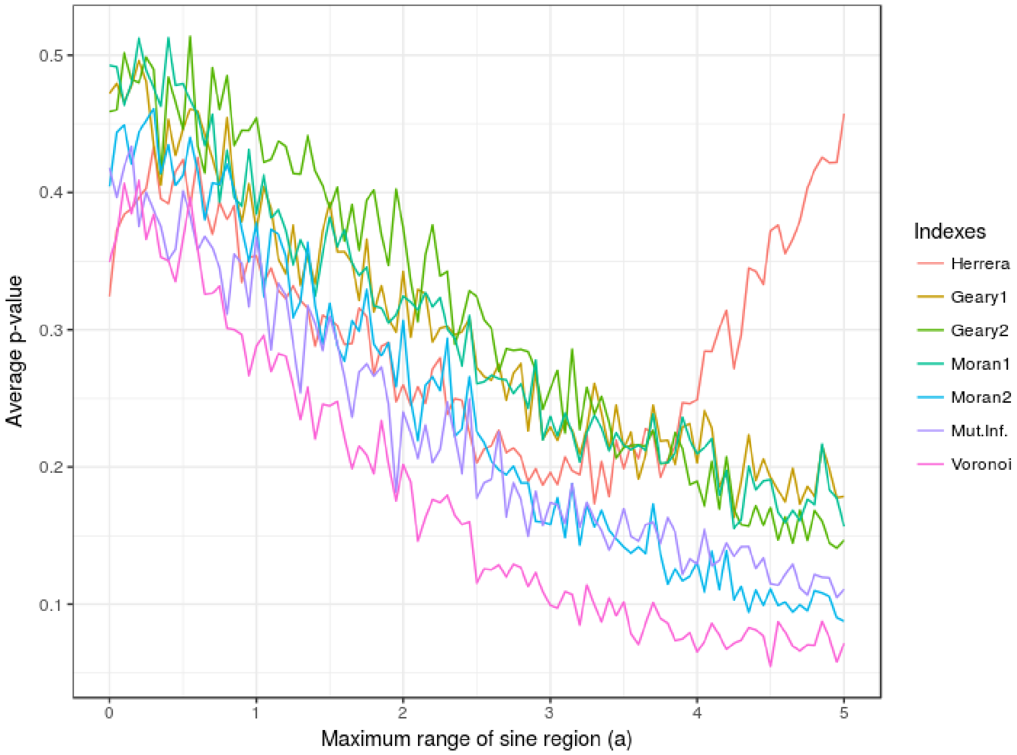


Figure 3. Variation of average p-values as a function of a corresponding to Moran's I, Geary's C, Mutual Information M, Herrera's D and the Voronoi-based index E. Number of points = 300; randomization iterations = 200; Monte Carlo iterations = 200.

4.2. Telephone Social Network Community Distribution in Cities

In [18] three social networks (in Portugal, Spain and France) were constructed derived from the communication activity between phone users, based on the CDRs registered by a telephone operator in those three countries. Then, communities were computed in those networks by applying the Louvain algorithm [17]. Next, also in [18], the geographical distribution of those communities along the telephone towers was computed by assigning to each tower the prevailing social network community among the phone users geographically associated with such tower. Finally, the relationship between communities and their geolocations was analyzed at country and city scales (Lisbon, Madrid, and Paris). The dependency of the communities on the geolocation was so high at country scale when compared to the city scale that an independence hypothesis was formulated at the city scale, claiming that the community distribution in cities is not affected by geolocation. We show now that the indexes presented in this work allow us to shed some light on this hypothesis. Table 2 shows the estimated p-values of the different indexes when applied to the data provided in [18] corresponding to the geographical distribution of communities along the telephone towers of the three cities (only best results between z_1 and z_2 are displayed for Moran's I and Geary's C). Please note that the extremely small p-values have been estimated by assuming an approximate Gaussian distribution of the corresponding index p-value and estimating its standard deviation via, again, a randomization procedure.

All indexes (except for Moran's I and Geary's C in Lisbon) suggest a strong dependency of communities on geolocation for all the cities. It is worth mentioning that Herrera's index, provided that a rigorous p-value analysis is performed, does also detect such dependency, even though it is weaker than the dependencies detected at the country scales. Hence, it was the lack of a p-value analysis in [18], where only absolute values of the index where relatively compared at both scales, that lead to a misleading independence hypothesis at the city scale.

For illustrative purposes, the community distribution along the towers for Paris is shown in Figure 4a, where each point presents a tower location and the corresponding color indicates the prevailing social network community in such tower. Figure 4b represents the corresponding Voronoi tessellation and Figure 4c displays the cells colored according to their corresponding y value. Finally, Figure 4d shows the respective results after shuffling the label values among the points, for comparative purposes. Visual inspection suggests that the number of pieces is clearly larger in the randomly shuffled case, this fact being in accordance with the extremely p-value obtained in Table 2.

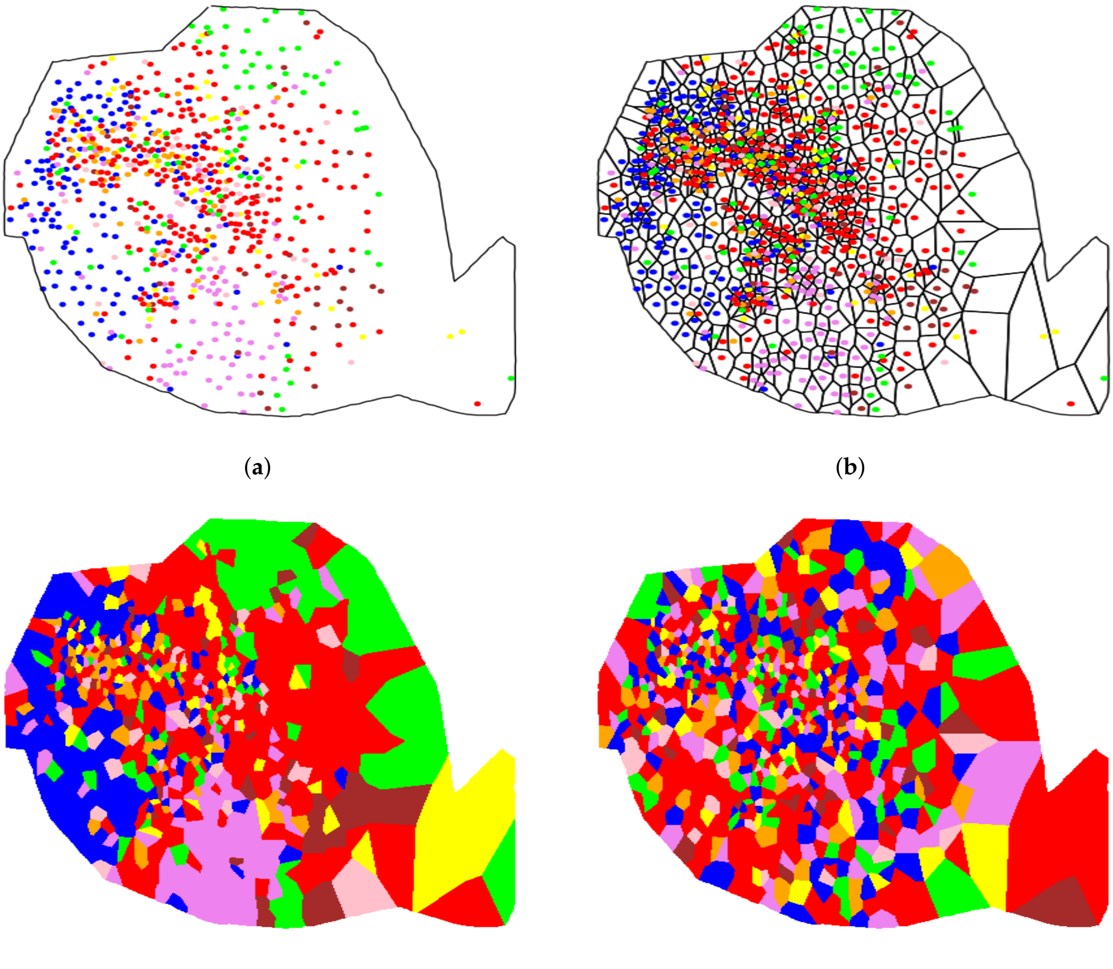


Figure 4. Analysis of Paris: (**a**) Community distribution along the telephone towers. (**b**) Corresponding Voronoi tessellation. (**c**) Colored groups of Voronoi cells. (**d**) Colored groups after randomly shuffling color labels.

Table 2. Estimation of p-values corresponding to the different indexes when testing the distribution of communities in Lisbon, Madrid and Paris.

	Paris	**Lisbon**	**Madrid**
Moran I	10^{-124}	0.34	0.002
Geary C	10^{-18}	0.37	0.0016
Mut. Inf. M	10^{-82}	0.0007	0.046
Herrera D	10^{-36}	0.0036	0.000034
Voronoi E	10^{-111}	10^{-15}	0.0045

5. Discussion

The ability of the presented indexes to evaluate the role of geolocation in the categorical variable depends on the size and shape of regions where the categories may be distributed. The new index based on the Voronoi tessellation, besides being directly applicable, has shown also a quite robust and efficient performance (with computational complexity $n \log n$) when compared with other indexes. It would be of interest to study further the topological properties of this new index, specifically the invariance against deformations beyond the group of motions in the plane. These techniques can be applied in many real scenarios as the one illustrated with the distribution of communities in cities.

6. Conclusions

Several indexes have been evaluated for assessing the relationship between geolocation (latitude and longitude) and a categorical variable, given a corresponding sample set. The direct application of the new index based on a Voronoi tessellation has proven to be robust when applied to different examples, including the geolocation of communities in some communication networks derived from CDRs.

Author Contributions: Both authors, P.J.Z. and M.Á.H.-M., contributed in the conceptualization, methodology, formal analysis, data curation and funding acquisition; P.J.Z. developed the software and wrote the original manuscript draft.

Funding: The present work has been partially supported by projects MTM2015-67396-P and MTM2017-84098-P from Ministerio de Economía y Competitividad (MINECO), Spain.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CDRs	Call Detail Records
ANOVA	Analysis of variance
MANOVA	Multivariate Analysis of Variance

References

1. Diggle, P.J. *Statistical Analysis of Spatial Point Patterns*; Academic Press: Cambridge, MA, USA, 1983.
2. Cressie, N. Statistics for spatial data. *Terra Nova* **1992**, *4*, 613–617. [CrossRef]
3. Davis, J.C.; Sampson, R.J. *Statistics and Data Analysis in Geology*; Wiley: New York, NY, USA, 1986; Volume 646.
4. White, M.J. The measurement of spatial segregation. *Am. J. Sociol.* **1983**, *88*, 1008–1018. [CrossRef]
5. Morisita, M. Measuring of the dispersion of individuals and analysis of the distributional patterns. *Mem. Fac. Sci. Kyushu Univ. Ser. E* **1959**, *2*, 5–23.
6. Clark, P.J.; Evans, F.C. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* **1954**, *35*, 445–453. [CrossRef]
7. Moran, P.A. The interpretation of statistical maps. *J. R. Stat. Society. Ser. B* **1948**, *10*, 243–251. [CrossRef]
8. Geary, R.C. The contiguity ratio and statistical mapping. *Inc. Stat.* **1954**, *5*, 115–146. [CrossRef]
9. Ripley, B.D. The second-order analysis of stationary point processes. *J. Appl. Probab.* **1976**, *13*, 255–266. [CrossRef]
10. Candia, J.; González, M.C.; Wang, P.; Schoenharl, T.; Madey, G.; Barabási, A.L. Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A Math. Theor.* **2008**, *41*, 224015. [CrossRef]
11. Zufiria, P.J.; Pastor-Escuredo, D.; Úbeda-Medina, L.; Hernandez-Medina, M.A.; Barriales-Valbuena, I.; Morales, A.J.; Jacques, D.C.; Nkwambi, W.; Diop, M.B.; Quinn, J.; et al. Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PLoS ONE* **2018**, *13*, e0195714. [CrossRef] [PubMed]
12. Beraldi, P.; Bruni, M.E. A probabilistic model applied to emergency service vehicle location. *Eur. J. Oper. Res.* **2009**, *196*, 323–331. [CrossRef]

13. Alvaro-Hermana, R.; Fraile-Ardanuy, J.; Zufiria, P.J.; Knapen, L.; Janssens, D. Peer to peer energy trading with electric vehicles. *IEEE Intell. Transp. Syst. Mag.* **2016**, *8*, 33–44. [CrossRef]
14. Bin, S.; Yuan, L.; Xiaoyi, W. Research on data mining models for the internet of things. In Proceedings of the Image Analysis and Signal Processing (IASP), Zhejiang, China, 9–11 April 2010; pp. 127–132.
15. Luong, N.C.; Hoang, D.T.; Wang, P.; Niyato, D.; Kim, D.I.; Han, Z. Data collection and wireless communication in Internet of Things (IoT) using economic analysis and pricing models: A survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 2546–2590. [CrossRef]
16. Chen, C.P.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [CrossRef]
17. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
18. Herrera, C. Socio Geographical Patterns Inferred From Mobile Phone Records. Ph.D. Thesis, Universidad Politécnica de Madrid, Madrid, Spain, 2017.
19. McDonald, J.H. *Handbook of Biological Statistics*; Sparky House Publishing: Baltimore, MD, USA, 2009; Volume 2.
20. Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
21. Gretton, A.; GyĂśrfi, L. Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* **2010**, *11*, 1391–1423.
22. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
23. Pál, D.; Póczos, B.; Szepesvári, C. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–9 December 2010; pp. 1849–1857.
24. Sricharan, K.; Raich, R.; Hero, A.O. Estimation of nonlinear functionals of densities with confidence. *IEEE Trans. Inf. Theory* **2012**, *58*, 4135–4159. [CrossRef]
25. Sricharan, K.; Raich, R.; Hero III, A.O. Empirical estimation of entropy functionals with confidence. *arXiv* **2010**, arXiv:1012.4188.
26. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In Proceedings of the International Conference on Algorithmic Learning Theory, Singapore, 8–11 October 2005; pp. 63–77.
27. Sejdinovic, D.; Sriperumbudur, B.; Gretton, A.; Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.* **2013**, *41*, 2263–2291. [CrossRef]
28. Székely, G.J.; Rizzo, M.L.; Bakirov, N.K. Measuring and testing dependence by correlation of distances. *Ann. Stat.* **2007**, *35*, 2769–2794. [CrossRef]
29. García, J.E.; González-López, V.A. Independence tests for continuous random variables based on the longest increasing subsequence. *J. Multivar. Anal.* **2014**, *127*, 126–146. [CrossRef]
30. Ross, B.C. Mutual information between discrete and continuous data sets. *PLoS ONE* **2014**, *9*, e87357. [CrossRef] [PubMed]
31. Zufiria, P.J.; Hernandez-Medina, M.A. Characterizing the Spatial Distribution of Geolocated Categorical Values. *Int. J. Appl. Phys. Math.* **2019**, *9*, 47–53. [CrossRef]
32. Garson, G.D. *Testing Statistical Assumptions*; Statistical Associates Publishing: Asheboro, NC, USA, 2012.
33. Light, R.J.; Margolin, B.H. An analysis of variance for categorical data. *J. Am. Stat. Assoc.* **1971**, *66*, 534–544. [CrossRef]
34. Leibovici, D.G.; Bastin, L.; Jackson, M. Higher-order co-occurrences for exploratory point pattern analysis and decision tree clustering on spatial data. *Comput. Geosci.* **2011**, *37*, 382–389. [CrossRef]
35. Cliff, A.D.; Ord, K. Spatial autocorrelation: A review of existing and new measures with applications. *Econ. Geogr.* **1970**, *46*, 269–292. [CrossRef]
36. Burridge, P. On the Cliff-Ord test for spatial correlation. *J. R. Stat. Soc. Ser.* **1980**, *42*, 107–108. [CrossRef]

37. Claramunt, C. A spatial form of diversity. In Proceedings of the International Conference on Spatial Information Theory, Ellicottville, NY, USA, 14–18 September 2005; pp. 218–231.

38. Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S.N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*; John Wiley & Sons: New York, NY, USA, 2009; Volume 501.