Communication

On the Optimal Point of the Weighted Simpson Index

José Pinto Casquilho * and Helena Mena-Matos *

Posted Date: 28 December 2023

doi: 10.20944/preprints202312.2137.v1

Keywords: Weighted Simpson index; Lagrange multiplier method; Critical point; Minimum value; Harmonic mean; Inversion problem

Communication

On the Optimal Point of the Weighted Simpson Index

José Pinto Casquilho [1,*] and Helena Mena-Matos [2,*]

[1] Postgraduate and Research Program, Universidade Nacional Timor Lorosa'e (UNTL), República Democrática de Timor-Leste; jose.casquilho@untl.edu.tl

[2] Centro de Matemática da Universidade do Porto (CMUP) e Faculdade de Ciências da Universidade do Porto (FCUP), Portugal; mmmatos@fc.up.pt

* Correspondence: josecasquilho@gmail.com; mmmatos@fc.up.pt

Abstract: In this short communication, following a brief introduction, we undertake a comprehensive analytical study of the weighted Simpson index. Our primary emphasis concerns the precise determination of the optimal point (minimizer) coordinates and of the minimum value of the index, a differentiable convex function, which is related to the harmonic mean concept. Furthermore, we address and solve the inversion problem and show the tight connection between both approaches. Last, we give some insights and final remarks on this subject.

Keywords: weighted Simpson index; Lagrange multiplier method; Critical point; minimum value; harmonic mean; inversion problem

MSC: 49N15; 90C46; 90C51; 92B99

1. Introduction

The Simpson index concerning a population distributed among k categories or classes is defined as

$$S = \sum_{i=1}^{k} p_i^2,$$

where p_i denotes the probability (or proportion of occurrences) of class i. So, one has $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$, and therefore S is defined on a $k-1$ simplex. This index equals the probability that two elements taken at random from the population of interest belong to the same category or class. The value of Simpson's index ranges from $1/k$ to 1, with 1 representing no diversity, so, the larger the value of S the lower the diversity. The name "Simpson index" roots from the influential 1949 paper by Edward Hugh Simpson entitled "Measurement of Diversity" [1], wherein he introduced what he called a measure of concentration defined in terms of population constants, with the minimum concentration equaling the maximum diversity. The Simpson index became a widely used quantitative metric in ecological and biodiversity studies as a tool for assessing and quantifying the diversity and evenness of species within ecological communities. It also applies to other biological problems, including biomedical sciences, such as measuring diversity concerning immunity in response to viral infections (e.g., [2]).

However, it is also acknowledged that the original mathematical concept formulation was used in cryptanalysis as far back as the 1920s and 30s — therein named *probability of monographic coincidence* — by the American cryptanalysts William Friedman and Solomon Kullback (e.g., [3]). It is relevant to note that the Italian statistician Corrado Gini had already applied the quantity $1 - \sum_{i=1}^{k} f_i^2$ as early as 1912. He defined the index with relative frequencies f_i computed from large samples, referring to it as an *index of mutability for disconnected (qualitative) variables* [4]. This quantity became later known as the "Gini-Simpson index", a name adopted in the 1980s by the eminent statistician C. R. Rao (e.g., [5,6]), who restated it with probabilities as $G = \sum_{i=1}^{k} p_i(1-p_i) = 1 - \sum_{i=1}^{k} p_i^2$. For instance, Jian et al. [7] consider a "Livelihood Simpson Index" which in fact is a Gini-Simpson index. Obviously, given a

 doi:10.20944/preprints202312.2137.v1

probability distribution $(p_1, \dots, p_k)$, the Simpson and Gini-Simpson indices correspond to complementary events, verifying $S + G = 1$.

The use of a weighted version of the Simpson index appears to have been first reported in 1992, when Nowak and May [8] conceived the effective immune response against the virus population composed of different strains in the context of HIV infections, then revisited two years later [9].

Some refer to the weighted Simpson index when they are actually dealing with the weighted Gini-Simpson index (e.g., [10,11]). The weighted Gini-Simpson index is defined as $GS_w = \sum_{i=1}^{k} w_i p_i (1 - p_i)$, a concave function, differentiable in the interior of a simplex, with an identifiable maximum value for which a method to determine the optimal point (maximizer) was framed based on the fact that one is dealing with feasibility values associated with the constraints of the simplex [12] – namely, that the optimal coordinates must verify $p_i^* \geq 0$ – what was not taken into account in [11] and may lead to miscalculations.

Yet, Kasulo and Perrings used a price-weighted Simpson index [13] to assess scenarios relative to the connection between the diversity of catch in a multi-species fishery and profit maximizing regimes. And one can still find an inverse form of the weighted Simpson index in [14], which the authors clarify could be interpreted as a weighted version of Hill's number N_2, concerning the unifying notation that was proposed therein [15]. Also, one can see that J. Ma used a symmetric form of the weighted Simpson index, building what he named a "comprehensive weighted Gini-Simpson index" [16].

In recent years, there has been a panoply of diversity, phylogenetic, and dissimilarity indices with a focus on biology (e.g., [17,18]), most of them associated with developments related to Rényi measures of entropy [19], including general reviews on the subject (e.g.,[20]), and a plethora of recent applications (e.g., [21]). There are also many publications within the scope of diversity in the social sciences – a process that had a cornerstone with the reference work of Patil and Taillie [22] addressing linguistic diversity, industrial concentration and income inequality – with applications either in economics (e.g., [23]) or demographics, such as the weighted Rényi's entropy for lifetime distributions [24].

However, with regard to the weighted Simpson index, it seems that an analytical study addressing the optimal point (minimizer) and the optimal value (minimum), has not yet been published.

2. The weighted Simpson index

Herein, we define the weighted Simpson index, concerning a population distributed among k categories or classes, as:

$$S_w = \sum_{i=1}^{k} w_i p_i^2 \tag{1}$$

where p_i denotes the probability (or proportion of occurrences) of class i and $w_i > 0$ is a weight assigned to that class, altogether defining a vector of positive real values $w = (w_1, \dots, w_k)^t$. For now, we have decided not to impose any extra conditions on the weights, leaving this matter to be discussed later. Our current focus is on understanding the broader context.

Weights allow to consider various features for the classes. In the context of biodiversity, these features may be related for example to environmental benefits, conservation importance, vulnerability or economic value of species – a subject that was already emphasized at least since the beginning of the 1980s, exemplified with biomass and other importance values [25]. The environmental benefits may include the ecological roles and contributions of a species to its ecosystem. Conservation importance further adds depth to the evaluation, addressing the urgency and priority assigned to preserving certain species to maintain ecosystem stability. Vulnerability is another important factor that weights help account for. Vulnerable species—those in danger of going extinct—need to be given particular attention during the assessment process. Economic value represents yet another dimension where weights come into play. The assessment of species in terms of their economic contributions, whether through ecosystem services, medicinal properties, or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2023 doi:10.20944/preprints202312.2137.v1

commercial value, demands a careful weighting to reflect their overall importance. Yet, one should be aware of the complexities and entanglements associated with a community structure when dealing with several trophic levels, where, in addition to competition, there may occur symbiosis and predator-prey interactions conveying a hierarchical structure in a networked ecological community (e.g., [26]).

2.1. Optimizing the weighted Simpson index

The weighted Simpson index (1) is defined on the simplex $\Delta^{k-1} = \{(p_1, \dots, p_k) \in \mathbb{R}^k : p_i \geq 0, \sum_{i=1}^k p_i = 1\}$ and is a convex function. So, S_w attains its (global) maximum at the boundary of Δ^{k-1} and its (global) minimum at the interior of Δ^{k-1}. Clearly, the maximum is attained when all except one of the p_i are zero, and therefore one has $max\, S_w = \max_i w_i$.

The minimum can be assessed with the method of Lagrange multipliers for finding the minima of S_w subject to the equality constraint $\sum_{i=1}^k p_i = 1$. As S_w is differentiable in the interior of Δ^{k-1}, one can build the Lagrangian function

$$L_w(p_1, \dots, p_k; \lambda) = S_w(p_1, \dots, p_k) + \lambda\left(\sum_{i=1}^k p_i - 1\right),$$

and find its extrema. Equating partial derivatives to zero, we get:

$$\begin{cases} \frac{\partial L_w}{\partial p_i} = 0 \\ \frac{\partial L_w}{\partial \lambda} = 0 \end{cases} \Leftrightarrow \begin{cases} 2w_i p_i + \lambda = 0 \\ \sum_{i=1}^k p_i - 1 = 0 \end{cases} \quad (2)$$

From (2) we conclude that for any specific j the relationship $w_j p_j = -\lambda/2 \Leftrightarrow p_j = -\frac{\lambda}{2w_j}$ holds. Now, using the constraint $\sum_{j=1}^k p_j = 1$ one has the following equivalence $1 = -\frac{\lambda}{2}\sum_{i=1}^k \frac{1}{w_i} \Leftrightarrow 1 = w_j p_j \sum_{i=1}^k \frac{1}{w_i}$ and we get the optimal coordinates of the minimum point given by:

$$p_j^* = \frac{1}{w_j\left(\sum_{i=1}^k \frac{1}{w_i}\right)} \qquad \text{for } j = 1, \dots k. \quad (3)$$

Also, using (3) one can evaluate the minimum value of the weighted Simpson index (1) as follows:

$$S_w^* = S_w(p_1^*, \dots, p_k^*) = \sum_{j=1}^k w_j \left(\frac{1}{w_j\left(\sum_{i=1}^k \frac{1}{w_i}\right)}\right)^2 = \frac{1}{\left(\sum_{i=1}^k \frac{1}{w_i}\right)^2}\sum_{j=1}^k \frac{1}{w_j} = \frac{1}{\sum_{i=1}^k \frac{1}{w_i}} \quad (4)$$

2.2. Some further comments on the minimizer

Note that the minimum value of the weighted Simpson index (4) is related to the harmonic mean of the weights $\mathrm{H}(w)$ by $S_w^* = \mathrm{H}(w)/k$. The name "harmonic mean" is said to have been proposed by Archytas and Hippasus and adopted by Nicomachus in accordance with the view of the geometrical harmony of the cube, because it has 12 edges, 8 vertices, and 6 faces, and 8 is the mean of 12 and 6 according to the theory of harmonics [27] (pp. 85-86), meaning the harmonic mean of those quantities. In general, the harmonic mean of k nonzero numbers is the reciprocal of the arithmetic mean of the reciprocals of those numbers, and is appropriate for averaging rates over constant numerator units [28]. As a typical example, if a set of investments are invested at different interest rates, and they all give the same income, the unique rate at which all of the capital tied up in those investments must be invested to produce the same revenue as given by the set of investments, is equal to the harmonic mean of the individual rates [29] (p. 240).

The special case of all weights being equal to 1 leads to $p_i^* = 1/k, \forall i$ and to the minimum value $S_w^* = 1/k$ as can be seen from expressions (3) and (4), and also because in this case $S_w = S$ with S being the (unweighted) Simpson index whose minimum is $1/k$.

Rewriting the optimal coordinates in (3) as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 December 2023 doi:10.20944/preprints202312.2137.v1

$$p_j^* = \frac{1}{1 + w_j \left(\sum_{i \neq j} \frac{1}{w_i}\right)}$$

it is straightforward to conclude that the weights are driving forces of the values of the coordinates of the minimizer, operating reciprocally: when the weight attached to a specific class increases and all the others keep invariant, the corresponding optimal coordinate decreases; and when a weight attached to another class increases, all the others being invariant, the original class increases its optimal coordinate.

If one considers the valuation of the classes of a distribution in the usual sense of importance assessment with an ordering of positive real numbers, expecting that $v_m > v_n$ would promote the result $p_m^* > p_n^*$, then one should be aware that the weights associated with the optimal point (3) would not be the values $\{v_m, v_n\}$ but could possibly be conceived like $w_m = 1/v_m$ and $w_n = 1/v_n$ instead.

2.3. Normalization

The use of normalized indices in applications is important for several reasons. Normalized indices provide a standardized scale, usually ranging from 0 to 1, or 0% to 100%, irrespective of the specific scale or magnitude of the data. This standardization allows for direct comparisons between different datasets or populations and remains consistent across different contexts and scales.

In the case of the weighted Simpson index this can be done in a classic way defining the normalized weighted Simpson index as

$$s_w = \frac{S_w - \min(S_w)}{\max(S_w) - \min(S_w)} = \frac{S_w \sum_{i=1}^k \frac{1}{w_i} - 1}{\max(w_i) \sum_{i=1}^k \frac{1}{w_i} - 1}$$

However, this normalization eliminates the effect of the number of classes in the distribution (e.g., species in a community). For example, in the case of all weights equal to 1, the normalized weighted Simpson index of a population with k species uniformly distributed is always 0 and thus independent of the number of species. The fact that normalized indices of diversity can be misleading has already been mentioned by several authors (e.g., [11]).

In the case of a weighted index, it may be relevant to normalize the weights so that the index becomes dimensionless and independent of the order of magnitude of the weights. For the weighted Simpson index, this normalization can be done, for example, by imposing the condition $\sum_{i=1}^k w_i = 1$. This corresponds to dividing each non-normalized weight w_i by the sum of all the weights. As $S_{\beta w} = \beta S_w$ for $\beta > 0$, this normalization procedure does not affect the results of the previous sections, namely, for the weighted Simpson index with normalized weights $S_{\boldsymbol{w}}$, the optimal coordinates of the minimum point are as in (3) and the value of the minimum of $S_{\boldsymbol{w}}$ is as in (4).

2.4. The inverse problem

The inverse procedure relative to the weighted Gini-Simpson was formulated recently [30]. Now, we consider the analogous problem concerning the weighted Simpson index, stated as: *g*iven a minimum point of S_w denoted by $(p_1^*, \ldots, p_k^*)$, verifying both $0 < p_j^* < 1$ and $\sum_{j=1}^k p_j^* = 1$, what would be a set of weights able to generate that solution? The answer to this question follows straightforward: recalling (3), the weights must be chosen to be inversely proportional to the optimal coordinates p_j^* with the proportionality constant equal to S_w^* (4), as we can see rewriting:

$$p_j^* = \frac{1}{w_j \left(\sum_{i=1}^k \frac{1}{w_i}\right)} \Leftrightarrow \frac{1}{p_j^*} = w_j \left(\sum_{i=1}^k \frac{1}{w_i}\right) \Leftrightarrow \frac{1}{p_j^*} = w_j \frac{1}{S_w^*} \Leftrightarrow w_j = \frac{S_w^*}{p_j^*}.$$

For non-normalized weights, there are infinitely many solutions to the inverse problem, parameterized by the minimum S_w^*. For normalized weights, using the condition $\sum_{j=1}^k w_j = 1$, one gets $\sum_{j=1}^k \frac{S_w^*}{p_j^*} = 1 \Leftrightarrow S_w^* = 1 \Big/ \left(\sum_{j=1}^k \frac{1}{p_j^*}\right)$ and so the weights must be chosen as:

$$w_i = \frac{1}{p_i^*\left(\sum_{j=1}^{k}\frac{1}{p_j^*}\right)} \text{ for } i = 1, \dots, k. \quad (5)$$

It should be noted that from the condition that the sum of the weights equals 1 it follows that $\sum_{i=1}^{k}\frac{1}{p_i^*} = \sum_{i=1}^{k}\frac{1}{w_i}$.

3. Final remarks

We have presented a detailed analytical study of the optimization problem associated with the weighted Simpson index. The core result is that at the optimal point one has $w_i p_i^* = w_j p_j^*$ for $i, j = 1, \dots, k$, and also, that for all i, one gets $w_i p_i^* = S_w^*$, the minimum value of the index. So, there is a trade-off between the weights and the optimal probabilities (or proportions of occurrences) in what could be seen as an equilibrium condition. The fact that Nowak [8,9] has used the weighted Simpson index as a Lyapunov function to assess an antigenic diversity threshold, seems compatible with an equilibrium point perspective, which could also be used in a broad sense concerning different problems within several scientific fields.

Furthermore, for a random variable W with values corresponding to the previous weights $\{w_i\}_{i=1,\dots,k}$, and probability function given by $\Pr[W = w_i] = p_i^*$, with p_i^* defined as in (3), computing the mean value of W entails $E[W] = \sum_{i=1}^{k} w_i p_i^* = k\, S_w^*$, which equals the harmonic mean of the weights, meaning $E[W] = \mathrm{H}(w)$.

Author Contributions: The first author conceived a draft solving the direct optimization problem and the second author revised and built the solution for the inverse problem. Both authors wrote together the paper.

Funding: This research received no specific funding. The second author was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the projects with reference UIDB/00144/2020 and UIDP/00144/2020.

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

References

1. Simpson, E. Measurement of diversity. *Nature* **1949**, *163*, 688. https://doi.org/10.1038/163688a0
2. Davis, C. L.; Adler, F. R. Mathematical models of memory CD8+ T-cell repertoire dynamics in response to viral infections. *B Math Biol.* **2013**, *75*, 491-522. https://doi.org/10.1007/s11538-013-9817-6
3. Österreicher, F.; Casquilho, J. A. P. On the Gini-Simpson index and its generalization – a historic note. *S Afr Stat J* **2018**, *52*(2), 129-137. https://doi.org/10.37920/sasj.2018.52.2.2
4. Gini, C. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. C. Cuppini: Bologna, Italy, 1912. https://www.byterfly.eu/islandora/object/librib:680892#mode/2up
5. Rao, C. R. Diversity and dissimilarity coefficients: a unified approach. *Theor Popul Biol* **1982**, *21*, 24-43. https://doi.org/10.1016/0040-5809(82)90004-1
6. Rao, C. R. Diversity: its measurement, decomposition, apportionment and analysis. *Sankhya: Indian J Stat* **1982**, *11-A*(1), 1-22. http://library.isical.ac.in:8080/xmlui/bitstream/handle/10263/504/82.E01.pdf?sequence=1&isAllowed=y
7. Jiang, X.; Yin, G.; Lou, Y.;Xie, S.; Wei, W, The impact of transformation of farmer's livelihood on the increasing labor costs of grain plantation in China. *Sustainability* **2021**, *13*, 11637. https://doi.org/10.3390/su132111637
8. Nowak, M. A; May, R. M. Coexistence and competition in HIV infections. *J Theor Biol* **1992**, *159*(3), 329-342. https://doi.org/10.1016/S0022-5193(05)80728-3
9. Nowak, M. A. The evolutionary dynamics of HIV infections. In *First European Congress of Mathematics, Paris, July 6-10, 1992. Progress in Mathematics*, Joseph, A.; Mignot, F.; Murat, F.; Prum, B.; Rentschler, R., Eds. Birkhäuser Verlag, Basel, Switzerland, 1994; Volume 120, pp. 311-326. https://doi.org/10.1007/978-3-0348-9112-7_13
10. Subburayalu, S.; Sydnor, T.D. Assessing street tree diversity in four Ohio communities using the weightd Simpson index. *Landscape Urban Plan* **2012**, *106*(1), 44-50. https://doi.org/10.1016/j.landurbplan.2012.02.004
11. Guiaşu, R. C.; Guiaşu, S. Conditional and weighted measures of ecological diversity. *Int J Uncertain Fuzziness Knowl Based Syst* **2003**, *11*(3), 283-300. https://doi.org/10.1142/S0218488503002089
12. Casquilho, J. P. A methodology to determine the maximum value of weighted Gini-Simpson index. *SpringerPlus* **2016**, *5*, 1143.

https://doi.org/10.1186/s40064-016-2754-8
13. Kasulo, V.; Perrings, C. Fishing down the value chain: Biodiversity and access regimes in freshwater fisheries – the case of Malawi. *Ecol Econ* **2006**, *59*(1), 106-114. https://doi.org/10.1016/j.ecolecon.2005.09.029
14. Guiaşu, R. C.; Guiaşu, S. The weighted Gini-Simpson index: Revitalising an old index of biodiversity. *Int J Ecol* **2012**, 478728. http://doi.org/10.1155/2012/478728
15. Hill, M. O. Diversity and evenness: A unifying notation and its consequences. *Ecology* **1973**, *54*, 427-432. https://doi.org/10.2307/1934352
16. Ma, J. Generalized grey target decision method based on the Gini-Simpson index involving mixed attributes and uncertain numbers. *Data Tech Appl* 2019, *53*(4), 484-500. https://doi.org/10.1108/DTA-02-2019-0019
17. Grabchak, M.; Marcon, E.; Lang, G.; Zhang, Z. The generalized Simpson's entropy is a measure of biodiversity. *PLoS ONE* **2017**, *12*(3), e0173305. https://doi.org/10.1371/journal.pone.0173305
18. Ricotta, C.; Szeidl, L.; Pavoine, S. Towards a unifying framework for diversity and dissimilarity coefficients. *Ecol Indic* **2021**, *129*, 107971. https://doi.org/10.1016/j.ecolind.2021.107971
19. Rényi, A. On the dimension and entropy of probability distributions. *Acta Math Acad Scient Hung* **1959**, *10*, 193-215. https://doi.org/10.1007/BF02063299
20. Xu, S.; Böttcher, L.; Chou, T. Diversity in biology: definitions, quantification and models. *Phys Biol* **2020**, *17*, 031001. http://doi.org/10.1088/1478-3975/ab6754
21. Garnier, J.; Lafontaine, P. Dispersal and good habitat quality promote neutral genetic diversity in metapopulations. *B Math Biol* **2021**, *83*, 20. https://doi.org/10.1007/s11538-020-00853-5
22. Patil, G.P.; Taillie, C. Diversity as a concept and its measurement. *J Am Stat Assoc* **1982**, *77*(379), 548-567.
23. Xu, S.; Peskin, C. S. The impact of universal recycling on the evolution of economic diversity. *PLoS ONE* **2022**, *17*(1), e0262184. https://doi.org/10.1371/journal.pone.0262184
24. Nourbakhsh, M.; Yari, G. Weighted Rényi's entropy for lifetime distributions. *Commun Stat-Theor M* **2017**, *46*(14), 7085-7098. https://doi.org/10.1080/03610926.2016.1148729
25. Lyons, N. I. Comparing diversity indices based on counts weighted by biomass or other importance values. *Am Nat* **1981**,*118*(3), 438-442. https://doi.org/10.1086/283836
26. Willis, A. D.; Martin, B. D. Estimating diversity in networked ecological communities. *Biostatistics* **2022**, *23*(1), 207-222. https://doi.org/10.1093/biostatistics/kxaa015
27. Heath, T. L. *A History of Greek Mathematics* (Vol. 1). Clarendon Press: Oxford, England, 1921.
28. Komić, J. Harmonic Mean. In *International Encyclopedia of Statistical Science*; Lovric, M., Eds.; Springer: Berlin, Heidelberg, Germany, 2011, pp. 622-624. https://doi.org/10.1007/978-3-642-04898-2_645
29. Dodge, Y. *The Concise Encyclopedia of Statistics*; New York: Springer, 2008, pp. 239-241. https://doi.org/10.1007/978-0-387-32833-1
30. Casquilho, J. P. On the weighted Gini-Simpson index: Estimating feasible weights using the optimal point and discussing a link with possibility theory. *Soft Comput* **2020**, *24*, 17187–17194. https://doi.org/10.1007/s00500-020-05011-6