Revue économique
https://doi.org/10.3917/reco.703.0305 **(French published version)**

Measuring Biodiversity and Spatial Structuring of Economic Activity Through Entropy

Eric Marcon[1*]

Abstract
Measures of spatial concentration and specialization in economics are very similar to those of biodiversity and species valence in ecology. Methodological developments are more advanced in ecology, which motivates this interdisciplinary transfer work. Entropy is the fundamental notion, derived from statistical physics and information theory, used to measure concentration and specialization. The notion of effective number, which is a number of categories in a simplified ideal distribution, is introduced. The decomposition of the total diversity of a distribution (the global location in economics) into absolute and relative concentration or specialization and replication is presented. These methods provide a comprehensive and robust theoretical framework for measuring spatial structuring in discrete space.

Keywords
entropy, Hill numbers, diversity, valence, concentration, specialization

[1]UMR EcoFoG, AgroParistech, CNRS, Cirad, INRA, Université des Antilles, Université de Guyane.
Campus Agronomique, 97310 Kourou, France.
***Corresponding author**: eric.marcon@agroparistech.fr, https://ericmarcon.github.io/

Contents

1. Introduction

Research on the spatial structure of economic activity has been mainly concerned with spatial concentration as a source of positive externalities (Marshall, 1890; Weber, 1909; Krugman, 1991), which goes hand in hand with specialization (Houdebine, 1999; Cutrini, 2010). Numerous measures of spatial concentration applicable to discrete data (as opposed to continuous-space measures discussed for example by Marcon and Puech, 2017) have been developed (for a review, see for example Combes and Overman, 2004) but a comprehensive methodological framework linking concentration, specialization, and inequality measures in general is lacking, although it has been sketched several times (Brülhart and Traeger, 2005; Mori et al., 2005; Bickenbach and Bode, 2008; Cutrini, 2010).

In parallel, the measurement of biological diversity, which has become biodiversity (Wilson and Peter, 1988) and, to a lesser extent, of the valence of species (Levins, 1968) have been the subject of an abundant literature in statistical ecology (Pielou, 1975; Patil and Taillie, 1982; Magurran, 1988, etc). It was largely inspired by information theory (Shannon, 1948) and statistical physics (Dewar and Porté, 2008). Entropy-based measures of diversity are the state of the art in ecology (Marcon, 2017).

In economics, Theil (1967) proposed measures of inequality and spatial concentration similar to Shannon's entropy, but subsequent methodological advances have lagged behind those in biodiversity measurement. The objective of this paper is to transfer the latest developments in biodiversity measurement to the discipline of geographical economics to complement its definitions of spatial concentration and specialization. The numerous borrowings of methods between distant disciplines will be highlighted.

Entropy and its properties will be presented in the next section. The application of these methods to the measurement of spatial concentration and specialization will follow, before a final synthesis section devoted to joint diversity, a framework allowing the complete decomposition of location measures.

2. Methods

2.1 Similar issues and opposing concepts

The methods presented here have been developed from the biodiversity literature. Ecologists need to measure the *diversity* of a community of living things, com-

posed of several species whose numbers are known. A less discussed question concerns the *valence* of species, i.e. for a given species the diversity of environments in which it is able to establish itself.

This notion was formalized under the name of niche width by Levins (1968), in the sense that the ecological niche is the set of conditions necessary for the development and reproduction of a living being. To fix ideas and without loss of generality, the examples treated here will concern trees in a forest. Each tree belongs to one and only one species, and the number of individuals of each species is known. The species are located in a taxonomy: they are grouped by genera and the genera by families. Finally, the forest is divided geographically into plots, which in turn are divided into sub-plots. We will limit ourselves here to the simplest measures of diversity, not taking into account the greater or lesser differences between species, taxonomic for example.

In economic geography, the question probably most dealt with is that of *spatial concentration* (Ottaviano and Puga, 1998; Combes and Gobillon, 2015), a source of positive externalities (Baldwin and Martin, 2004). It is very similar to the species valence of ecologists, but opposite: high concentration means low valence[1]. Specialization* (Amiti, 1997) is similarly the opposite notion of diversity. The examples treated here in economics will concern industrial establishments in European countries provided by the freely accessible EuroStat database. The establishments have a number of employees, which allows them to be weighted. They belong to a sector of activity, here according to the NUTS nomenclature, which is a taxonomy similar to that of biological species, and their location by country can be detailed by regions (according to the NACE nomenclature) and their subdivisions.

Specialization and spatial concentration (Cutrini, 2010), like diversity and valence (Gregorius, 2010) are mathematically related: the existence of highly concentrated sectors implies the existence of regions specialized in that sector. A synthetic approach can be developed: Cutrini (2010) defines the "global localization" for this purpose, which will be generalized.

2.2 Data and notations

The data have been chosen for their accessibility and simplicity: the aim here is to present methods rather than to deal in detail with complex economic issues. The applications will be based on the numbers of employees per industrial sector in 25 European countries in 2015. The data are available online at EuroStat [2], in the file *SBS data by NUTS 2 regions and NACE Rev. 2*.

The nomenclature of economic sectors is the NACE (Statistical Classification of Economic Activities in the European Communities) in its revision 2. Only the industrial sectors (NACE code: C) have been retained. Sectors C12 (Manufacture of tobacco products), C19 (Manufacture of coke and refined petroleum products), C21 (Manufacture of basic pharmaceutical products and pharmaceutical preparations) and C30 (Manufacture of other transport equipment) were removed because of missing data in major countries (e.g. C30 in Belgium).

Of the 30 countries available, Cyprus, Malta, Ireland, Luxembourg and Slovenia were removed because they had too much missing data. The selection of the data is therefore a compromise to keep the essential information, which is debatable but sufficient for the methodological demonstration of this article.

After filtering, the data are presented in the form of a table (called a contingency table) whose 19 rows are the industrial sectors and 25 columns the selected countries. Each cell of the table contains the number of employees in the sector and the country considered, without missing data.

The sectors are indicated by the letter s and the countries by the letter i. The number of employees per sector and country is noted $n_{s,i}$. The marginal values are noted n_i (the number of employees in country i, all sectors combined) and n_s (the number in sector s, all countries combined). To simplify the writing, the level of aggregation corresponding to all sectors will be called "the industry" and that corresponding to all countries will be called "Europe": n_s will thus be called the number of employees working in the s sector in Europe. The total number of employees is $n = \sum_s n_s = \sum_i n_i$, equal to 27,419,407. The relative sizes of the countries and sectors are shown in the Appendix. The probability that a randomly selected individual works in sector s and country i is denoted $p_{s,i}$ and estimated by his or her observed frequency $p_{s,i} = n_{s,i}/n$ (to lighten the notation, the empirical frequency is denoted as the theoretical probability rather than $\hat{p}_{s,i}$). The marginal probabilities are denoted p_s and p_i; they are estimated by n_s/n and n_i/n respectively. Finally, the probabilities will also be considered by sector or by region: $p_{s|i} = p_{s,i}/p_i$ is the probability for a person from country i to work in sector s. The sum of these probabilities is 1 for each sector or region: $\sum_s p_{s|i} = \sum_i p_{i|s} = 1$.

The vector of probabilities $p_{s|i}$ of all sectors in country i is denoted $\mathbf{p_{s|i}}$. Similarly, $p_{i|s}$ is the probability, in the chosen sector s, that a person works in country i and $\mathbf{p_{i|s}}$ is the vector of country probabilities for sector s. The probability matrix whose elements are $p_{s,i}$ is denoted $\mathbf{P}$.

The data and the R code (R Core Team, 2018) needed to reproduce the full results are in the appendix. The code makes extensive use of the *entropart* package (Marcon and Hérault, 2015b) dedicated to biodiversity measurement.

2.3 Entropy as a measure of uncertainty

The notions being established, the task is now to translate them into operational measures allowing to com-

[1] A species of low valence is said to be *specialized* in ecology, but this vocabulary is not used here to avoid confusion with regional specialization.

[2] http://ec.europa.eu/eurostat/web/regions/data/database

pare the diversity of different biological communities (such as the trees of a forest, without the example being limiting) or the specialization of industrial regions, to give a concrete, easily understandable meaning to these measures, and to characterize their properties in order to be able to use them for example in models.

Biological diversity is an important determinant of ecosystem functioning (Chapin et al., 2000). Among many measures developed according to needs (Peet, 1974), the interest of Shannon (1948)'s entropy has been argued in particular by Pielou (1975) in a reference book. In econometrics, the work of Davis (1941) and especially Theil (1967) opened the way. Theil's well-known index is the difference between Shannon's entropy and its maximum possible value, which illustrates the opposition of the approaches presented above as well as the convergence of the methods.

Entropy is, among other things, a measure of uncertainty that it is time to formalize. Let us define an experiment (for example the sampling of a tree at random in a forest) whose set of possible outcomes (the species to which it belongs) is known. The results are noted r_s where the index s takes all possible values between 1 and S, the number of possible results. The probability of obtaining r_s is p_s, and $\mathbf{p_s} = (p_1, p_2, \ldots, p_S)$ is the set (mathematically, the vector) of probabilities of obtaining each result. Obtaining the result r_s is not very surprising if p_s is large: it brings little additional information compared to the simple knowledge of probabilities. On the other hand, if the species r_s is rare (p_s is small), its result is surprising. The notion of information, defined by Shannon, is identical to that of surprise, more intuitive. We therefore define an information function, $I(p_s)$, decreasing as the probability increases, from $I(0) = +\infty$ (or possibly a strictly positive finite value) to $I(1) = 0$ (the observation of a certain result brings no surprise).

Entropy is defined as the average of the information provided by all possible outcomes of the experiment. As each outcome has probability p_s to be realized, the average over all possible outcomes is the weighted average of $I(p_s)$. Entropy is defined as

$$H(\mathbf{p_s}) = \sum_s p_s I(p_s).$$

2.3.1 Shannon's entropy

Shannon used the information function $I(p_s) = -\ln p_s$ for its mathematical properties. It can be written as $I(p_s) = -\ln(1/p_s)$. The inverse of the probability, $1/p_s$, will be called *rare*[3]: a very rare species has a probability close to 0.

The information function used by Shannon is therefore the logarithm of rarity.

The term "entropy" was introduced by Clausius in 1865 (*Memoir IX* in Clausius, 1868) for his new formulation of the second principle of thermodynamics stated by Carnot (1824). Its Greek etymology means *transformation* because the second principle concerns the variation of entropy. Boltzmann characterized the entropy of a complex system (a gas, whose each particle can have several possible states) in 1877 (Sharp and Matschinsky, 2015). Shannon (1948) finally showed that the number of possible states of a system is analogous to the number of messages of a chosen length that can be created by assembling the letters of an alphabet whose letter frequencies are fixed. Shannon's entropy is, except for a multiplicative constant, equal to Boltzmann's entropy normalized by the length of the message, of which it is independent. This fundamental property allows him to describe the complexity of a system not only by the possible number of its states, but more simply by the relative frequency of its components, giving birth to the theory of information.

The relevance of entropy as a measure of diversity follows directly from this: a system is all the more diverse that it can have a large number of possible states or, equivalently, that it is difficult to predict the state in which it is, or that it has a high entropy.

2.3.2 Generalized entropy

Many alternative information functions can be considered, including the most exotic ones like $I(p_s) = \cos(p_s\pi/2)$ (Gregorius, 2014).

Among them, three families of parametric functions have become established: the generalized entropy of the inequality literature (Shorrocks, 1980), the entropy of Rényi (1961), which was widely used until the 2000s for the measurement of biodiversity, and, more recently, the HCDT entropy detailed here.

Tsallis (1988) proposed a generalized entropy in statistical physics for systems that do not meet the properties required by Boltzmann theory. It had been defined by Havrda and Charvát (1967) in cybernetics and then rediscovered, notably by Daróczy (1970) in information theory, hence its name, *HCDT* entropy (see Mendes et al. (2008), page 451, for a complete history).

Its mathematical form is:

$${}^qH(\mathbf{p_s}) = \frac{1}{q-1}\left(1 - \sum_{s=1}^{S} p_s^q\right),$$

where q is an arbitrary parameter. When $q = 1$, the formula does not apply but the limit of ${}^qH(\mathbf{p_s})$ when $q \to 1$ is the Shannon entropy, which is thus retained as a definition of ${}^1H(\mathbf{p_s})$.

Its interest appears more clearly by defining a generalization of the logarithm function, the deformed logarithm of order q (Tsallis, 1994) as

$$\ln_q x = \frac{x^{1-q} - 1}{1-q}.$$

Again, $n_q x$ tends to the natural logarithm when q tends to 1. The HCDT entropy is then written as a generalization of the Shannon entropy:

$${}^qH(\mathbf{p_s}) = \sum_s p_s \ln_q(1/p_s)$$

[3] Patil and Taillie (1982) use the term *rarity* in the sense of *information*, but this definition has not been adopted in the later literature.

The deformed logarithm is a function which, as its name indicates, deforms the natural logarithm function by changing its curvature but respecting, whatever q is, $\ln_q 1 = 0$ and the limit $+\infty$ when $x \to \infty$. Its value at $x = 0$ is negative but finite for $q < 1$. For $q \geq 1$, this is not the case: $\ln_q x \to -\infty$ when $x \to 0$. By varying the parameter q around 1, the information function $\ln_q(1/p_s)$ attributes respectively a greater or lesser surprise to the rare species (i.e. whose rarity, $1/p_s$, is great) when q increases or decreases.

At this point, we have a simple and general definition: the entropy (of order q) of a system is the average surprise brought by the observation of one of its individuals; the surprise is the logarithm (of order q) of the rarity. A biological community is all the more diverse as it is surprising (i.e. as its entropy is large). A region is all the more specialized as its entropy is low.

Three values of q are particularly interesting:

- $q = 0$: entropy is richness, i.e., S, the number of species or sectors, minus 1;
- $q = 1$: the entropy is Shannon's entropy. In econometrics, $S - 1 H$ is the Theil index;
- $q = 2$: the entropy is Simpson (1949)'s biodiversity index, i.e. the probability that two randomly chosen individuals belong to a different species. In econometrics, its complement to 1, i.e. the probability that two individuals belong to the same sector, is the Herfindahl index, or Herfindahl-Hirschman (Hirschman, 1964), which measures here specialization.

Negative values of q give a species a greater importance the rarer it is, whereas at $q = 0$ all species contribute equally to the entropy (they are simply counted, whatever their probability). The interest of these values is therefore limited. Since their mathematical properties are poor (Marcon et al., 2014), they are in practice not used. Values of q greater than 2 are little used because they neglect too much the species which are not the most frequent.

2.3.3 From entropy to diversity

Entropy has a physical meaning: it is a quantity of surprise; it is thus much more than an index, which is only an arbitrary value that must respect an order relation to allow comparisons. However, with the exception of orders 0 and 2, the value of entropy has no intuitive interpretation. Hill numbers address this lack.

The desire of Hill (1973) was to make diversity indices intelligible after Hurlbert (1971)'s noted paper entitled "The non-concept of specific diversity". Hurlbert criticized the diversity literature for being too abstract and remote from biological realities, in particular by providing examples in which the order of communities was not the same according to the diversity index chosen.

The number of Hill of order q is the number of equiprobable species giving the same value of entropy as the observed distribution, in other words an *effective number* of species, also called *number equivalent*. The concept was rigorously defined by Gregorius (1991), after Wright (1931) who had first defined the effective size of a population in genetics: given a characteristic variable (here, entropy) depending only on a numerical variable (here, the number of species) in an ideal case (here, equiprobability of species), the effective number is the value of the numerical variable for which the characteristic variable is that of the data set.

Formally, they are simply the deformed exponential of the HCDT entropy (Marcon et al., 2014). The deformed exponential function of order q is the reciprocal function of the deformed logarithm, whose value is

$$e_q^x = [1 + (1-q)x]^{1/(1-q)}.$$

The Hill number of order q, simply called *diversity of order q* (Jost, 2006) is thus

$${}^qD(\mathbf{p_s}) = e_q^{{}^qH(\mathbf{p_s})}.$$

The explicit formulation from the probabilities is:

$${}^qD(\mathbf{p_s}) = \left(\sum_s p_s^q\right)^{1/(1-q)}.$$

These results had already been obtained with another approach by MacArthur (1965) and taken up by Adelman (1969) in the economic literature. Also, the inequality measure of Atkinson (1970) is very similar to Hill numbers.

The rigorous use of vocabulary removes any ambiguity (Jost, 2006): the terms diversity, specialization, valence and concentration will be reserved for Hill numbers, and they will not be used for entropy (which may be called *index* of diversity or concentration).

2.3.4 Diversity profiles

Since diversity is expressed in the same unit (a number of species) whatever its order, it is possible to plot a diversity profile, that is the value of ${}^qD$ as a function of q. The curves of two communities may cross each other because the weight of rare species decreases with increasing q. If this is not the case, the order relationship between the communities is well defined (Tothmeresz, 1995).

2.4 The decomposition of entropy

The notion of diversity β was introduced by Whittaker (1960) as the degree of differentiation of biological communities. The question addressed here is the decomposition of diversity from aggregate data (the diversity of economic sectors in Europe) to a more detailed level (by country). The diversity of the most aggregated level was called γ by Whittaker, the average diversity of the detailed levels α, and the differentiation between the detailed levels β. It is clear that the γ and α diversities are similar in nature: only the

level of detail of the data differs. In contrast, the characterization of β diversity has generated controversy (Ellison, 2010).

In economics, the decomposition of inequality measures has followed a parallel path to that of ecologists (Bourguignon, 1979). That of spatial concentration has remained limited to Theil's entropy (Mori et al., 2005; Cutrini, 2010) with the notable exception of Brülhart and Traeger (2005) who used the generalized entropy of Shorrocks (1980).

Jost (2007) showed that the decomposition of entropy is additive: β entropy is the difference between the entropies γ and α. Marcon et al. (2012) then interpreted the β entropy as the additional information brought by the knowledge of the disaggregated distributions ($\mathbf{p_{s|i}}$ for each country i) in addition to that of the aggregated data ($\mathbf{p_s}$ for the whole of Europe), i.e. a relative entropy. The divergence of Kullback and Leibler (1951) is well known to economists as Theil's relative entropy (Conceição and Ferreira, 2000). The difference between the first-order *gamma* entropy and the average of the first-order entropies of the disaggregated distributions is the average of the corresponding Kullback-Leibler divergences (Rao and Nayak, 1985), referred to by statistical physicists as the "Jensen-Shannon divergence". Marcon et al. (2014) generalized this result to all orders of the HCDT entropy: β entropy is the average over all countries of the generalized Kullback-Leibler divergence between the $\mathbf{p_{s|i}}$ and $\mathbf{p_s}$ distributions, itself defined as the average over all sectors of the information gain brought by knowledge of the disaggregated distribution:

$$ {}^{q}_{\beta}H(\mathbf{P}) = \sum_i p_i \sum_s p_{s|i}[\ln_q(1/p_{s|i}) - \ln_q(1/p_s)] $$

.

Like γ and α entropy, β entropy can be transformed into an effective number which is the number of communities of the same weight, with no common species, that would have the same β entropy as the actual communities. The diversity decomposition is multiplicative: γ diversity is the product of α and β diversities.

The complete decomposition is finally a product of effective numbers: the diversity of the assembly of several biological communities, called γ diversity is an effective number of species; it is the product of the effective number of species per community (α diversity) by the effective number of communities (β diversity). It will be applied in this article to the economy of the European countries: the effective number of economic sectors of Europe (γ) is the product of the average effective number of sectors per country (α) by an effective number of countries (β).

Similarly, the valence of an economic sector at an aggregate level (manufacturing industry) is an effective number of countries (γ), which can be decomposed into an effective number of countries per sector at a less aggregate level (α) multiplied by an effective number of sectors (β).

The decomposition will be limited here to a single level of disaggregation of the data. It can be repeated: countries can be broken down into regions, regions into counties… The effective number of economic sectors in Europe (γ) can then be decomposed into an effective number of countries (β_1) times an effective number of regions (β_2) times an effective number of counties (β_3) times an effective number of sectors per county (α). The hierarchical decomposition of diversity has been addressed by Marcon et al. (2012); Richard-Hansen et al. (2015); Pavoine et al. (2016), among others.

2.5 Joint diversity: mutual information and replication

We have seen that entropy can be used from both diversity and valence perspectives (equivalently: specialization and spatial concentration). The data are the same and can be represented in the contingency table whose rows represent, for example, the industrial sectors while the columns represent the countries, each cell of the table providing the abundance (a number of establishments or employees) of a sector in a country.

The diversity of countries is calculated by treating each column of the table, the valence of sectors by treating each row. The diversity ${}^{q}D(\mathbf{p_s})$ of the whole of Europe (defined as the aggregation of the countries) is obtained, like the valence of the aggregated sectors ${}^{q}D(\mathbf{p_i})$, from the marginal probabilities The diversity ${}^{q}D(\mathbf{p_{s,i}})$ of the data set, all sectors and countries combined, is of great interest, especially theoretically for the Shannon entropy (Faddeev, 1956; Baez et al., 2011): it is called joint diversity (Gregorius, 2010).

The difference between the joint entropy and the sum of the marginal entropies (that of all sectors and that of all countries), ${}^{q}H(\mathbf{p_{s,i}}) - {}^{q}H(\mathbf{p_s}) - {}^{q}H(\mathbf{p_i})$, is called the mutual information. The Shannon entropy (but not the HCDT entropy of order different from 1) of two independent systems adds up: if country membership is independent of sector membership, i.e. if the probability $p_{s,i}$ is simply the product of the probabilities p_s and p_i, then the mutual Shannon information is zero. In other words, the mutual information is the additional entropy brought by the non-independence of the rows and columns of the array. It is equal to the two entropies β, the one of the diversity and the one of the valence. These properties are only valid for the Shannon entropy. They have been used in different forms in the literature (for example Cutrini, 2009; Chao et al., 2013; Haedo and Mouchart, 2017).

Regardless of the order considered, Gregorius (2010) showed that joint diversity provides important additional information about the distribution of abundances that is not captured by the diversity decomposition already presented. The example of biodiversity is used here to simplify the presentation. The *alpha* diversity is the number of equiprobable species in a typical community. The β diversity is the number of these typical communities, equiprobable and without common species. The γ diversity is the product of the two previous ones, a number of equiprobable species resulting

from the assembly of the communities. Each species appears only in one community in this representation. The identical replication of the communities does not modify the diversities α, β and γ, it is actually a property required for diversity measures (Hill, 1973). In contrast, joint diversity is multiplied by the number of replications (Marcon, 2017): the ratio of joint diversity to β diversity measures replication as an effective number, the number of replications of communities.

Replication has few practical applications in ecology because the available data are typically samples of the communities being studied. Their replication reflects the sampling effort, which is a choice of the experimenter. When the data are exhaustive or, more generally, when the marginal probabilities of communities are interpretable as their sizes, replication is as important an information as diversity.

3. Spatial concentration and specialization

The methods presented so far, from physics and statistical ecology, have interesting applications in economics. Two questions will be addressed: the measurement of the spatial concentration of economic activities and the decomposition of joint diversity.

The spatial concentration of economic activities is an important topic in the literature (Combes and Gobillon, 2015). The first step in understanding the economic phenomena involved is the characterization of the concentration. A major step was taken by Ellison and Glaeser (1997) who clearly laid down the principle of a relative measure (the geographical distribution of an industrial sector is compared to that of the size of the regions where it is considered, a principle summarized under the title of the dartboard approach) and that of the statistical test of the observed distribution against its value under an appropriate null hypothesis: a uniform and independent distribution. These features are lacking in earlier concentration indices, such as the Gini index (Gini, 1912; Ceriani and Verme, 2012), whose observed value can only be compared to its possible extremes.

The central statistic of the Ellison and Glaeser index for sector s is, with our notations, $G_s = \sum_i (p_{i|s} - p_i)^2$, i.e., the sum of the squares of the deviations between country i's share of the total workforce in sector s and country i's share of the industry, all sectors combined. In mathematical terms, G is the distance L^2 between the observed distribution of the s sector and its expected distribution (Haedo and Mouchart, 2017), that of industry in general.

The relative index of Theil (1967) is sometimes used for the same purpose (Cutrini, 2009): it also measures the gap between the observed and the expected distribution, but with another metric: the Kullback-Leibler divergence.

The HCDT entropy allows to unify and extend these approaches. The concentration, absolute and relative, will be considered first. Specialization will follow.

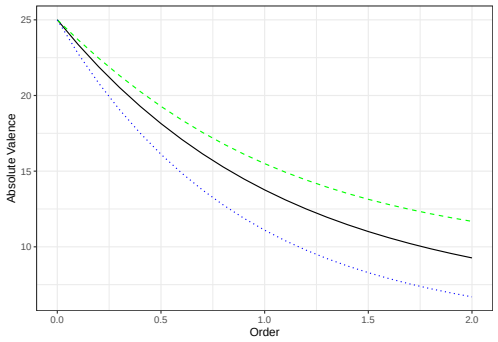

Figure 1. Absolute valence profiles of industry (solid curve, black), the C10 sector (long-dotted line, green) and the C20 sector (short-dotted line, blue)

3.1 Spatial concentration

3.1.1 Valence and absolute concentration

The valence of sector s, ${}^qD(\mathbf{p_{i|s}})$ is the effective number of countries it occupies. Valence can be calculated for any level of sectoral grouping, here for the entire industry (NACE code C) or by detailed sector. A valence profile can be plotted for each sector. At low orders, the valence gives high importance to countries with low occupancy. At $q = 0$, the valence is simply the number of countries in which the sector is present. At large orders, only majority occupied countries contribute to the valence.

Figure 1 shows the valence profiles of the entire industry and the C10 (Food Manufacturing) and C20 (Chemical Manufacturing) sectors that deviate the most from the entire industry of all sectors studied. Valence is measured in effective numbers of countries. All sectors are present in all countries (the data are highly aggregated) so the valence of order 0 is always equal to the maximum possible, 25. At order 2, at the other end of the curves, 9.3 countries occupied by the same number of employees would suffice to obtain the same level of valence as observed for the whole industry.

Valence is the opposite concept to concentration. A simple transformation of valence values allows them to be translated into concentration levels that are more in line with the economic culture. The complement of valence to the number of countries is a good measure of concentration as the actual number of countries left behind by the sector under study. It can be normalized by the number of countries minus 1 to obtain a value between 0 and 1 shown in Figure 2.

The value of the concentration is the proportion of countries left behind (in effective numbers). A value of 0 means that all countries are occupied, a value of 1 that the whole sector is concentrated in one country. The chemical industry, C20, is much more concentrated than industry in general, while the food industry, C10, is much less concentrated. These results are valid at all orders, except near 0, when the presence of

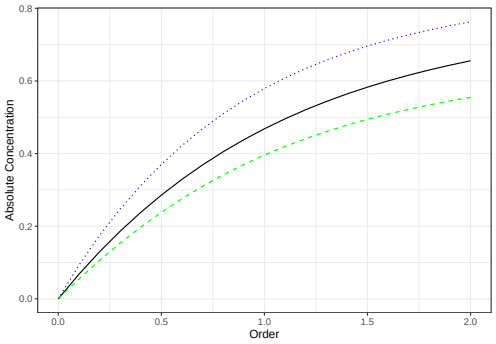

Figure 2. Absolute concentration profiles of industry (solid curve, black), sector C10 (long-dotted line, green) and sector C20 (short-dotted line, blue).

the sector alone counts, regardless of its abundance.

This measure of valence or concentration is absolute (Brülhart and Traeger, 2005): it does not compare the actual number of sectors to any external reference. For its interpretation, a comparison to another absolute measure (concentration at a more aggregated level) is necessary (Marcon and Puech, 2017).

3.1.2 Relative concentration

Absolute valence was calculated at the level of disaggregated sectors (C10 and C20) and at the level of the entire industry, whose numbers were obtained by aggregating those of the sectors. Using the terminology of the biodiversity decomposition, the absolute valence of the whole industry is the γ valence, equal to the product of the α valence (the average of that of the disaggregated sectors) and the β valence, the effective number of equiprobable sectors sharing no country.

The decomposition of entropy is the same, but the γ entropy is the sum of the entropies α and β. The β entropy is, as we have seen, the generalized Jensen-Shannon divergence between the distribution of each sector and the aggregate distribution of the industry. At particular orders $q = 1$ and $q = 2$, this divergence is the average, weighted by the size of sectors, of Theil's relative entropy and Ellison and Glaeser's G_s statistics. These classical indices of spatial concentration are generalized Kullback-Leibler divergences, in other words β entropies of particular orders, giving different importance to countries with small numbers: the Ellison and Glaeser index, of order 2, takes into account only the dominant establishments

The β entropy measures relative concentration, not relative valence: the α and β entropies have fundamentally different properties. It integrates a reference (the distribution of the industry for all sectors) and thus has an expected value, 0, if the distribution of the considered industry is identical to the reference one.

Its value cannot be interpreted simply: one must resort to the effective number of sectors, whose interpretation is intuitive.

In the framework of the decomposition presented above, the average relative concentration is the ratio of the γ valence to the α valence: it applies to all sectors but does not give information on a particular sector.

It must therefore be detailed for each sector: the relative concentration of sector s is defined as the ratio between the absolute valence of the whole industry (γ) and its own absolute valence:

$$ {}^qC_s = {}^qD(\mathbf{p_i}) / {}^qD(\mathbf{p_{i|s}}). $$

This is an effective number of sectors: if all sectors had a valence equal to the effective number of countries ${}^qD(\mathbf{p_{i|s}})$, it would take ${}^qC_s$ to obtain an industry with a valence of ${}^qD(\mathbf{p_i})$ effective countries.

The value of the relative concentration can be seen in figure 1: it is equal to the ratio between the values of the valence profiles of the industry and the sector considered. For the chemical industry (C20), it varies from 1 (at order 0) to 1.4 at order 2: 1.4 effective valence sectors that of the C20 sector, i.e. 6.7 countries, would form an industry whose valence would be that observed for the European industry, $6.7' = 9.3$ effective countries.

The concentration is less than 1 for the food industry (C10): $0.79 = 9.3 = 9.3 =$ 'r format(upCy[*which*(*upC*10x==2)]/u
In other words, the C10 sector is relatively dispersed.

Relative concentration and absolute valence (figure 1) are related: their product is the absolute valence of the whole of the sectors, taken as a reference. Absolute concentration (figure 2) thus goes hand in hand with relative concentration, but the information they provide is different.

In the economic literature, relative entropy has been used to measure relative concentration by Brülhart and Traeger (2005). Mori et al. (2005) used the Kullback-Leibler divergence between the distribution of a sector and that of the area (instead of the number of employees working in the industry) of regions in Japan to measure the topographic (not relative) concentration of sectors. Rysman and Greenstein (2005) proposed a test of the relative concentration of a sector based on the likelihood ratio of the distributions of the sector and the whole industry, which is simply the Kullback-Leibler divergence (see Mori et al., 2005, for a detailed presentation of the links between the two approaches). Alonso-Villar and Del Río (2013) proposed a generalized entropy decomposition but were limited in practice to order 1.

Theil's relative entropy has been used to compare the evolution of spatial concentration over time (e.g., Cutrini, 2010) since it does obey an order relation like any entropy. Finally, Bickenbach et al. (2013) have combined the Theil index (absolute) and the relative Theil index to better describe spatial concentration by applying them to different economic sectors (industry and services).

3.2 Specialization

The specialization measure works in exactly the same way as the concentration measure, swapping the role of the rows and columns of the contingency table.

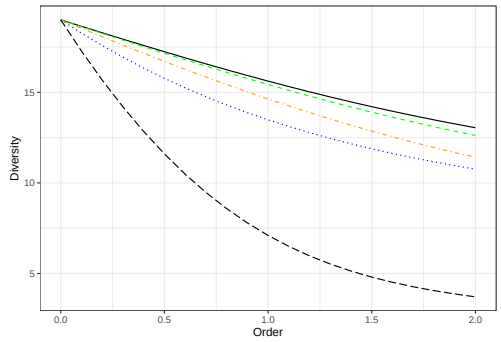

Figure 3. Diversity profiles of Europe (solid curve, black), Italy (long-dotted line, green), France (alternating dotted line, orange), Germany (short-dotted line, blue) and Iceland (very-long-dotted line, black).

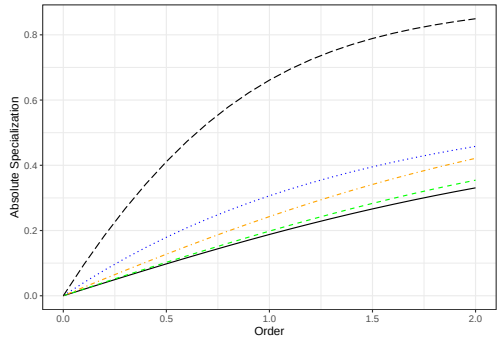

Figure 4. Absolute specialization profiles of Europe (solid curve, black), Italy (long-dotted line, green), France (alternating dotted line, orange), Germany (short-dotted line, blue) and Iceland (very-long-dotted line, black).

Figure 3 shows the absolute diversity profiles for Italy, Germany, France, and Iceland and for Europe. Diversity is the effective number of equiprobable sectors that would provide the same diversity as observed. As before, the level of aggregation of the data is such that all sectors are represented in all countries: richness, i.e. the diversity of order 0 is equal to the number of sectors. All countries are less diverse than Europe: they are therefore all specialized to varying degrees. Italy is not very specialized, Germany is more specialized than France, and Iceland is the most specialized country in Europe with less than 5 effective industrial sectors of order 2, three times less than Italy. The food industry, C10, employs almost half of the employees working in industry in Iceland.

Diversity can be transformed into absolute specialization, as valence was into concentration, to obtain the figure 4.

Finally, relative specialization is the ratio between the absolute diversity of the whole of Europe and that of each country, visible in figure 3. Iceland is the most relatively specialized country, with a value at order 2 of 3.5 effective countries.

3.3 Significance tests

Two approaches are possible to test the concentration or specialization profiles. For the sake of clarity, and without loss of generality, the aim here is to test the specialization of Italy against the null hypothesis that it would not be different from that of the whole of Europe.

The first formalization of the null hypothesis is that the distribution of industrial sectors in Italy is the same as that of the whole of Europe. The test then concerns the value of the generalized Kullback-Leibler divergence between the distribution of sectors in Italy and the distribution of sectors aggregated at European level. This is the approach, for spatial concentration, of Mori et al. (2005) at order 1. The average for all s sectors of Ellison and Glaeser's G_s statistic is equal to the average of the second-order divergences between the sector distribution and the industry-wide distribution (Marcon, 2017, section 12.4). The interpretation of the divergence is not robust (Jost, 2007): the statistic tested is the β entropy but in some cases its value is constrained by that of the α entropy regardless of the γ entropy.

The alternative formalization is that the specialization of Italy is equal to that of a country of the same size whose distribution of sectors would be that of the whole of Europe. The statistic tested is an effective number (absolute diversity or relative specialization, equivalently). This is the approach chosen here because it does not suffer from the problem of dependence of α and β entropies.

The test is performed by bootstrapping, i.e. by randomly generating a large number of new data corresponding to the null hypothesis and by computing the statistic of interest. The data are simulated by 1000 draws in a multinomial distribution whose parameters are the number of employees working in Italy and the probabilities of the sectors at European level. The specialization of each simulation is computed for orders from 0 to 2, by intervals of 0.1. The quantiles corresponding to 2.5% and 97.5% of the simulated specializations constitute the limits of the confidence envelope of the statistic under the null hypothesis, to which the real specialization is compared.

The details of the test are presented in the Appendix. The null hypothesis is not rejected at order 0: the specialization of Italy is identical to that of Europe since in both cases all sectors are present. From order 0.1, the null hypothesis is rejected: Italy is more specialized than Europe.

The variability of the simulated specialization is extremely low because the numbers are large and the employees are redistributed independently of each other by the multinomial distribution. For this reason, Mori et al. (2005), using similar data, choose to test the spatial concentration of establishments by ignoring their employment numbers. A much better null hypothesis is that establishments are randomly distributed, but with their actual size, which greatly increases the un-

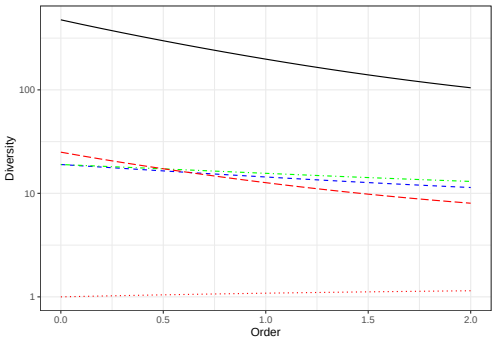

Figure 5. Diversity profiles of European countries: joint diversity (solid curve, black), intra-country α diversity (blue dotted line), effective number of countries (β diversity, average relative specialization: short-dotted line, red), diversity of Europe (γ: long-dotted line, green) and replication of countries (long-dotted line, red). The diversity scale is logarithmic.

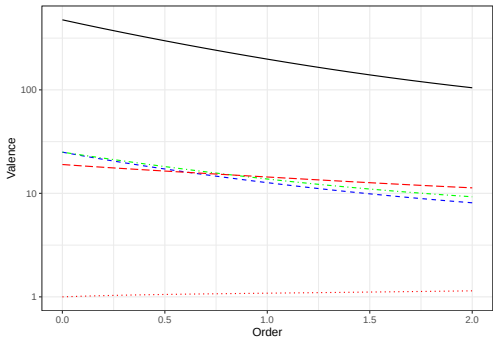

Figure 6. Valence profiles of industry sectors: joint diversity (solid curve, black), intra-sector valence (α: blue dotted line), effective number of sectors (relative concentration, β valence: short-dotted line, red), whole industry valence (γ: long-dotted line, green), and sector replication (long-dotted line, red). The scale is logarithmic.

certainty about the simulated specialization. Individual data on establishments, or at least on their size distribution, are needed to go further. With the data available here, all the profiles presented in figures 1 through 4 are significantly different from each other as early as order 0.1.

4. Joint Diversity

The sector and country contingency table allows us to decompose the total diversity and derive several interesting insights. To preserve the properties of the decomposition, diversity and valence will not be transformed into specialization and concentration.

Cutrini (2010) called the mutual information of the contingency table the relative concentration of the sectors, equal to the relative specialization of the regions measured by the Kullback-Leibler divergence (Theil's relative index). This is only part of the information provided by the data, β diversity or valence, and this approach is only valid at order 1. Joint diversity exploits all the information by combining α and β diversity or valence (together forming γ diversity) and replication.

4.1 Diversity (specialization) of countries

The joint diversity is decomposed into the product of α diversity, the effective number of industry sectors per country, β diversity, the effective number of countries, and replication, the number of replicas of these effective countries. γ diversity the product of α and β, is that of Europe. The decomposition is valid at all orders of diversity, and presented in the form of profiles (figure 5).

The profiles are plotted on a logarithmic scale because the values are of different orders of magnitude and also because the multiplicative decomposition becomes additive in this form: the height of the diversity joined on the figure is the sum of the heights of the *alpha* and *beta* diversities and of the replication.

The diversity of Europe (γ, green curve) has already been presented in figure 2. It is very close to the α, intra-country diversity (blue curve): the effective number of countries (β, relative specialization, red curve) varies from 1 (at order 0) to 1.1 at order 2. This value is very small: the maximum possible is the number of countries, 25, if they do not share any sectors; in fact only 19 because the number of sectors here is less than the number of countries. The countries thus have a certain level of absolute specialization, but it is identical to that of the whole of Europe: their relative specialization is very low. Relative specialization increases with the order considered, i.e. by progressively neglecting the smallest sectors: the countries are a little more different from each other when considering only the most important sectors.

The replication of countries is therefore high: from 25 at order 0 (all countries contain all sectors) to 8 at order 2.

4.2 Valence (concentration) of sectors

Figure 6 shows the valence decomposition of industry sectors.

The results are similar to those for specialization. The sectors are concentrated in absolute terms, but their relative concentration is very low and replication is high.

This large replication of countries and sectors shows that at this level of data aggregation, European industry has little variation in structure between countries or sectors.

5. Conclusion

Information theory, statistical physics and statistical ecology have developed methods to define and rigorously measure uncertainty, diversity and heterogeneity in general. The methods presented here unify and

extend widespread approaches in economics: the measurement of spatial concentration and specialization by entropy, and its decomposition. The main contributions are a clearer mathematical framework, the systematic use of generalized entropy and the quantification of heterogeneity by effective numbers that allow a clear interpretation of the quantities considered.

Not all methodological possibilities have been explored. An important aspect of biodiversity measurement is its estimation from sampled rather than exhaustive data (Marcon, 2015), opening the possibility of assessing concentration or specialization from surveys rather than from public or commercial databases. Functional or phylogenetic diversity (Marcon and Hérault, 2015a) would also allow for the differentiation of sectors from each other in assessing specialization or the proximity of occupied regions in measuring spatial concentration.

References

Adelman, M. A. (1969). Comment on the "H" Concentration Measure as a Numbers-Equivalent. *The Review of Economics and Statistics 51*(1), 99–101.

Alonso-Villar, O. and C. Del Río (2013). Concentration of Economic Activity: An Analytical Framework. *Regional Studies 47*(5), 756–772.

Amiti, M. (1997). Specialisation Patterns in Europe. Technical report.

Atkinson, A. B. (1970, sep). On the measurement of inequality. *Journal of Economic Theory 2*(3), 244–263.

Baez, J. C., T. Fritz, and T. Leinster (2011). A characterization of entropy in terms of information loss. *Entropy 13*(11), 1945–1957.

Baldwin, R. E. and P. Martin (2004). Agglomeration and regional growth. In J. V. Henderson and J.-F. Thisse (Eds.), *Handbook of Urban and Regional Economics*. Amsterdam: Elsevier. North Holland.

Bickenbach, F. and E. Bode (2008). Disproportionality Measures of Concentration, Specialization, and Localization. *International Regional Science Review 31*(4), 359–388.

Bickenbach, F., E. Bode, and C. Krieger-Boden (2013). Closing the gap between absolute and relative measures of localization, concentration or specialization. *Papers in Regional Science 92*(3), 465–480.

Bourguignon, F. (1979). Decomposable Income Inequality Measures. *Econometrica 47*(4), 901–920.

Brülhart, M. and R. Traeger (2005). An Account of Geographic Concentration Patterns in Europe. *Regional Science and Urban Economics 35*(6), 597–624.

Carnot, S. (1824). *Réflexions sur la puissance motrice du feu*. Paris: Bachelier.

Ceriani, L. and P. Verme (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality 10*(3), 421–443.

Chao, A., Y.-T. Wang, and L. Jost (2013). Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution 4*(11), 1091–1100.

Chapin, F. S. I., E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack, and S. Díaz (2000). Consequences of changing biodiversity. *Nature 405*(6783), 234–242.

Clausius, R. (1868). *Théorie mécanique de la chaleur*. Paris: Eugène Lacroix.

Combes, P.-P. and L. Gobillon (2015). The empirics of agglomeration economies. In G. Duranton, J. V. Henderson, and W. C. Strange (Eds.), *Handbook of Urban and Regional Economics*, Volume 5, Chapter 5, pp. 247–348. Amsterdam: Elsevier.

Combes, P.-P. and H. G. Overman (2004). The spatial distribution of economic activities in the European Union. In J. V. Henderson and J.-F. Thisse (Eds.), *Handbook of Urban and Regional Economics*, Volume 4, Chapter 64, pp. 2845–2909. Amsterdam: Elsevier. North Holland.

Conceição, P. and P. Ferreira (2000). The Young Person's Guide to the Theil Index: Suggesting Intuitive Interpretations and Exploring Analytical Applications. Technical report, Austin, Texas.

Cutrini, E. (2009). Using entropy measures to disentangle regional from national localization patterns. *Regional Science and Urban Economics 39*(2), 243–250.

Cutrini, E. (2010). Specialization and Concentration from a Twofold Geographical Perspective: Evidence from Europe. *Regional Studies 44*(3), 315–336.

Daróczy, Z. (1970). Generalized information functions. *Information and Control 16*(1), 36–51.

Davis, H. T. (1941). *The theory of econometrics*. Bloomington, Indiana: The Principia Press.

Dewar, R. C. and A. Porté (2008). Statistical mechanics unifies different ecological patterns. *Journal of theoretical biology 251*(3), 389–403.

Ellison, A. M. (2010). Partitioning diversity. *Ecology 91*(7), 1962–1963.

Ellison, G. and E. L. Glaeser (1997). Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. *Journal of Political Economy 105*(5), 889–927.

Faddeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Mat. Nauk 1*(67), 227–231.

Gini, C. (1912). *Variabilità e mutabilità*. Bologna: C. Cuppini.

Gregorius, H.-R. (1991). On the concept of effective number. *Theoretical population biology 40*(2), 269–83.

Gregorius, H.-R. (2010). Linking Diversity and Differentiation. *Diversity 2*(3), 370–394.

Gregorius, H.-R. (2014). Partitioning of diversity : the "within communities" component. *Web Ecology 14*, 51–60.

Haedo, C. and M. Mouchart (2017). A stochastic independence approach for different measures of concentration and specialization. *Papers in Regional Science in press*.

Havrda, J. and F. Charvát (1967). Quantification method of classification processes. Concept of structural alpha-entropy. *Kybernetika 3*(1), 30–35.

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology 54*(2), 427–432.

Hirschman, A. O. (1964). The Paternity of an Index. *The American Economic Review 54*(5), 761–762.

Houdebine, M. (1999). Concentration Géographique des Activités et Spécialisation des Départements Français. *Economie et Statistique 326-327*(6-7), 189–204.

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology 52*(4), 577–586.

Jost, L. (2006). Entropy and diversity. *Oikos 113*(2), 363–375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology 88*(10), 2427–2439.

Krugman, P. (1991). *Geography and Trade*. London: MIT Press.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Levins, R. (1968). *Evolution in Changing Environments: Some Theoretical Explorations*. Princeton University Press.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews 40*(4), 510–533.

Magurran, A. E. (1988). *Ecological diversity and its measurement*. Princeton, NJ: Princeton University Press.

Marcon, E. (2015). Practical Estimation of Diversity from Abundance Data. *HAL 01212435*(version 2).

Marcon, E. (2017). *Mesures de la Biodiversité*. Kourou, France: UMR EcoFoG.

Marcon, E. and B. Hérault (2015a). Decomposing Phylodiversity. *Methods in Ecology and Evolution 6*(3), 333–339.

Marcon, E. and B. Hérault (2015b). entropart, an R Package to Measure and Partition Diversity. *Journal of Statistical Software 67*(8), 1–26.

Marcon, E., B. Hérault, C. Baraloto, and G. Lang (2012). The Decomposition of Shannon's Entropy and a Confidence Interval for Beta Diversity. *Oikos 121*(4), 516–522.

Marcon, E. and F. Puech (2017). A Typology of Distance-Based Measures of Spatial Concentration. *Regional Science and Urban Economics 62*, 56–67.

Marcon, E., I. Scotti, B. Hérault, V. Rossi, and G. Lang (2014). Generalization of the Partitioning of Shannon Diversity. *Plos One 9*(3), e90289.

Marshall, A. (1890). *Principle of Economics*. London: Macmillan.

Mendes, R. S., L. R. Evangelista, S. M. Thomaz, A. A. Agostinho, and L. C. Gomes (2008). A unified index to measure ecological diversity and species rarity. *Ecography 31*(4), 450–456.

Mori, T., K. Nishikimi, and T. E. Smith (2005). A Divergence Statistic for Industrial Localization. *The Review of Economics and Statistics 87*(4), 635–651.

Ottaviano, G. I. P. and D. Puga (1998). Agglomeration in the global economy: A survey of the new economic geography. *The World Economy 21*(6), 707–731.

Patil, G. P. and C. Taillie (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association 77*(379), 548–561.

Pavoine, S., E. Marcon, and C. Ricotta (2016). 'Equivalent numbers' for species, phylogenetic or functional diversity in a nested hierarchy of multiple scales. *Methods in Ecology and Evolution 7*(10), 1152–1163.

Peet, R. K. (1974). The measurement of species diversity. *Annual review of ecology and systematics 5*, 285–307.

Pielou, E. C. (1975). *Ecological Diversity*. New York: Wiley.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. and T. K. Nayak (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory 31*(5), 589–593.

Rényi, A. (1961). On Measures of Entropy and Information. In J. Neyman (Ed.), *4th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, USA, pp. 547–561. University of California Press.

Richard-Hansen, C., G. Jaouen, T. Denis, O. Brunaux, E. Marcon, and S. Guitet (2015). Landscape patterns influence communities of medium- to large-bodied vertebrate in undisturbed terra firme forests of French Guiana. *Journal of Tropical Ecology 31*(5), 423–436.

Rysman, M. and S. Greenstein (2005). Testing for agglomeration and dispersion. *Economics Letters 86*, 405–411.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal 27*, 379–423, 623–656.

Sharp, K. and F. Matschinsky (2015). Translation of Ludwig Boltzmann's paper "on the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium". *Entropy 17*(4), 1971–2009.

Shorrocks, A. F. (1980, apr). The Class of Additively Decomposable Inequality Measures. *Econometrica 48*(3), 613.

Simpson, E. H. (1949). Measurement of diversity. *Nature 163*(4148), 688.

Theil, H. (1967). *Economics and Information Theory*. Chicago: Rand McNally & Company.

Tothmeresz, B. (1995). Comparison of different methods for diversity ordering. *Journal of Vegetation Science 6*(2), 283–290.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics 52*(1), 479–487.

Tsallis, C. (1994). What are the numbers that experiments provide? *Química Nova 17*(6), 468–471.

Weber, A. (1909). *Über den Standort der Industrien*. Tübingen. English translation edited in 1971, "Theory of the location of industries", Russell & Russell.

Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs 30*(3), 279–338.

Wilson, E. O. and F. M. Peter (Eds.) (1988). *Biodiversity*. Washington, D.C.: The National Academies Press.

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics 16*(2), 97–159.