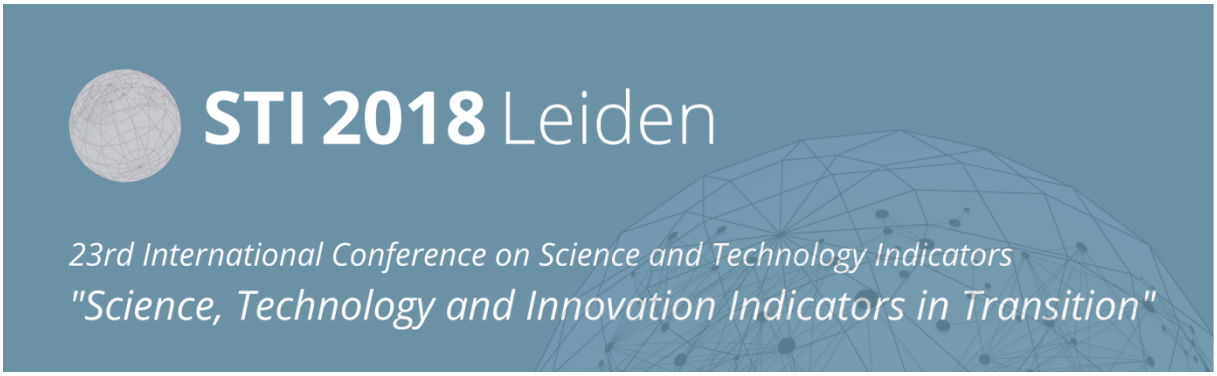
A new method for diversity measurement: Taking similarity between cells seriously

Rousseau, R.



STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

A new method for diversity measurement: Taking similarity between cells seriously

Ronald Rousseau

Ronald.Rousseau@kuleuven.be
ECOOM, KU Leuven, Naamsestraat 61, Leuven 3000 (Belgium)

Ronald.Rousseau@uantwerpen.be
Faculty of Social Sciences, University of Antwerp, Middelheimlaan 1, Antwerp 2020 (Belgium)

Introduction

Diversity is an attribute of any system whose elements may be apportioned into categories or cells (Jones & Leonard, 1989). In (Rousseau & Van Hecke, 1999) we stated that the main components of biodiversity are species richness and evenness while in Nijssen et al. (1998) we recalled that the Lorenz curve provides a truthful representation of evenness. This observation was already made – implicitly – in Solomon (1979) and Patil and Taillie (1979). Lorenz curves of the arrays (6,3,1) and (6,6,3,3,1,1) coincide and hence have the same evenness. Thus measures derived from the Lorenz curve do not take the number of cells into account. It was realized that combining species richness and evenness is not a simple endeavour as many well-known diversity measures such as the Gini index or the coefficient of variation do not take species richness into account and are only acceptable if the number of species, N, is fixed. Yet, the Hirschman-Simpson-Herfindahl index or the repeat rate (Rousseau, 2018) does take variety and evenness into account. This is one of the reasons why nowadays the Rao-Stirling index, an extension of the repeat rate which, moreover, takes similarity between cells into account, is popular as a diversity index used in interdisciplinarity studies (Rafols & Meyer, 2010).

True diversity in the sense of Jost

Jost (2009) convincingly argued that when studying (bio)diversity one should be able to give a meaning to sentences such as "Diversity has decreased by 50%". Diversity measures satisfying the following six properties are those for which reasoning in terms of ratios or percentages is possible. Jost refers to them as true diversities. These six requirements are:

1. Symmetry (anonymity). A diversity function is symmetric in its arguments.
2. Zero output independence. Adding a species with zero abundance does not change the diversity.
3. Transfer principle (Dalton, 1920). Transferring a unit abundance from a rarer species to a more common species should not increase diversity.
4. Homogeneity (scale invariance). Diversity depends only on species relative frequencies and not on their absolute abundances.
5. Replication principle. Suppose m communities have identical sets of species abundances, but no species are shared between any of the communities. All m communities necessarily

have the same diversity D_0. Suppose we pool all m communities. Then the diversity of the m pooled equally diverse, equally large, completely distinct communities must be $m.D_0$.
6. Normalization. If the diversity measure is applied to S equally common species, its value is S.

The following Hill-type (Hill, 1973) measures are "true diversities":

$$\left(\sum_{i=1}^{N} p_i^q\right)^{1/(1-q)} \tag{2}$$

where q is a parameter with values ranging from 0 to infinity (the cases q=1 and q=∞ are obtained as limits). For q = 2, this is: $\left(\sum_{i=1}^{N} p_i^2\right)^{-1} = \frac{1}{\sum_{i=1}^{N} p_i^2}$ which is the reciprocal of the repeat rate (Rousseau, 2018).

Well-known measures such as the Simpson diversity index $1-\sum_{i=1}^{N} p_i^2$ and the Shannon entropy measure: $-\sum_{i=1}^{N} p_i \ln(p_i)$, where p_i denotes the relative abundance in cell i (for instance: species i) do not meet these requirements. As such, when using the Simpson measure or the Shannon entropy, it makes no sense to state that diversity has increased by 40%. Such statements can only be made in the context of measures that satisfy Jost's requirements.
In the next sections we will point out that, contrary to what we stated in (Rousseau & Van Hecke, 1999) taking species richness (or variety) and evenness (or balance) into account does not suffice for a full characterization of diversity.

Taking similarity into account
There is a third, maybe even more fundamental, aspect for studying diversity. This third aspect is called disparity and characterizes how different species (cells) are between each other. The why and the how of taking these three aspects into account has been thoroughly studied – in a general context – by Stirling (2007). This led to the proposal to use Rao's quadratic entropy measure (Rao, 1982):

$$R = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} d_{ij} p_i p_j = \langle P, D * P \rangle \tag{3}$$

as a diversity measure. Here D is a dissimilarity matrix with elements $(d_{ij})_{ij}$. If $0 \leq d_{ij} \leq 1$ and considering similarity as the antonym of disparity, we can form a corresponding similarity matrix S, with $s_{ij} = 1 - d_{ij}$. D*P denotes the product of the disparity matrix D and the array P. The symbol <.,.> denotes the standard scalar or inner product of two arrays. Rao describes this index as the expected dissimilarity between two individuals selected randomly with replacement, where d_{ij} is the dissimilarity (disparity) between species i and j, brought together in a disparity matrix D. This measure and the related Integration Score have been used in a number of interdisciplinarity studies such as (Porter & Rafols, 2009; Rafols & Meyer, 2010; Leydesdorff & Rafols, 2011, Wagner et al., 2011). Indeed, interdisciplinarity, or with a more general term, knowledge integration, involves the use of information coming from a diversity of fields or subfields. A similar formula has been used in studying nucleotide sequences (Nei & Li, 1979).

Other suggestions for taking similarity or its opposite, disparity, into account

Unfortunately, in the same way as the Simpson index is not a true diversity measure Rao's quadratic measure is not a true diversity when disparity is taken into account. A family of measures which is, was proposed by Leinster and Cobbold (2012):

$$ {}^{q}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij} p_j \right)^{q-1} \right)^{\frac{1}{1-q}} \tag{4} $$

where S = (s_{ij}) is a similarity matrix and q is a parameter with values ranging from 0 to infinity (again, the cases q=1 and q=∞ are obtained as limits). We assume that these similarities s_{ij} are such that $s_{ii} = 1$ for all i and $0 \leq s_{ij} = s_{ji} \leq 1$. For the special case q=2, this leads to:

$$ {}^{2}D^{S} = \left(\sum_{i=1}^{N} p_i \left(\sum_{j=1}^{N} s_{ij} p_j \right) \right)^{-1} = \frac{1}{\sum_{i,j=1}^{N} s_{ij} p_i p_j} = \frac{1}{\langle P, S * P \rangle} \tag{5} $$

We note from formula (5) that $ {}^{2}D^{S} = \frac{1}{1-R} = \frac{1}{1-\langle P, D * P \rangle}$, where R is Rao's quadratic entropy measure. For this reason we refer to this measure as the modified Rao-Stirling diversity measure. A practical application of this measure in the case of diversity of article references is elaborated in (Zhang et al., 2016). These $^{q}D_S$ measures have the following nine properties (Leinster & Cobbold, 2012):

LC1. Effective number: the diversity of a community of N equally abundant, totally dissimilar species (the similarity matrix is the identity matrix) is N.
LC2. Modularity: suppose that the community is partitioned into m subcommunities, with no species shared between subcommunities, and with species in different subcommunities being totally dissimilar. Then the diversity of the community is entirely determined by the sizes and diversities of the subcommunities.
LC3. Replication: if, moreover, these m subcommunities are of equal size and equal diversity, D_0, then the diversity of the whole community is mD_0.
LC4. Symmetry (anonymity): diversity is unchanged by the order in which species happen to be listed originally.
LC5. Absent species: diversity is unchanged by adding a new species of abundance 0.
LC6. Identical species: if two species are identical, then merging them into one leaves the diversity unchanged.
LC7. Monotonicity: when the similarities between species increase, diversity decreases.
LC8. Naive model: when similarities between species are ignored ($s_{ij} = 0$, $i \neq j$), diversity is greater than when they are taken into account.
LC9. Range: the diversity of a community of N species is between 1 and N.

Proofs that the $^{q}D_S$ satisfy these properties for all q are provided in (Leinster and Cobbold, 2012). Interestingly, we note that Leinster and Cobbold (2012) do not mention balance or the Lorenz curve in their study of true diversity measures. Their list of properties leads to the question: is this a list of requirements for proper diversity measures or just a list of observed properties of the set of $^{q}D_S$ measures? We tend to think that this is not a list of strict requirements, but intend to elaborate on this in a later study.

The monotonicity of balance requirement

Stirling (2007) drew a list of ten desirable features for a general diversity heuristic Δ and pointed out that no measure can meet all requirements. As stated earlier he opted for the quadratic entropy as a reasonable compromise.

Inspired by Stirling's list we formulate the following requirements: monotonicity of variety, balance and disparity. These three requirements are of the form: if two aspects are given then Δ increases with the remaining third. For instance, monotonicity in balance implies that for given variety and disparity, the diversity measure should increase monotonically with balance. We recall that in (Rousseau, 2018) we gave a counterexample for which the array with the larger evenness or balance has the smaller diversity value as measured with the Rao-Stirling measure (with the same variety and disparity). This counterexample showed that the Rao-Stirling measure does not meet the "monotonicity of balance" requirement. Clearly, the same statement holds for its "true diversity" variant (Zhang et al., 2016).

In our opinion this result is perfectly natural and actually shows the role played by similarity, see next section. In the previous sections we provided an overview of known facts about diversity. In the next ones we suggest a new way to measure diversity, provide an example and come to a conclusion.

Taking similarity seriously

We claim that when similarity is taken into account one must take this fact seriously. By this we mean that the observations, say array X, are just a point of departure, but the real array of interest is S*X = Y. We recall that this array was introduced in (Rousseau et al., 2017) in a study of the composition of peer review panels. It is called a similarity-adapted array. Now we transform Y to its normalized form P_Y (sum of coordinates equal to one i.e. an L_1-norm). Then Hill-type (Hill, 1973) measures of P_Y can be calculated, leading to a family of true diversities.

The new method

Based on the above reflections we propose the following approach to diversity measurement

1) Determine the aim of the investigation and collect data.

2) Determine the number of cells (N), i.e. variety.

This number N can be the number of observed cells or a theoretical number (which then includes empty cells). When studying butterflies in the park one probably has to use the number of different observed species; unless this is done already for many years, then one could use the number of species ever recorded. When composing expert panels or in interdisciplinarity studies using WOS Subject Categories as cells it seems natural to use all categories and hence include empty cells. This leads to an N-dimensional array X.

3) Take similarity into account.

We assume that an NxN similarity matrix S is known. Then one forms the similarity-adapted array S*X=Y. Note that variety determines the dimensions of this similarity matrix. Through this operation some empty cells may not be empty anymore. This operation implies that taking similarity of empty cells into account is logical in some circumstances, typically in interdisciplinarity studies, and makes no sense in other (studying butterflies: if a similar butterfly has not been observed it just is not present).

4) Taking evenness into account (for the transformed array!)

Only now and taking (dis)similarity serious (not as something added at the end) balance or evenness is taken into account. As a first step we transform Y to its normalized form P_Y (sum of coordinates equal to one). Then Hill-type measures can be calculated, leading to a family of true diversities.

An example.
We observed the array $(10,8,4)^t$, which is actually $(10,8,4,0)^t$ as one cell for which items could be present, was empty. The 4x4 similarity matrix is: $\begin{pmatrix} 1 & 0.4 & 0.5 & 0.5 \\ 0.4 & 1 & 0.6 & 0.6 \\ 0.5 & 0.6 & 1 & 0.3 \\ 0.5 & 0.6 & 0.3 & 1 \end{pmatrix}$.
Based on actual observations (N=3) and using the modified Rao-Stirling diversity measure leads to a value of 1.502. Using the new order of operations an using the reciprocal of the repeat rate yields a value of 2.995. Applying these calculations for N=4 yields again 1.502 (a consequence of property LC5) for the modified Rao-Stirling diversity and 3.947 for the reciprocal of the repeat rate.

Conclusion
We proposed a new way of measuring diversity in any field. Because of this new proposal, mathematical requirements (axioms) for diversity measures do not apply to the original observations (the array X), but to the similarity-adapted array (Y) and its normalized form (P_Y). The main difference between the two approaches is the treatment of empty cells.

References
Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal*, 30, 348-361.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2), 427-432.

Jost, L. (2009). Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics*, 68(4), 925-928.

Jones, G.T. & Leonard, R.D. (1989). The concept of diversity: an introduction. In R.D. Leonard & G.T. Jones (Eds.), *Quantifying Diversity in Archaeology* (pp. 2-3). Cambridge: Cambridge University Press.

Lambshead, P.J.D., Platt, H.M. & Shaw, K.M. (1983). Detection of differences among assemblages of marine benthic species based on an assessment of dominance and diversity. *Journal of Natural History (London)*, 17(6), 859- 874.

Leinster, T. & Cobbold, C.A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, 93(3), 477-489.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics*, 5(1), 87-100.

Nei, M., & Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Science USA*, 76(10), 5269-5273.

Nijssen, D., Rousseau, R. and Van Hecke, P. (1998). The Lorenz curve: a graphical representation of evenness. *Coenoses*, 13(1), 33-38.

Patil, G.P. & Taillie, C. (1979). An overview of diversity. In J.F. Grassle, G.P. Patil, W. Smith & C. Taillie (Eds.), *Ecological Diversity in Theory and Practice*, (pp. 3-27). Fairland (MD): International Co-operative Publishing House.

Porter, A. L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, 81(3), 719-745.

Rafols, I. & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(3), 263–287.

Rao, C.R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1), 24-43.

Rousseau, R. (2018). The repeat rate: From Hirschman to Stirling. *Scientometrics*, 116(1), 645–653

Rousseau, R., Guns, R., Rahman, A.I.M.J. & Engels, T.C.E. (2017). Measuring cognitive distance between publication portfolios. *Journal of Informetrics*, 11(2), 583-594.

Rousseau, R. & Van Hecke, P. (1999). Measuring biodiversity. *Acta Biotheoretica* 47(1), 1-5.

Solomon, D.L. (1979). A comparative approach to species diversity. In J.F. Grassle, G.P. Patil, W. Smith & C. Taillie (Eds.), *Ecological Diversity in Theory and Practice*, (pp. 29-35). Fairland (MD): International Co-operative Publishing House.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707-719.

Wagner, C. S., Roessner, J.D., Bobb, K., Klein, J. T., Boyack, K.W., Keyton, J., Rafols, I. & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14-26.

Zhang, L., Rousseau, R. & Glänzel, W. (2016). Diversity of references as an indicator for interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology*, 67(5), 1257–1265.