Spatial entropy for biodiversity and environmental data: The R-package SpatEntropy

Spatial entropy for biodiversity and environmental data: the R-package SpatEntropy

Linda Altieri, Daniela Cocchi, Giulia Roli

Department of Statistical Sciences, University of Bologna, via Belle Arti, 41, 40126, Bologna, Italy.

{linda.altieri; daniela.cocchi; g.roli}@unibo.it

Abstract

Entropy measures are standard tools in environmental and ecological sciences to describe the heterogenity of data. This paper reviews a selection of spatial entropy indices, some of which are very recent, suitable to deal with spatial data on variables presenting a finite number of categories. A special focus is given on biodiversity data, but methods can be applied to any other environmental phenomena. The new R package SpatEntropy is here introduced to compute spatial entropy measures in practice. The extension from traditional entropy measures to their spatial version is a unique feature of the package, which is able to work with both areal and point data. A practical part is also presented, where two types of environmental data are considered, regarding trees biodiversity and urban expansion, respectively. The package SpatEntropy is run over these dataset and shown to represent a user-friendly and helpful tool for new package users.

Keywords: Spatial entropy; Shannon's entropy; biodiversity; tree species; urban expansion; R package

1 Software and data availability

Name of Package/Software: SpatEntropy package of R software.

Title of Package: Spatial Entropy Measures.

Date First Available: 2018-02-28 (Version 0.1.0).

Date Updated Version: 2021-04-07 (Version 2.0-1).

Developers: Linda Altieri, Daniela Cocchi, Giulia Roli.

Maintainer: Linda Altieri.

Contact: linda.altieri@unibo.it

Required Software: R (>= 3.5.0), a free software environment for statistical computing and graphics, available at https://cran.r-project.org/.

Required Package: spatstat (>= 2.0-0), available at https://cran.R-project.org/package=spatstat.

Automatically Imported Packages: spatstat.geom, spatstat.core, spatstat.linnet.

Package Availability: the SpatEntropy package is free and downloadable from the CRAN repository at https://cran.R-project.org/package=SpatEntropy.

Package Reference Manual: available for Version 2.0-1 at https://cran.r-project.org/web/packages/SpatEntropy/SpatEntropy.pdf

Package Description: The heterogeneity of spatial data presenting a finite number of categories can be measured via computation of spatial entropy. Functions are available for the computation of the main entropy and spatial entropy measures in the literature. They include the traditional version of Shannon's entropy (Shannon, 1948), Batty's spatial entropy (Batty, 1974), Batty's LISA entropy (O'Neill et al., 1988), Li and Reynolds' contagion index (Li and Reynolds, 1993), Karlstrom and Ceccato's entropy (Karlström and Ceccato, 2002), Leibovici's entropy Leibovici (2009), Parresol and Edwards' entropy (Parresol and Edwards, 2014) and Altieri's entropy (Altieri et al., 2018a). Full references for all measures can be found under the topic 'SpatEntropy'. The package is able to work with lattice and point data. The updated version works with the updated spatstat package (>= 2.0- 0). It also provides a more intuitive framework for all functions, including improved examples, and new data. The speed of most functions has also been substantially increased.

Data Availability: the SpatEntropy package includes two datasets. The first one, raintrees, is a marked point pattern dataset about four rainforest tree species available at http://www.ctfs.si.edu. It is a ppp object (see package spatstat) with 7251 points, containing: window, an object of type owin (see package spatstat), i.e. the 1000x500 metres observation area, x, y, the numeric vectors with points' coordinates, marks, a character vector matching the tree species to the data points. This dataset documents the presence of tree species over Barro Colorado Island, Panama. Barro Colorado Island has been the focus of intensive research on lowland tropical rainforest since 1923 (http://www.ctfs.si.edu). Research identified several tree species over a rectangular observation window of size 1000x500 metres; the tree species constitute the point data categorical mark. This dataset presents 4 species with different spatial configurations: Acalypha diversifolia, Chamguava schippii, Inga pezizifera and Rinorea sylvatica. The second dataset included in the package, turin, is a lattice dataset with Turin Urban Morphological Zones. Values are either 0 (non-urban) or 1 (urban). Pixels outside the administrative borders are classified as NA. This raster/pixel/lattice dataset comes from the EU CORINE Land Cover project (EEA, 2011) and is dated 2011. It is the result of classifying the original land cover data into urbanised and nonurbanised zones, known as *Urban Morphological Zones* (UMZ). UMZ data are useful to identify shapes and patterns of urban areas, and thus to detect what is known as urban sprawl. Turin's metropolitan area is extracted from the European dataset and is composed by the municipality of Turin and the surrounding municipalities: Beinasco, Venaria Reale, San Mauro Torinese, Grugliasco, Borgaro Torinese, Collegno, Pecetto Torinese, Pino Torinese, Moncalieri, Nichelino, Settimo Torinese, Baldissero Torinese, Rivoli, Orbassano. The dataset is made of 111x113 pixels of size 250x250 metres. Many other data examples can be found in the just mentioned package spatstat, as well as in other R packages devoted to biodiversity and environmental studies (e.g. vegan).

2 Introduction

2.1 Entropy as a biodiversity and environmental measure

Measuring biodiversity with appropriate syntheses is a challenging task (see, e.g., Hoskins et al., 2020, and Drechsler, 2020). The proposals of indicators range from the simplest indices denoting species richness (see, e.g., Fisher et al., 2012) towards more articulated proposals related to the abundance of species in ecological communities (see, e.g., Scarnati et al., 2009). Under this perspective, entropy indices certainly represent an important and efficient tool widely used for properly measuring biological diversity (Leinster and Cobbold, 2012). Shannon's entropy (Shannon, 1948) is the most common and successful formulation, as it is able to synthesize several concepts in a single number: entropy, information, heterogeneity, surprise, contagion. The flexibility of such index and its ability to describe any kind of data, including categorical variables, motivate its diffusion across several applied fields for data description and interpretation (Frosini, 2004). Over the last decades, the concept of entropy has been further developed and generalized in engineering and statistics (Cover and Thomas, 2006), also extending to the case of continuous variables, firstly dealt by Rényi (1961) and Batty (1974, 1976, 2010, 2014).

In the wide context of environmental studies, such as geography, ecology, biology and landscape analyses, a large use of entropy measures has been made. In such cases, the researcher often deals with spatial data, i.e., data that are georeferenced as points or areas, and, as a consequence, also entropy measures need to be revised to include this spatial information. At this regard, a number of works are available in the literature, aiming at building a spatial entropy index. They can be ascribed to three main approaches. The first starts with Batty (1974, 1976, 2010, 2014) who defined a spatial entropy measure which evaluates the distribution of an event over an area, allowing for the unequal space partition into sub-areas. Later, Karlström and Ceccato (2002) modified the initial proposal in order to satisfy the property of additivity in terms of decomposition of the global index into local components, firstly introduced by Theil (1972) and then defined by Local Indices of Spatial Association (LISA) criteria (Anselin, 1995). The second approach to spatial entropy includes space based on a suitable transformation of the study variable to account for the distance between realizations (co-occurrences); the first proposal was made by O'Neill et al. (1988) for contiguous couples of realizations, then extended by Leibovici (2009) and Leibovici et al. (2014) to further distances and general degrees of co-occurrences. Contagion indices (Li and Reynolds, 1993; Parresol and Edwards, 2014) are also based on this view: spatial contagion is the opposite of entropy. As for the third approach, a set of spatial entropy measures has been presented by Altieri et al. (2018a, 2019a,b), starting from the co-occurrence approach but overcoming some undesirable features of the previous measures. According to this framework, Shannon's entropy of the transformed variable is decomposed into the information due to space and the remaining one, once space is considered. The proposal solves the problem of preserving additivity and disaggregating results, allowing for partial and global syntheses.

2.2 Available software packages for entropy

The `R` software (`R` Core Team, 2017) is certainly one of the most flexible options for performing statistical analysis, in particular when spatial data have to be managed. `R` is employed by a wide part of the global community of statisticians and has the great advantage of being open source, i.e. it can be downloaded freely and anyone can contribute to the software by submitting packages of functions for additional features. Packages undergo severe checks by the R-team, which ensure the quality of the available material. The software main drawback is that its traditional interface is not very user-friendly; this can be solved by installing an interface such as RStudio (`https://www.rstudio.com/`).

As far as entropy measures are concerned, several `R` packages on CRAN (`https://cran.r-project.org`) are available. The most common ones are `entropart`, `entropy` and `EntropyEstimation`. They all allow traditional entropy computation and decomposition into its two terms known as mutual information and conditional entropy, generally referred to in information theory applications (Cover and Thomas, 2006). In particular, `entropart` provides functions to calculate the alpha, beta and gamma diversity indices of communities, typical of ecological studies, together with Simpson's evenness index. The package also offers an example dataset about tree species, but its data format is very specific of the package and not easy to understand for `R` beginners. The further package `entropy` is more focused on various estimators of entropy, a topic which does not constitute the core of the present work, nor of many biodiversity studies. In particular, it allows to compute the Hausser and Strimmer shrinkage estimator, the maximum likelihood estimator and its Miller-Madow correction, various Bayesian estimators including the NSB estimator and the Chao-Shen estimator. It also provides functions for estimating the Kullback-Leibler divergence, and for the partition of entropy between mutual information and residual entropy. Its functions start from the observed counts of species, without any considerations about the spatial location of data; the package does not offer any data example. Lastly, `EntropyEstimation` partially overlaps with both the aforementioned packages, as it includes functions for the estimation of Shannon's entropy, variants of Renyi's entropy, mutual information, Kullback-Leibler divergence and generalized Simpson's indices; no data example is available. Further `R` packages, though containing the word "entropy" in the description, all deal with different data contexts or different fields of study (e.g. sampling entropy, see Altieri and Cocchi, 2021a).

Despite a wide literature on spatial entropy measures, as illustrated in Section 2.1, to our knowledge there is no `R` package offering an automatic computation of these indices. Given the potential and the crucial role of spatial entropy measures in real applications, this lack needs to be overcome by providing a practical and intuitive support for new users in managing spatial information, in form of point or areal data, and obtaining fast and interpretable results in the contexts of biodiversity and environmental studies.

2.3 `SpatEntropy`: a tool for easy computation of spatial entropy measures

This work aims at introducing the new `R` package `SpatEntropy` (an improved version of the original Altieri et al., 2018b), which collects functions for the computation of all the spatial entropy indices mentioned in Section 2.1. They include the traditional version of Shannon's entropy and Shannon's entropy of a transformation of the study variable X, known in the literature as Z, which considers pairs of observations at a given distance over space. Moreover, if offers computation of Batty's spatial entropy and its modified LISA version. Following the observation pairs approach, the package provides functions for computing O'Neill's entropy for adjacent couples and its transformations: Li and Reynolds' relative contagion index and Parresol and Edwards' entropy. In addition, the generalized version of Leibovici's entropy can be computed, together with the further generalization to a set of spatial entropy indices that we call Altieri's entropies. The package is able to work with both grid and point data, and makes extensive use of the widely known `spatstat` package (version $\geq 2.0-0$) functions and data structures (Baddeley et al., 2015). `SpatEntropy` allows usage by non-statisticians, provided they have basic knowledge of `R`; the minimum effort is requested from the user, and all functions are sided by data examples. Additionally, real datasets are provided for performing, as examples, the studies introduced in the present paper. Many other data examples can be found in the just mentioned package `spatstat`, as well as in other `R` packages devoted to biodiversity and environmental studies (e.g. `vegan`).

The present paper illustrates how to make use of `SpatEntropy` focusing on a main example concerning ecological data and witnessing the need to assess biodiversity via an entropy index that includes information about the data spatial structure. This example, consisting of a marked point pattern dataset about four rainforest tree species, is available in the `SpatEntropy` package under the name `raintrees`. This dataset documents the presence of tropical tree species over Barro Colorado Island, Panama. Barro Colorado Island has been the focus of intensive research on lowland tropical rainforest since 1923 (`http://www.ctfs.si.edu`). Research identified several tree species over a rectangular observation window of size 1000×500 meters; the tree species constitute the point data categorical mark, i.e. a label attached to each point of the pattern. The `raintrees` dataset presents 4 species with different spatial configurations: *Acalypha diversifolia* (*acaldi*), *Chamguava schippii* (*cha2sc*), *Inga pezizifera* (*ingape*) and *Rinorea sylvatica* (*rinosy*). The overall dataset has a total number of 7251 points and is shown in Figure 1.

One further example, which is synthetically presented along the paper, highlights both the ability of the measures proposed in this paper to deal with different types of spatial data, and the wide potential of using properly calibrated entropy measures. The package `SpatEntropy` offers the binary areal dataset `turin`, which comes from the EU CORINE Land Cover project (EEA, 2011). Turin's metropolitan area is composed by the main municipality of Turin and its 15 surrounding municipalities, that constitute its commuting belt. The dataset is a rectangular matrix

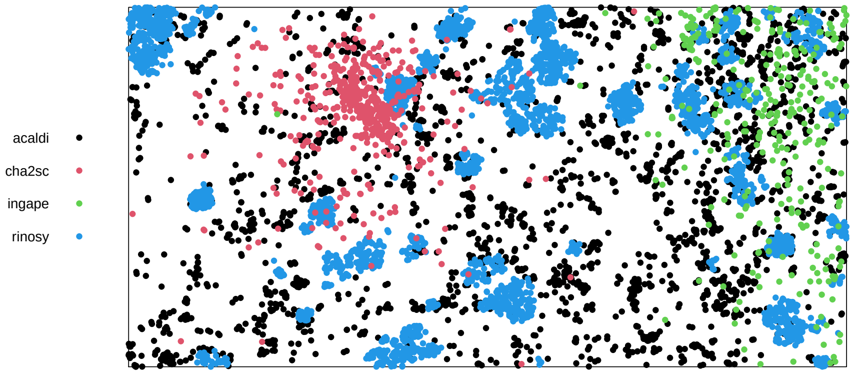

Figure 1: Rainforest tree data.

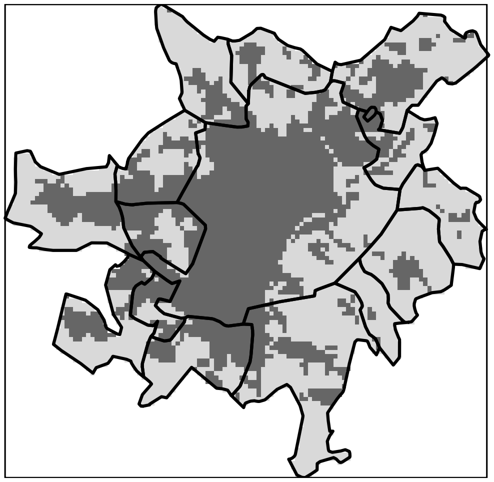

Figure 2: Turin urban data.

of 111 × 113 observations, each corresponding to a pixel of size 250 × 250 meters. Pixels within the area administrative borders are classified as 1 (urban) or 0 (non-urban), while pixels outside the border are missing values. The dataset is shown in Figure 2, where a dark pixel is urban and a light gray pixel is non-urban. In the package, the dataset is accompanied by the objects `turinW`, i.e. the enclosing rectangular window, and `turinTess`, which contains the names and administrative borders of all municipalities.

The paper develops as follows. In Section 3, the principles for introducing space in traditional entropy measures are presented to review the leading spatial entropy indices available in the literature. The application of the new `R` package `SpatEntropy` on the main running example of trees' biodiversity is carried on in Section 4. Section 5 uses `SpatEntropy` and illustrates the results on the second example, i.e. areal data of urban sprawl. Finally, Section 6 concludes the work.

3 From traditional to spatial entropy measures

The traditional concept of entropy dates back to Shannon's formula (Shannon, 1948), which in Information Theory (Cover and Thomas, 2006) quantifies the average amount of information brought by a discrete random variable X, taking values x_i in a set of I outcomes, according to the probability mass function (pmf) p_X. In biodiversity applications, X represents the variety of life considered, i.e., the levels of biodiversity commonly discussed (genetic, species and ecosystem diversity). As an example, in the rainforest tree data introduced above, X classifies the trees into the $I = 4$ different species x_1 to x_4, i.e. *Acalypha diversifolia*, *Chamguava schippii*, *Inga pezizifera* and *Rinorea sylvatica*.

Shannon's entropy of X is then defined as

$$H(X) = \sum_{i=1}^{I} p(x_i) \log \left(\frac{1}{p(x_i)} \right) \tag{1}$$

and represents the expected value of the so-called information function $I(p(x_i)) = \log(1/p(x_i))$; entropy measures the information or, in other words, the surprise, coming from observing realizations. Intuitively, outcomes with a very low probability of occurrence (i.e., hardly observable species) increase the entropy value, while outcomes very likely to occur (i.e., more observable species) give a small contribution to entropy. Thus, entropy, or analogously information and surprise, will be larger when the observed outcomes are not likely to occur. Entropy ranges in $[0, \log(I)]$ and its maximum value is achieved when X is uniformly distributed.

In many situations, entropy is seen as a descriptive measure and the pmf of X is built using the relative frequencies of the observed categories. When the goal is to make inference on the underlying process determining the observed outcomes, entropy can also be seen as an estimator $\widehat{H}(X)$.

In its standard use, the pmf is estimated by the so-called plug-in estimator (Paninski, 2003). In such case the descriptive measure and the estimator coincide: $\widehat{p}(x_i) = n_i/n$ substitutes the probabilities $p(x_i)$ with the observed relative frequencies over n realizations. In particular, referring to the rainforest tree example the frequencies are the ratios of the count of each species over the total number of observed trees. In the pllied part of the paper, all entropy measures are computed by substituting the elements of the unknown probability distribution with the observed relative frequencies.

A major drawback of Shannon's entropy is that it does not account for the spatial location of occurrences, so that datasets with identical pmf but very different spatial configurations share the same entropy value. In biodiversity this point is crucial, as the location of the variety of life considered is available (often as point coordinates), as well as strongly related to the research questions (the main aim is to quantify the spread of biodiversity over a certain space). Indeed, in the rainforest tree data example the interest lies in measuring the diversity across species but also in quantifying the spread of species according to the spatial location of the trees.

The following Sections present the mainstream approaches to build spatial entropy measures. Most indices imply the formal definition of a neighbourhood (Cressie, 1993). The simplest way of representing a neighbourhood system over n spatial units is via an adjacency matrix A (Anselin, 1995), i.e., a square matrix whose elements indicate whether pairs of units are neighbours: $a_{uu'} = 1$ if $u' \in \mathcal{N}(u)$, that is the neighbourhood of area u, $a_{uu'} = 0$ otherwise, and $a_{uu} = 0$ by definition. Spatial units may be points, defined via precise coordinate pairs, or, alternatively, areas, identified via common representative coordinate pairs, such as the area centroids. In biodiversity, the concept of spatial neighbourhood generalizes to an idea of similarity, so that some species are closer (e.g. in a biological, taxonomic or logical sense) than others (Leinster and Cobbold, 2012). Once the adjacency matrix is suitably specified, computations proceed the same way for any similarity system.

3.1 Batty's entropy

Batty's spatial entropy (Batty, 1974, 1976, 2010, 2014) is useful for evaluating the heterogeneity in the distribution of a phenomenon over an area. It is particularly appropriate when the observation area is exogenously partitioned into sub-areas (such as municipality administrative boundaries for a region). The use of a meaningful exogenous area partition is very important, since conclusions are heavily affected by the partition itself.

Let a phenomenon of interest F occur over an observation window of size T partitioned into G areas of size T_g. This defines G dummy variables identifying the occurrence of F over a generic area g, with $g = 1, \ldots, G$. Note that the phenomenon is denoted by the symbol F, a presence/absence variable, different from X, a generic categorical variable. Examples of F are: the presence of a single species in biodiversity studies, like in the case of point data on rainforest trees, or the observation of urban versus non-urban patches in city expansion studies, like in areal data on binary Turin's dataset. Given that F occurs over the whole window, its occurrence in area g takes place with probability p_g, where $1 - p_g = \sum_{g' \neq g} p_{g'}$ and $\sum_g p_g = 1$. The phenomenon intensity is obtained as $\lambda_g = p_g / T_g$, where T_g is the area size, and is assumed constant within each area. Batty's spatial entropy is then defined with respect to these quantities as

$$H_B(F) = \sum_{g=1}^{G} p_g \log\left(\frac{T_g}{p_g}\right). \tag{2}$$

It expresses the average amount of information brought by the occurrence of F in any area in the observation window, accounting for unequal space partition through the multiplicative component T_g. Note that $p_g \neq p(x_i)$ as the latter is the probability of occurrence of a category (i.e. a species) anywhere over the window, while p_g expresses the probability of occurrence of the phenomenon under study over a sub-area.

Analogously to Shannon's entropy, which is high when the I categories of X are equally represented over a (non spatial) data collection, Batty's entropy is high when F is equally intense over the G areas partitioning the observation window (i.e., when $\lambda_g = \lambda$ for all g). Batty's entropy $H_B(F)$ reaches a minimum value equal to $\log(T_{g^*})$ when $p_{g^*} = 1$ and $p_g = 0$ for all $g \neq g^*$, with g^* denoting the area with the smallest size. The maximum value of Batty's entropy is $\log(T)$, reached

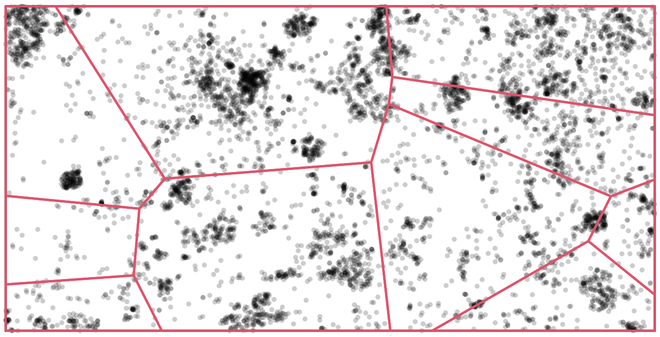

Figure 3: Example of partition into 10 sub-areas for rainforest tree dataset.

when the intensity of F is the same over all areas, i.e., $\lambda_g = 1/T$ for all g.

In the case of point data, Batty's spatial entropy aggregates information over each sub-area into a single number p_g, thus neglecting the detailed information about the points coordinates. This represents an unpleasant pitfall.

For the rainforest tree data, Batty's entropy may be computed if the aim lies on measuring the entropy observed across sub-areas of interests, controlling for their dimensions. In such case, entropy is computed either on the overall dataset, discarding the information about the different kinds of species (see, e.g., Figure 3, where $G = 10$), or by considering each category of species separately. In the latter case, the starting variable X, covering $I = 4$ species, is transformed into a dummy variable indicating the presence/absence of one species of interest over the sub-areas; if the interest lies in the entropy levels of each species over the sub-areas, 4 dummy variables are needed and 4 Batty's entropies are obtained and compared, without the possibility to have a unique measure.

3.2 Batty's LISA entropy

Karlström and Ceccato (2002) made a challenging proposal to exploit the contribution of neighbourhood in Batty's entropy index, following the Local Indices of Spatial Association (LISA) theory (Anselin, 1995). They consider an adjacency matrix for each of the G sub-areas introduced in Section 3.1, i.e. $A = \{a_{gg'}\}_{g,g'=1,\dots,G}$. The elements on the diagonal of the adjacency matrix A are $a_{gg} = 1$, i.e., each area neighbours itself. Then, the elements $a_{gg'}$ of this matrix are included to weight the probability of occurrence of F in a given spatial unit g:

$$\widetilde{p}_g = \sum_{g'=1}^{G} a_{gg'} p_{g'}. \tag{3}$$

Mimicking Batty's entropy, Batty's LISA entropy index is then defined as

$$H_{LISA}(F) = \sum_{g=1}^{G} p_g \log\left(\frac{1}{\widetilde{p}_g}\right). \tag{4}$$

i.e. it fixes $T_g = 1$ and defines the information function as $I(\widetilde{p}_g) = \log(1/\widetilde{p}_g)$.

The main purpose of Batty's LISA index lies again on measuring and comparing entropies at different levels of biodiversity (e.g., separately by species) across the sub-areas of interests, but, instead of controlling for their dimensions, a sub-area neighbourhood is included in the computation (where, again, neighbourhood can be extended to a more general meaning than purely spatial adjacency). This approach further allows to obtain local entropies

$$L_g = p_g \log\left(\frac{1}{\widetilde{p}_g}\right) \tag{5}$$

which represent the specific contribution of each sub-area and preserve the additive properties of LISA theory in deriving the global index $H_{LISA}(F)$.

The maximum of $H_{LISA}(F)$ does not depend on the choice of the neighbourhood and is $\log(G)$. As A tends to the identity matrix, $H_{LISA}(F)$ tends to Batty's spatial entropy (2), with equality in the case of $T_g = 1$ for all g.

3.3 O'Neill's entropy and contagion indices

Another way to build a spatial entropy measure relies on defining a new categorical variable Z, where each realization identifies ordered couples (x_i, x_j) of occurrences of X over space. Order preservation within couples regards considering the relative spatial location of the observations. Conventionally, if order is preserved the couple (x_i, x_j) implies that the observation carrying the j-th category occurs at the right or below the observation carrying the i-th category. Under this criterion, such couple is different from (x_j, x_i). For I categories of X, the new variable Z has $R = I^2$ categories. The attention moves from the computation of (1), namely $H(X)$, to an index of the same form, i.e. Shannon's entropy of Z, $H(Z)$, based on the pmf p_Z.

In biodiversity, entropy measures based on Z are useful when the variable of interest has two or more categories, e.g. species, and when the goal is to understand how the presence of a species at one location affects neighbouring outcomes, i.e. the adjacent presence of the same or a different species. Intuitively, when the variable is strongly spatially associated, neighbouring outcomes are closely related and the surprise (and thus, the entropy) in observing data decreases. Defining the type of neighbourhood to consider is crucial, so that different proposals based on this choice are available in the literature.

O'Neill et al. (1988) proposed an early spatial entropy index by defining a neighbourhood based on contiguity, i.e. areal units sharing a border. Shannon's entropy (1) is computed in this case for the subset of values of the variable Z made of contiguous couples. Such couples are identified by non-zero elements in a suitable adjacency matrix, the contiguity matrix C. The subset of couples of contiguous realizations is denoted by $Z|C$, with pmf $p_{Z|C}$. Then, the associated Shannon's entropy is

$$H_O(Z|C) = \sum_{r=1}^{I^2} p(z_r|C) \log\left(\frac{1}{p(z_r|C)}\right). \tag{6}$$

This measure ranges from 0 to $\log(I^2)$ and quantifies the residual amount of entropy associated to the variable of interest, once the influence of the spatial configuration in terms of contiguity has been taken into account. When it is close to 0, the heterogeneity of data mostly depends on its spatial structure and the residual entropy is low. When it is close to its maximum, data do not present a strong (positive or negative) spatial correlation and are more similar to a randomly scattered scheme.

A direct derivation of this approach is based on the concept of contagion, which is the opposite of entropy. Indeed, the Relative Contagion index RC (Li and Reynolds, 1993) is

$$H_{RC}(Z|C) = 1 - H_O^{norm}(Z|C) = 1 - \frac{1}{\log(I^2)} \sum_{r=1}^{I^2} p(z_r|C) \log\left(\frac{1}{p(z_r|C)}\right) \tag{7}$$

where the second term is the normalized entropy of $Z|C$, obtained via the multiplication of (6) by

$1/\log(I^2)$. Its complement to 1 measures relative contagion: the higher the spatial contagion between categories of $Z|C$, the lower the spatial entropy. Normalization has the advantage of allowing comparison across datasets presenting a different number of categories. giusto non andare a capo? If one wants to account for the number of categories of X when computing the contagion index, non-normalized measures should be computed in order to distinguish among contexts with different numbers of categories. For this reason, Parresol and Edwards (2014) suggest an unnormalized version of (7):

$$H_P(Z|C) = -H_O(Z|C) = \sum_{r=1}^{I^2} p(z_r|C)\log(p(z_r|C)) \tag{8}$$

thus ranging from $-\log(I^2)$ to 0. Both the Relative Contagion and Parresol-Edward's index have an opposite interpretation than O'Neill's entropy: increasing departures from 0 denote a stronger spatial structure in the data which affects the heterogeneity.

O'Neill's entropy and its modified versions are conceived for areal data, as they require information on sharing a border. A practical example on entropy measures based on contiguity relies on Turin's urban data. In the case of biodiversity point data, such as the tree species dataset, this approach implies a pixelization of the observations, which is a sensible choice, as it arbitrarily assignes a spatial dimension and a block structure to points, with great loss of information. Point data may be analysed using the proposal of distance, rather than contiguity, between observations. In such a situation, more suitable measures can be actually applied, such as the ones presented in the next Sections.

3.4 Leibovici's entropy

Following the idea of O'Neill, Leibovici (2009) and Leibovici et al. (2014) propose a richer measure of entropy by extending $H_O(Z|C)$ in two directions. Firstly, Z can now represent not only couples, but also triples and further degrees m of co-occurrences. The authors develop the case of ordered co-occurrences, so that the number of categories of Z is $R_m = I^m$. Secondly, space is now allowed to be continuous, so that areal as well as point data may be considered and associations may not coincide with contiguity: the concept of distance between occurrences generalizes the concept of contiguity.

Once a distance d of interest is fixed, then co-occurrences are defined for each m and d as the m-th degree simultaneous realizations of X at any distance $d^* \leq d$, i.e., distances are considered according to a cumulative perspective. To this aim, an adjacency hypercube A_d is built and a subset of all co-occurrences Z is selected conditional on it, i.e. $Z|A_d$ or, for the sake of simplicity, $Z|d$. Then, Leibovici's spatial entropy is again a version of Shannon's entropy defined on the pmf $p_{Z|d}$

$$H_L(Z|d) = \sum_{r=1}^{I^m} p(z_r|d) \log \left(\frac{1}{p(z_r|d)} \right). \tag{9}$$

The derivation of O'Neill's entropy (6) is straightforward as it is the special case when $m = 2$ and d is defined for contiguous co-occurrences. In applications, only co-occurrences of degree $m = 2$, i.e. couples of realizations, are usually considered; therefore, unless otherwise specified, from now on we refer to the case of couples ($m = 2$). Leibovici's approach finally allows to include the coordinates' information for point data in the computation of the entropy. If we again consider the example about the spatial pattern of the species, 16 couples can be observed: (*Acalypha*, *Acalypha*), (*Acalypha*, *Chamguava*), (*Acalypha*, *Inga*), (*Acalypha*, *Rinorea*), (*Chamguava*, *Acalypha*), ... and so on. Once a distance d has been fixed, only couples of species (x_i, x_j) located at a distance lower than d are isolated and enter the computation of $H_L(Z|d)$. Clearly, the choice of d affects this measure of entropy. Thus, in real applications, different scenarios for d need to be explored and compared. Overcoming this feature is among the reasons why a further generalization of this approach has been proposed by Altieri et al. (2018a) and described in the following Section.

3.5 Altieri's set of entropy measures

Altieri et al. (2018a, 2019a, 2019b) developed a more general approach defining a set of spatial entropy measures starting from O'Neill and Leibovici's indices. Such set should be employed when the interest lies in understanding the role of the spatial configuration in determining the entropy of a (biodiversity) variable, not only at one isolated specific distance but also at a global level or at several distance ranges simultaneously. In addition, it should be used when the influence of space needs to be quantified as a percentage of the entropy (e.g., in the rainforest tree example, how much the global entropy across species is due to the spatial location of the trees). From a statistical perspective this approach constitutes a more sophisticated tool that allows more flexibility and interpretability than other available measures of entropy.

The important starting point is a different way of computing the variable Z with respect to previuos definitions: the general degree m of co-occurrences is defined discarding the order within co-occurrences. For the sake of simplicity, let's consider only the case of $m = 2$, where, under this approach, pairs, not couples, of realizations are considered, i.e., the relative spatial location of the two realizations is irrelevant. The reason for this choice is two-fold. Firstly, ordering occurrences is not sensible in spatial statistics, where spatial configurations are not generally assumed to have a direction. Secondly, discarding the order ensures a one-to-one correspondence between Shannon's entropy of X and Z. Note that, when order is discarded, the number of categories of Z is smaller. See Altieri et al. (2019a) for further details on discarding the order. The gap between the two options grows as I increases, and induces a different computational burden for large datasets, adding a practical advantage to the choice of discarding the order.

The second step of this approach to spatial entropy is to introduce a discrete variable W, that represents space by classifying the distances at which the two occurrences of a pair take place. Classes w_k are defined, with $k = 1,\dots,K$, covering all possible distances within the observation window. Each distance class w_k implies the choice of a corresponding adjacency matrix A_k, which identifies pairs where the two realizations of X lie at a distance belonging to the range w_k.

In biodiversity contexts, considering different choices of distances at which pairs (e.g., pairs of observable species) occur is crucial, as well as quantifying the whole contribution of space on the global entropy measure.

Thanks to the introduction of W and following the basis of Information Theory (Cover and Thomas, 2006), the total entropy of Z may be decomposed as

$$H(Z) = \sum_{r=1}^{R} p(z_r) \log\left(\frac{1}{p(z_r)}\right) = MI(Z,W) + H(Z)_W \tag{10}$$

where, in the spatial context defined by the variable W, $MI(Z,W)$ is the spatial mutual information and is the part of entropy of Z due to the spatial configuration W, while $H(Z)_W$ is the spatial global residual entropy and represents the remaining information brought by Z after space has been taken into account. The more Z depends on W, i.e. the more the realizations of X are spatially associated (e.g., nearby observations of the same species), the higher the spatial mutual information. Conversely, when the spatial association among the realizations of X is weak (e.g., observations of a species tend to have neighbours of different species), the entropy of Z is mainly due to spatial global residual entropy.

The entropy $H(Z)$ is a stable reference value, while its two components $MI(Z,W)$ and $H(Z)_W$ vary and are able to assess the role of space for datasets with different spatial configurations. For the sake of interpretation and diffusion of the results, a proportional measure of spatial mutual information is:

$$MI_{prop}(Z,W) = MI(Z,W)/H(Z) \tag{11}$$

which ranges in $[0,1]$ and quantifies the contribution of space according to the total entropy of Z.

The overall value of $MI(Z,W)$, however, is often negatively influenced by what happens at large distance ranges, where scarce spatial correlation is usually present. Hence, spatial mutual information for the whole dataset may be low even when a clustered pattern occurs (e.g., groups of nearby realizations belonging to same species are observed). The variable W helps in overcoming this drawback, since K subsets of realizations of Z, denoted by $Z|w_k$, are available and spatial mutual information can be also decomposed as

$$MI(Z,W) = \sum_{k=1}^{K} p(w_k) PI(Z|w_k) \tag{12}$$

that is a weighted sum of partial terms $PI(Z|w_k)$, each denoted spatial partial information and defined as

$$PI(Z|w_k) = \sum_{r=1}^{R} p(z_r|w_k) \log\left(\frac{p(z_r|w_k)}{p(z_r)}\right). \tag{13}$$

Each partial term $PI(Z|w_k)$ quantifies the contribution to the departure from independence of each conditional distribution $p_{Z|w_k}$, i.e. the contribution of the k-th distance range to the global mutual information between Z and W. As regards biodiversity, each partial term measures the degree of association across species over space at each distance range. If this association is high, then pairs of observations belonging to the same species are more likely, that is a low level of biodiversity is detected. Conversely, a high biodiversity situation is revealed when the spatial associations of species is low, i.e. pairs of different species are more often observed. The interest usually lies on short distance ranges, but all depends on research focus and on the definition/dimensions of the observation window.

Analogously to (12), for spatial residual entropy the following additive decomposition holds:

$$H(Z)_W = \sum_{k=1}^{K} p(w_k) H(Z|w_k), \tag{14}$$

where the terms $H(Z|w_k)$, denoted as partial residual entropies, measure the partial contributions to the enmtropy of Z once the spatial configuration is controlled:

$$H(Z|w_k) = \sum_{r=1}^{R} p(z_r|w_k) \log\left(\frac{1}{p(z_r|w_k)}\right). \tag{15}$$

Each term $H(Z|w_k)$ thus represents a Leibovici's entropy index (9) on unordered couples, which is now included in a wider and complete framework for spatial entropy measures. In biodiversity, a high value for $H(Z|w_k)$, especially at short distance ranges, is a hint for declaring large levels of biodiversity.

4 A biodiversity study with `SpatEntropy`: the rainforest tree data

This Section offers a practical guide to the `SpatEntropy` package via the rainforest tree data example. In what follows, the functions and the possible options are described following the application; more detailed information is available in the package manual downloadable at `https://CRAN.R-project.org/package=SpatEntropy`, where a number of toy examples are included to further practice of the package. For simplicity, many results are presented by rounding to two decimal places, while `R` provides more precise values.

After `SpatEntropy` and its dependence package `spatstat` (`>2.0-0`) are installed from CRAN (`https://cran.r-project.org`), the graphical representation of data (as in Figure 1) may be produced with the commands

```
R> data(raintrees)
R> plot(raintrees, main="", pch=16, cols=1:4)
```

As introduced in Section 1, the variable X here represents the tree species, with $I=4$ categories, briefly coded as $x_1 =$ *acaldi*, $x_2 =$ *cha2sc*, $x_3 =$ *ingape* and $x_4 =$ *rinosy*. Consequently, the variable Z has $4^2 = 16$ categories when ordering within species couples is considered, or $\binom{5}{2} = 10$ categories when unordered pairs are built. The total number of trees is $N = 7251$, where $n_1 = 2678$ trees belong to species x_1, $n_2 = 544$ to species x_2, $n_3 = 311$ to species x_3 and $n_4 = 3718$ to species x_4. The relative amounts for each species are the frequencies that enter entropy computations (the input for Batty's entropy is different, as explained in Section 4.1). The four species have different spatial configurations: *acaldi* is evenly distributed over the region, *cha2sc* has a clustered configuration, *ingape* shows a tendency to grow on the right hand side of the area, and *rinosy* is multiclustered. Therefore, space is likely to play a role in the data heterogeneity.

As a starting point, the traditional version of Shannon's entropy is easily computed by

```
R> shannon(raintrees)
```

The function produces three output elements. The core information is returned by

```
$shann
[1] 1.04
```

which gives the value of Shannon's entropy for the four tree species. A comment on the entropy value may be facilitated by the relative value, obtained in the second element of output by the command

```
$rel.shann
[1] 0.75
```

This value states that the overall data entropy is 75% of the maximum possible entropy, in this case equal to $\log(4)$. Such measure is easier to interpret across disciplines and allows comparison to other datasets. A low value for the relative heterogeneity hints at a structure in the data, possibly of spatial nature, that decreases the level of chaoticity, but no deeper exploration is allowed by this traditional Shannon's entropy. The third element of the output is a table with the relative frequencies that enter the computation of $H(X)$ in place of the unknown probabilities:

```
$probabilities
category frequency
1 acaldi 0.37
2 cha2sc 0.08
3 ingape 0.04
4 rinosy 0.51
```

The entropy that is obtained as a first output is an estimate $\widehat{H}(X)$ of $H(X)$. If an uncertainty assessment is requested, the variance of Shannon's entropy can be estimated as $\widehat{V}[\widehat{H}(X)] = \widehat{H}(X)^{(2)} - (\widehat{H}(X))^2$ (Paninski, 2003), where $\widehat{H}(X)^{(2)} = \sum_{i=1}^{I} \frac{n_i}{n} \log\left(\frac{n}{n_i}\right)^2$. In the package, such variance is obtained with

```
R> varshannon(raintrees)
[1] 0.44
```

Afterwards, we illustrate what the different spatial entropy measures bring to the study of the biodiversity of rainforest tree data.

4.1 Results for Batty's entropy

As introduced in Section 3.1, the computation of Batty's entropy imposes two exogenous decisions: first, the definition of the phenomenon of interest, and second, the choice of suitable sub-areas. As a running example, let us compute entropy on all rainforest trees, i.e. on the presence/absence of trees without distinction among the 4 species. When there are no reasons for fixing a specific area partition, the package allows to randomly partition the observation window into G portions (and we propose the former $G = 10$ of Figure 3) by using

```
R> batty.ent=batty(unmark(raintrees), partition=10)
```

The function `unmark` discards the information about the species. The argument `partition` allows to choose G. In the function output, we first see

```
$areas.freq
area.id abs.freq rel.freq area.size
1   1   1160     0.16     56466.98
```

```
2   2   351       0.05      26545.91
3   3   704       0.10      77000.24
4   4   72        0.01      24134.30
⋮
```

reporting descriptive information about each sub-area: the number of trees (`abs.freq`), the corresponding relative frequency (`rel.freq`), which stands for p_g, and the sub-area size `area.size`, i.e. T_g. Then, the outputs

```
$batty
[1] 13.07
$rel.batty
[1] 0.99
```

provide Batty's entropy and its relative version, respectively. These results approach the maximum possible heterogeneity and are different from Shannon's entropy values. This happens because they take different perspectives: Batty's entropy says that, with the given area partition, rainforest trees are (almost) equally intensely distributed over the sub-areas.

For plotting the data according to the area partition (obtaining Figure 3), the following code is used:

```
R> plot(unmark(raintrees), pch=16, cex=0.6, main="")
R> plot(batty.ent$areas.tess, add=TRUE, border=2)
```

If the interest lies in focusing on a specific tree species, e.g. *Chamguava*, Batty's entropy can be computed by including the argument `category` into the command and by removing the function `unmark`

```
R>batty.ent=batty(raintrees, category="cha2sc", partition=10)
$batty
[1] 11.84
$rel.batty
[1] 0.90
```

In this case, the entropy value is lower than the one referred to all species, meaning that this particular species is not as evenly distributed as the overall dataset according to the area partition.

A more interesting approach is based on partitioning the area according to a covariate. Rainforest trees dataset is accompanied by the object `raintreesCOV`, i.e. a list of two environmental covariates: *soil elevation* and *slope*. These variables may affect the degree of biodiversity of species. Therefore, they can be used to define sub-areas based on their values. As an example, let's consider *slope*, which is a continuous variable, and 4 intervals according to its values. The categories of soil to derive these intervals are here obtained by using quantiles as breakpoints through the following code:

```
R> slopecut=cut(raintreesCOV$grad, breaks = quantile(raintreesCOV$grad,
   probs = (0:4)/4), labels = 1:4) # discretize the covariate
R> maskv=tiles=list(); for(ii in 1:nlevels(slopecut)) {
R> maskv[[ii]]=as.logical(c(slopecut$v)==levels(slopecut)[ii])
R> tiles[[ii]]=owin(xrange=data$window$xrange, yrange=data$window$yrange,
   mask=matrix(maskv[[ii]], nrow(slopecut$v))) }
```

The map of the original covariate *slope* for the whole area is shown in Figure 4, left panel, and is produced by:

```
R> par(mfrow=c(1,2))
R> plot(raintreesCOV$grad, main="", col=gray(seq(1,0,l=100)))
```

Figure 4, right panel, shows the area partition based on the above discretization of the covariate *slope* together with species *Chamguava*. It is obtained by running the following code:

```
R> plot(slopecut, main = "")
R> plot(split.ppp(raintrees)$cha2sc, add=TRUE, pch=16, cex=0.6, main="")
```

Once the partition is built, the relative frequencies of trees are computed for each discretized level of the variable *slope*, and Batty's entropy is derived by using the same command `batty()` and specifying the area partition. The following commands refer to species *Chamguava*:

```
R> slopetess=list(tiles=tiles, n=nlevels(slopecut))
R> batty(raintrees, category="cha2sc", partition=slopetess)
```

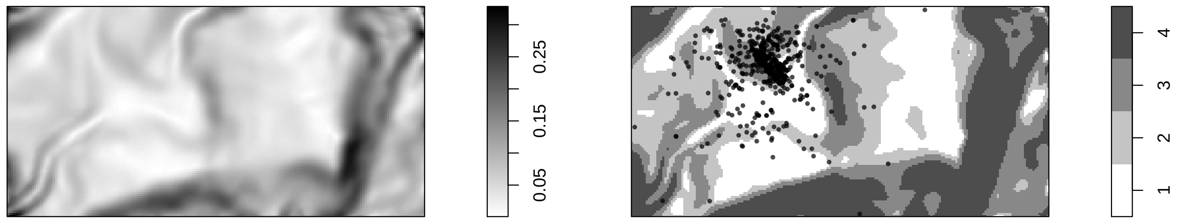

Figure 4: Left panel: original covariate *slope*. Right panel: discretized version of the covariate *slope* into 4 levels, together with species *Chamguava*.

```
$batty
[1] 12.83
$rel.batty
[1] 0.98
```

Interpretation is analogous to the former one, but something more can be said here. Firstly, the two partition options for Batty's entropy on the same species show how sensible the measure is to the chosen area partition: any interpretation must be carried out accounting for the choice. In addition, despite the evident tendency to spatial clustering, species *Chamguava* does not seem to have any dependence on the covariate *slope*: its entropy computed over the slope partition is high, close to the maximum, denoting that trees grow evenly across all covariate levels. A computation of Batty's entropy under a covariate-based partition can give hints on how biodiversity can be due to environmental covariates.

4.2 Results for Batty's LISA entropy

The LISA version of Batty's entropy can be built in the package `SpatEntropy` by using the same commands as before to obtain an area partition of interest. Moreover, this version needs a definition for the neighbourhood, that can be done by choosing either the number of neighbours for each sub-area, or a neighbourhood distance measured over the sub-areas' centroids. In the former case, consider as an example the case of 3 neighbours, a random partition in $G = 10$ sub-areas and the species *Chamguava*. Then, running the following command, one obtains

```
R> KC.ent=karlstrom(raintrees, category="cha2sc", partition=10, neigh=3)
$karlstrom
[1] 1.88
$rel.karl
[1] 0.82.
```

The following part of the output with details on the area partition is

```
$areas.freq
area.id abs.freq rel.freq neigh.mean area.size
1   1     0        0.00     0.01      57581.30
2   2     0        0.00     0.01      54517.04
3   3     307      0.56     0.20     101940.22
4   4     3        0.01     0.02     33719.95
⋮
```

that, with respect to the function output of Batty's entropy, has one new column with the average frequencies over the neighbouring areas, which stand for $\tilde{p}_g$ in the computations. Compared to Batty's entropy, given the area partition the consideration of the neighbouring probabilities decreases the relative entropy level. Other neighbourhood options may be tried for further comparison.

4.3 Results for Leibovici's entropy

Leibovici's entropy refers to Z, i.e. the variable which defines co-occurrences of rainforest trees. As anticipated in Section 3.3, O'Neill's entropy, that was the first entropy measure based on such variable, is unsuitable when point data are considered. The computation of Shannon's entropy of Z for unordered pairs of rainforest trees:

```
R> shannonZ(raintrees)
```

produces the same output of `shannon()`, but in this case the probability table reports the 10 frequencies of all unordered pairs of tree categories runs as:

```
$probabilities
pair           abs.freq   rel.freq
1 acaldiacaldi 3584503    0.14
2 cha2sccha2sc 147696     0.01
3 ingapeingape 48205      0.002
⋮
$shannZ
[1] 1.67
$rel.shannZ
[1] 0.72
```

Again, the entropy value is far from the maximum value, as can be seen from the relative version, and hints at a data structure that cannot be captured without spatial information. On the contrary, Leibovici's entropy, which, by default, works with ordered couples and needs the specification of a distance of interest d among the trees forming each pair, is much more appropriate. An example with $d = 10$ meters follows. The command is

```
R> leibovici(raintrees, ccdist=10)
```

where the argument `ccdist` allows to specify the value of d in the same unit of measurement as the observation area (metres in the case of raintrees). Results are in the output

```
$leib
[1] 1.12
$rel.leib
[1] 0.40
```

The value for Leibovici's entropy for $d = 10$ is quite low and different both from Shannon's entropy and from Batty's entropies: they measure different aspects of entropy. For Leibovici's entropy, a low value detects an important role of space for the heterogeneity of the tree species, when the potential spatial relationship is measured up to 10 metres. This measure is very interesting but offers a picture limited to the chosen distance, so that computations should be repeated for a number of different values for ds to get a general understanding of the data behaviour.

4.4 Results for Altieri's set of entropy measures

The recent set of entropy measures proposed by Altieri et al. (2018a) defines an exhaustive framework which overcomes the drawbacks of Leibovici's entropy. This approach considers several ranges which cover all the obsevation window in order to gain information on the spatial data heterogeneity at all distances. As a starting point, the distance classes are chosen. In this example, 4 ranges are considered for the variable W as $w_1 = [0,1[$; $w_2 = [1,2[$; $w_3 = [2,10[$; $w_4 = [10,d_{max}]$, where $d_{max} = 1118.03$ (in metres), in coherence with the observation area. Then, the spatial entropy measures are obtained by running the following code

```
R> outp=altieri(raintrees, distbreak=c(1,2,10), verbose=TRUE)
```

where the argument `distbreak` fixes the internal breaks of the distance classes and `verbose` controls whether information about the computation progress is printed in the output (`TRUE`) or not (`FALSE`). This function is the most sophisticated of the package and, even if very efficient, may take up to a few minutes for large datasets. Therefore, we recommend to use the command `verbose=TRUE` in order to check the progress of the computation. At the end of the process, the generated output is very detailed. First, it provides the quantities used to compute the set of entropy measures: distance breaks, the distribution of the distance classes, the number of pairs (in total and per class) and tables with the absolute and relative frequencies for all distance classes. Then, the global values for Spatial Mutual Information $MI(Z,W)$, Spatial Residual Entropy $H(Z)_W$, and their sum, i.e. Shannon's entropy $H(Z)$ are returned. The ilustration of his part is omitted in the present text, and we refer to the package manual for further details.

Finally, partial entropies are yielded, in both absolute and relative terms:

```
$SPI.terms
class [0, 1] class [1, 2] class [2, 10] class [10, 100] class [100, 1118.03]
0.70         0.64         0.56          0.09            0.001
```

```
$rel.SPI.terms
class [0, 1] class [1, 2] class [2, 10] class [10, 100] class [100, 1118.03]
0.39         0.37         0.33           0.05            0.0007
$RES.terms
class [0, 1] class [1, 2] class [2, 10] class [10, 100] class [100, 1118.03]
1.10         1.07         1.12           1.77            1.65
$rel.RES.terms
class [0, 1] class [1, 2] class [2, 10] class [10, 100] class [100, 1118.03]
0.61         0.63         0.67           0.95            0.9993.
```

The `SPI.terms` are the spatial partial information terms $PI(Z|w_k)$ and show the contribution of space to the data heterogeneity, i.e. to biodiversity, at each distance class; the `RES.terms` are the partial residual entropies $H(Z|w_k)$ and measure the biodiversity of trees due to other sources of heterogeneity, once space is controlled. These quantities represent the most interesting output to interpret. In particular, focusing on each distance range and on relative versions (`rel.SPI.terms` and `rel.RES.terms`), where the two components sum to 1 for each distance range, we can identify the role of space. For instance, for trees lying at most 1 metre apart, 39% of the heterogeneity is due to the spatial correlation, i.e. to pairs of trees belonging to the same species; conversely, the remaining percentage (61%) detects the part of entropy attributable to pairs of tree of different species, i.e. the biodiversity level conditional on this distance. The residual terms $H(Z|w_k)$s increase when the distance ranges rise (because the role of space naturally decreases at higher distances); when they assume high and leading values also at lower distances, this implies that data do not present a strong spatial structure and that we can expect high levels of biodiversity in small sites too. Therefore, partial residual entropies represent the main goal as they measure biodiversity once the effect of space is controlled. To conclude, the several tools proposed by Altieri et al. assess a remarkable biodiversity across the rainforest trees of Barro Colorado Island by exploring the phenomenon under a complete and very flexible framework.

5 SpatEntropy in other environmental applications: Turin's urbanization study

Entropy measures are not only a standard tool in biodiversity studies, but also may be widely employed in environmental applications such as landscape, land cover and land use data. As a further example of usage of the SpatEntropy package, an environmental study for the binary grid data about urbanization in Turin is considered. As introduced in Section 2, land cover data X are managed and dichotomized into urbanized ($x_1 = 1$) and non-urbanized ($x_2 = 0$) zones, i.e. pixels of size 250×250 metres (see Figure 2). The aim is to measure the urban dispersion of the city by highlighting the differences in the use of both spatial entropy indices and SpatEntropy with respect to the previous application in biodiversity field. The main functions, already described in Section 4 are only briefly cited in what follow. A command is added regarding the computation of O'Neill's entropy, since the concept of contiguity is well-defined for grid data, while it could not be used in the previous data example.

```
R> shannon(turin)
R> shannonZ(turin)
R> batty(turin, cell.size=250, partition=turinTess)
R> karlstrom(turin, cell.size=250, partition=turinTess, neigh=2)
R> oneill(turin)
R> leibovici(turin, cell.size=250, ccdist=400)
R> altieri(turin, cell.size=250, distbreak=c(1,2,5)*cell.size)
```

The main results are collected in Table 1, where in the left-hand side the absolute entropy values are reported, while the right-hand side shows the relative versions which are straightforward to comment. Firstly, the two non-spatial Shannon entropies are computed: $H(X)$ and $H(Z)$. These both take values close to their maximum ($\log(2)$ and $\log(3)$, respectively), identifying the metropolitan area of Turin as a highly chaotic area.

The sub-areas for the computation of Batty's entropy are the administrative borders of the 15 municipalities forming the metropolitan area (object turinTess, introduced in Section 2.3);

		Entropy	Relative entropy
Shannon of X $H(X)$		0.688	0.993
Shannon of Z $H(Z)$		1.033	0.941
Batty $H_B(F)$		16.295	0.821
Batty+LISA $H_{LISA}(F)$	2 neigh	1.923	0.710
Batty+LISA $H_{LISA}(F)$	4 neigh	2.140	0.790
Batty+LISA $H_{LISA}(F)$	12 neigh	2.243	0.828
O'Neill $H_O(Z\|C)$		1.060	0.764
Leibovici $H_L(Z\|d)$	$d = 400$	1.094	0.789
Leibovici $H_L(Z\|d)$	$d = 1000$	1.224	0.883
Leibovici $H_L(Z\|d)$	$d = 10000$	1.377	0.993
Altieri $PI(H\|w_k)$	Option 1		
	w_1=[0; 250]	0.321	0.247
	w_2=]250; 500]	0.223	0.177
	w_3=]500; 1250]	0.113	0.094
	w_4=]1250; d_{max}]	<0.001	<0.001
Altieri $PI(H\|w_k)$	Option 2		
	w_1=[0; 400]	0.288	0.224
	w_2=]400; 1000]	0.142	0.116
	w_3=]1000; 10000]	0.019	0.018
	w_4=]10000; d_{max}]	0.015	0.015

Table 1: Results for all entropy measures on Turin data. Batty's and Batty's LISA entropies are computed using the municipalities' administrative borders as partition

Batty's LISA entropy is obtained by further considering the 2, 4 or 12 nearest neighbours based on the pixels' centroids. The argument `cell.size`, to be ignored for point data, defines the dimension of the pixels in the area measurement unit (250×250 metres). The relative versions of all indices can be compared: Batty's entropy is similar to Batty's LISA entropy with the most extensive neighbourhood (12 neighbours), while smaller neighbourhood systems decrease the level of entropy. These values are below their maximum compared to Shannon's entropies, likely because the main municipality of Turin is more intensely urbanized than the municipalities of its belt.

O'Neill's and Leibovici's entropies can be commented together, as the former is a special case of the latter when d equals the cell width (in this case 250 metres). The option `ccdist` defines the distance between pixel centroids considered for building pairs. Three versions of Leibovici's entropy are computed, for $d = 400, 1000, 10000$ metres. O'Neill's entropy takes a relatively small value, equal to 76% of its maximum, meaning that, when considering adjacent couples, the different categories are not evenly distributed, so the spatial configuration is not random. By exploring the output of `oneill`, which has a similar structure to the other functions' outputs, and in particular the table of relative frequencies, the majority of couples are shown to be homogeneous, of type *(urban, urban)* (48% of the total) or *(non-urban, non-urban)* (39% of the total), while heterogeneous couples are a minority. Consequently, the index shows that the urban tissue tends to be compact, with a different conclusion with respect to the traditional Shannon's entropy. As d increases, the entropy value approaches its maximum as couples of pixels lying farther apart are included and the measure becomes more similar to Shannon's entropy.

As regards Altieri's set of entropy measures, the spatial partial information terms $PI(H|w_k)$ here represent the quantities of interest, as they measure the urban dispersion or compactness for Turin's area: a high value for $PI(H|w_k)$, especially at short distance ranges, is a hint for a compact urban expansion, which is a desirable feature as dispersed cities are proved to be inefficient and may impede sustainable development (Altieri et al., 2019b). We explore two options of distance ranges with respect to spatial information terms, in order to show the sensitivity of results based on such choice. The first one defines W as $w_1 = [0, 250]$; $w_2 =]250, 500]$; $w_3 =]500, 1250]$; $w_4 =]1250, d_{max}]$. For the second option W is $w_1 = [0, 400]$; $w_2 =]400, 1000]$; $w_3 =]1000, 10000]$; $w_4 =]10000, d_{max}]$. Over the small distance ranges, a relevant role of the spatial structure in determining the data heterogeneity can be appreciated, that decreases for larger distances. In particular, focusing on the first distance class of the first option $[0, 250]$, i.e. pixels' contiguity, and on the relative version of $PI(H|w_k)$, we quantify a 24.7% of entropy due to the spatial correlation. If we move to the first class of the second option $[0,400]$ this percentage is 22.4%. The higher these values, the closer we get to maximum urban compactness. When medium distance ranges are considered, e.g. $]1000,10000]$ or $]500,1250]$, the relative spatial partial information terms become small, denoting the possibility of a future urban expansion in the metropolitan area. As a conclusion, this set of entropy measures reveals as altogether very exhaustive in explaining the urban structure of Turin's metropolitan area, which is characterized by a substantive and non-negligible compactness of the urban patches at distances up to 1/1.25 kilometres. Once more, the package `SpatEntropy` efficiently supported the analysis in a rapid and suitable way. The potential and applicability of this study on urbanization may be highlighted by works that compare entropy measures across urban areas over a region or country, so that results can be commented both in absolute and relative terms (Altieri and Cocchi, 2021b).

6 Concluding remarks

In this work, the `R` package `SpatEntropy` is first introduced as the unique and efficient statistical tool to compute a large set of spatial entropy measures. After a review of the measures available in the literature, `SpatEntropy` is applied to real data from biodiversity and environmental fields. In particular, a first dataset on rainforest trees of Barro Colorado Island has been considered in order to quantify the biodiversity of trees' species over space. The spatial entropy indices are then employed for measuring the urban sprawl of Turin's metropolitan area. The potential of the package we propose is shown on these examples in terms of computational efficiency and simplicity of use. These features make `SpatEntropy` a suitable toolkit to be exploited also by non-statisticians and new users.

The package `SpatEntropy` is able to easily deal with different kinds of variables and of spatial data: the rainforest tree example refers to point data over an observation window with respect to a multicategorical variable (the 4 species), while urbanization data on Turin are areal, i.e. it considers a grid of pixels is considered, and a binary variable (urbanized/non-urbanized) characterizes each pixel. The ability of the package in managing space comes from the integration and exploitation of the widely known R package spatstat, which also offers the possibility of experimenting SpatEntropy on additional available dataset. Point data are more precise than areal ones and this implies different ways of defining the neighbourhood: we show how to fix distances for grids, which are automatically built by SpatEntropy using pixel centroids. Multicategorical and dichotomic variables can be equally analyzed for both point and areal data.

The different spatial entropy measures can be easily built by the SpatEntropy package and then carefully compared as shown in the two applications' Sections. We focus on the flexible and complete perspective offered by Altieri's set of spatial entropy measures. Despite the approach of SpatEntropy implies a larger, but still feasible, computational effort, results are very exhaustive and informative of the phenomenon, at both global and local levels. In particular, we showed how for the rainforest tree data the leading quantities are the partial residual terms, which increase when biodiversity of species rises at appropriate distance ranges, i.e. by considering space. In particular, we observed a percentage equal to 61% of residual entropy at the smallest distance range (up to 1 metre), denoting a remarkable biodiversity level even when considering small portions of the whole observation area. Conversely, for land cover data on Turin's urbanization, the most informative quantities are the spatial partial terms measuring the degree of compactness of the metropolitan area, which is a desirable feature for an efficient and sustainable development: the city of Turin is shown to have a compact urban structure. The choice of the distance ranges depends on the research aim; different options might be easily and quickly tested thanks to the package flexibility. Results can be enriched by a comparison to the findings on similar datasets (e.g., other tree species or different cities), exploiting the corresponding versions of all indices provided by the package.

Other fields of applications cover any spatial dataset where syntethic measures of heterogeneity may be of interest. They span over natural phenomena such as earthquakes (Altieri et al., 2016) and wildfires (Leuenberger et al., 2018; Gelfand and Monteiro, 2014), polluting agents (Cameletti et al., 2014), meteorological events (De Caceres et al., 2018) and epidemiological data (Blangiardo et al., 2016; Greco and Trivisano, 2009). Several dataset on these contexts of application are also available in the SpatEntropy package as additional toy examples.

References

Altieri, L. and Cocchi, D. (2021a). Spatial sampling for non-compact patterns. *International Statistical Review* Online Version before inclusion in an issue, doi:https://doi.org/10.1111/insr.12445.

Altieri, L. and Cocchi, D. (2021b). Understanding the expansion of italian metropolitan areas: A study based on entropy measures. *Environment and Planning B: Urban Analytics and City Science* Online Version before inclusion in an issue, doi:https://doi.org/10.1177/23998083211012699.

Altieri, L., Cocchi, D., Greco, F., Illian, J. and Scott, E. (2016). Bayesian p-splines and advanced computing in r for a changepoint analysis on spatio-temporal point processes. *Journal of Statistical Computation and Simulation* 86: 2531–2545, doi: https://doi.org/10.1080/00949655.2016.1146280.

Altieri, L., Cocchi, D. and Roli, G. (2018a). A new approach to spatial entropy measures. *Environmental and Ecological Statistics* 25: 95–110, doi:https://doi.org/10.1007/s10651-017-0383-1.

Altieri, L., Cocchi, D. and Roli, G. (2018b). SpatEntropy: Spatial Entropy Measures. R package version 0.1.0.

Altieri, L., Cocchi, D. and Roli, G. (2019a). Advances in spatial entropy measures. *Stochastic Environmental Research and Risk Assessment* 33: 1223–1240, doi:https://doi.org/10.1007/s00477-019-01686-y.

Altieri, L., Cocchi, D. and Roli, G. (2019b). Measuring heterogeneity in urban expansion via spatial entropy. *Environmetrics* 30: 1–16, doi:https://doi.org/10.1002/env.2548.

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis* 27: 94–115, doi:https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

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

Batty, M. (1974). Spatial entropy. *Geographical Analysis* 6: 1–31, doi: https://doi.org/10.1111/j.1538-4632.1974.tb01014.x.

Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis* 8: 1–21, doi: https://doi.org/10.1111/j.1538-4632.1976.tb00525.x.

Batty, M. (2010). Space, scale, and scaling in entropy maximizing. *Geographical Analysis* 42: 395–421, doi:https://doi.org/10.1111/j.1538-4632.2010.00800.x.

Batty, M., Morphet, R., Masucci, P. and Stanilov, K. (2014). Entropy, complexity, and spatial information. *Journal of Geographical Systems* 16: 363–385, doi:https://doi.org/10.1007/s10109-014-0202-2.

Blangiardo, M., Finazzi, F. and Cameletti, M. (2016). Two-stage bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology* 18: 1–12, doi:https://doi.org/10.1016/j.sste.2016.03.001.

Cameletti, M., Lindgren., F., Simpson, D. and Rue, H. (2014). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis* 97: 109–131, doi:https://doi.org/10.1007/s10182-012-0196-3.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory. Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Cressie, N. (1993). *Statistics for Spatial Data. Revised Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

De Caceres, M., Martin-StPaul, N., Turco, M., Cabon, A. and Granda, V. (2018). Estimating daily meteorological data and downscaling climate models over landscapes. *Environmental Modelling & Software* 108: 186–196, doi:https://doi.org/10.1016/j.envsoft.2018.08.003.

Drechsler, M. (2020). Model-based integration of ecology and socio-economics for the management of biodiversity and ecosystem services: State of the art, diversity and current trends. *Environmental Modelling & Software* 134: 104892, doi:https://doi.org/10.1016/j.envsoft.2020.104892.

EEA (2011). Analysing and Managing Urban Growth. Tech. rep., Environmental European Agency, also available as http://www.eea.europa.eu/articles/analysing-and-managing-urban-growth.

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K. and Caley, M. J. (2012). A software tool for elicitation of expert knowledge about species richness or similar counts. *Environmental Modelling & Software* 30: 1–14, doi:https://doi.org/10.1016/j.envsoft.2011.11.011.

Frosini, B. V. (2004). *Descriptive Measures of Ecological Diversity*. Paris, France. http://www.eolss.net: In Environmetrics. Edited by J. Jureckova, A. H. El-Shaarawi in Encyclopedia of Life Support Systems (EOLSS), revised edn. 2006. Developed under the Auspices of the UNESCO, Eolss Publishers.

Gelfand, A. E. and Monteiro, J. V. D. (2014). Explaining return times for wildfires. *Journal of Statistical Theory and Practice* 8: 534–545, doi:https://doi.org/10.1080/15598608.2013.821047.

Greco, F. and Trivisano, C. (2009). A multivariate car model for improving the estimation of relative risks. *Statistics in Medicine* 28: 1707–1724, doi:https://doi.org/10.1002/sim.3577.

Hoskins, A. J., Harwood, T. D., Ware, C., Williams, K. J., Perry, J. J., Ota, N., Croft, J. R., Yeates, D. K., Jetz, W., Golebiewski, M., Purvis, A., Robertson, T. and Ferrier, S. (2020). Bilbi: Supporting global biodiversity assessment through high-resolution macroecological modelling. *Environmental Modelling & Software* 132: 104806, doi:https://doi.org/10.1016/j.envsoft.2020.104806.

Karlström, A. and Ceccato, V. (2002). A new information theoretical measure of global and local spatial association. *The Review of Regional Research (Jahrbuch für Regionalwissenschaft)* 22: 13–40, available as https://mpra.ub.uni-muenchen.de/6848/.

Leibovici, D. G. (2009). *Defining Spatial Entropy from Multivariate Distributions of Co-occurrences*. Berlin, Springer: In K. S. Hornsby et al. (eds.): COSIT 2009, Lecture Notes in Computer Science 5756, pp 392-404.

Leibovici, D. G., Claramunt, C., LeGuyader, D. and Brosset, D. (2014). Local and global spatio-temporal entropy indices based on distance ratios and co-occurrences distributions. *International Journal of Geographical Information Science* 28: 1061–1084, doi: https://doi.org/10.1080/13658816.2013.871284.

Leinster, T. and Cobbold, C. (2012). Measuring diversity: the importance of species similarity. *Ecology* 93(3): 477–489, doi:https://doi.org/10.1890/10-2402.1.

Leuenberger, M., Parente, J., Tonini, M., Pereira, M. G. and Kanevski, M. (2018). Wildfire susceptibility mapping: Deterministic vs. stochastic approaches. *Environmental Modelling & Software* 101: 194–203, doi:https://doi.org/10.1016/j.envsoft.2017.12.019.

Li, H. and Reynolds, J. F. (1993). A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 8: 155–162, doi:https://doi.org/10.1007/BF00125347.

O'Neill, R. V., Krummel, J. R., Gardner, R. H., Sugihara, G., Jackson, B., DeAngelis, D. L., Milne, B. T., Turner, M. G., Zygmunt, B., Christensen, S. W., Dale, V. H. and Graham, R. L. (1988). Indices of landscape pattern. *Landscape Ecology* 1: 153–162, doi: https://doi.org/10.1007/BF00162741.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation* 15: 1191–1254, doi:https://doi.org/10.1162/089976603321780272.

Parresol, B. R. and Edwards, L. A. (2014). An entropy-based contagion index and its sampling properties for landscape analysis. *Entropy* 16: 1842–1859, doi:https://doi.org/10.3390/e16041842.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rényi, A. (1961). *On measures of entropy and information*. Berkeley: In 4th Berkeley Symp. Math. Statist. Prob., vol. I, Univ. Calif. Press., pp 547-561.

Scarnati, L., Attorre, F., Farcomeni, A., Francesconi, F. and Sanctis, M. D. (2009). Modelling the spatial distribution of tree species with fragmented populations from abundance data. *Community Ecology* 10: 215–224, doi:https://doi.org/10.1556/comec.10.2009.2.12.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423 and 27(4):623–656, doi:https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Theil, H. (1972). *Statistical Decomposition Analysis: with applications in the social and administrative sciences*. Amsterdam: North-Holland Publishing Company.