Lorenz curves determine partial orders for diversity and similarity measurement in ecological applications
-----------------------------------------------

Ronald Rousseau

KHBO (Association K.U. Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400, Oostende, Belgium

& University of Antwerp, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
Email: ronald.rousseau@khbo.be

Abstract

Problems with the notions of diversity and evenness can be overcome by a more fundamental, mathematical approach. Such an approach is based on the Lorenz curve, of which different forms exist. Each Lorenz curve determines a partial order in the corresponding set of abundances (in general: set of items). A Gini index can be associated to each of these curves. The following aspects of evenness, inequality and diversity, each leading to a different Lorenz curve, and resulting in a different partial order, are studied:

- classical Lorenz curves and evenness measures
- intrinsic diversity profiles (or k-dominance curves) and associated measures of diversity
- generalised Lorenz curves, as introduced by Shorrocks, that are not scale invariant but take absolute numbers of abundances into account
- weighted Lorenz curves for comparisons with an internal or external standard
- a recently introduced Lorenz curve used for a source per source comparison of items.

It is further shown that Lorenz curves can be used for similarity measurement in the case of absence-presence data. Then the corresponding Gini index is nothing but the Jaccard similarity index.

We conclude that the Lorenz curve and Gini index are universal tools for studying many types of diversity. It is important though to be precise in stating the aim of diversity or evenness measurements, and carefully choose the corresponding Lorenz curve and Gini index.
1. Introduction

It is well known that the earth’s biodiversity is rapidly declining. In order to gauge this decline, and to evaluate the effectiveness of countermeasures it is of the utmost importance to measure biodiversity in an exact way. Unfortunately, there is no agreement on how to do this.

Precise definitions and a logical approach are called for. Moreover, we are convinced that different approaches, studying various aspects of biodiversity, are necessary. One must, however, be sure to use logically sound approaches. For instance, proposing diversity indices without a clear knowledge of what they are measuring and to which kind of ranking they give rise is certainly a bad idea (Rousseau & Van Hecke, 1999).

We will only discuss complete communities, studied at a precise moment in time. Dynamical aspects and sampling properties such as estimation problems fall outside the scope of this note.

2. Evenness

Species richness and evenness are two important notions related to biodiversity. Species richness refers to the number of different species present and is theoretically unambiguous. In practice, however, there might be serious difficulties as to when to consider two species as different, or whether or not to include extremely rare species.

In any case, evenness is an even more subtle notion. We define it as the relative apportionment of abundances among species, actually present or assumed to be present. Note that abundances can be expressed in terms of biomass, individuals or any other quantity appropriate in a particular investigation. We agree with Taillie (1979) and Gosselin (2001) that the classical Lorenz curve is the perfect representation for evenness ranking, being independent of the number of species concerned. Besides giving rise to a partial order satisfying the permutation invariance and the scale invariance property, this ranking also satisfies Dalton’s transfer principle and the replication property. We give a short explanation of these properties.

A species abundance array gives rise to precisely one classical Lorenz curve (while the opposite is not true at all). We assume that the reader is familiar with classical Lorenz curves (see Figure 1). We just recall that there are two versions: a concave one and a convex one. For the convex form species’ abundances are ranked from low to high; ranking from high to low yields the concave variant. These two approaches are mathematically equivalent. Considering the convex
form, we obtain a partial order where a Lorenz curve corresponds to a higher degree of evenness if it is situated above the other one. Highest evenness occurs for the diagonal line, corresponding with perfect evenness. As Lorenz curves may intersect, this construction yields a partial (not a complete or total) order, referred to as the Lorenz dominance order for abundance arrays. Permutation invariance refers to the fact that evenness is not a property of individuals but of a community as a whole. Scale invariance means that abundance arrays differing by a proportionality factor are considered equivalent. As Lorenz curves are drawn using proportions they are automatically scale invariant. This is an interesting property as evenness for abundances of biomass weighted in pounds or weighted in kilos must be the same. Dalton’s transfer principle (Dalton, 1920) states that when the biomass of a less abundant species decreases in favour of an already more abundant species, evenness must decrease. It has been shown that the Lorenz order meets this transfer principle. Finally, the replication property, also due to Dalton, says that the evenness of any replication of a community must be equal to the evenness of the original community. In other words: the evenness of the abundance array (2,3,5,8) is equal to that of (2,3,5,8,2,3,5,8) = (2,2,3,3,5,5,8,8), which is equal to that of (2,2,2,2,3,3,3,3,5,5,5,5,8,8,8,8,8). As replication has no influence on the Lorenz curve also this requirement is clearly satisfied.

Fig 1: Lorenz curve for evenness: X denotes the Lorenz curve of the array (1, 3, 5, 10), while D denotes the diagonal, representing perfect evenness.
An evenness function is a function that associates a non-negative number to each abundance array. It must moreover respect the Lorenz partial order. Examples of such functions are (Nijssen et al., 1998):

**The Gini evenness index: \( G' \)**

This index is equal to twice the area under the convex Lorenz curve, or 1 minus twice the area between the Lorenz curve and the diagonal. If \( \mathbf{X} = (x_j)_{j=1,...,N} \) denotes an abundance array, the Gini evenness index is calculated as:

\[
G'(\mathbf{X}) = \frac{2}{\mu N^2} \left( \sum_{j=1}^{N} (N+1-j) x_j \right) - \frac{1}{N}
\]

where the \( x_j \)'s are ranked from low to high and \( \mu \) denotes the mean of the set \( \{x_j\} \).

**The modified Simpson index: \( \Lambda \)**

The modified Simpson index is defined as:

\[
\Lambda(\mathbf{X}) = \frac{1}{N \sum_{j=1}^{N} \left( \frac{x_j}{S} \right)^2}
\]

where \( S \) denotes the sum of the \( \mathbf{X} \)-array: \( S = \sum_{j=1}^{N} x_j \).

**The reciprocal of the coefficient of variation: \( 1/V \)**

This evenness measure is defined as \( \frac{1}{V} = \frac{\mu}{\sigma} \), where \( \sigma \) denotes the standard deviation of the set \( \{x_j\} \) (the components of the abundance array \( \mathbf{X} \)).

Finally, we also mention the adapted Shannon-Wiener index: \( H' \)

\[
H'(\mathbf{X}) = \frac{1}{\ln(N) + \sum_{j=1}^{N} \left( \frac{x_j}{S} \ln \left( \frac{x_j}{S} \right) \right)}
\]
Note though that this index is not defined (or equal to $\infty$) for the perfect evenness situation. The Brillouin index (Magurran, 1991) does not satisfy the requirements for an evenness measure and hence must not be used here.

We conclude this part by stressing that striving for a total order representing evenness is a futile endeavour. A partial order is the best one can achieve.

3. Intrinsic diversity profiles: combining evenness with species richness

It is generally agreed (but see section 4) that diversity combines two aspects: species richness and evenness. Disagreement arises as to how these two aspects should be combined, and how to find a measure describing this combination. We have shown (Rousseau et al., 1999) that if species abundance is never zero, then the intrinsic diversity profiles, introduced by Patil and Taillie (1979), are the best possible. Under the term ‘best possible’ we mean that the resulting partial order is the strongest one among a number of reasonable candidates (Rousseau et al., 1999). Note also that intrinsic diversity profiles are equivalent to k-dominance curves as introduced by Lambshead et al. (1981). These curves can be drawn as follows: abundances are ranked from high to low. On the abscissa the cumulative number of species is shown (not normalised!). As ordinates the cumulative proportion of abundances is used. Figure 2 illustrates the concept. These profiles introduce a partial order denoted here as $\prec\prec$. If $X \prec\prec X_1$ then the intrinsic diversity profile of $X$ never lies under the profile for $X_1$. In this partial order perfect evenness for $N$ species is less diverse than perfect evenness for $N + 1$ species (while their evenness is the same).
Fig. 2 Intrinsic diversity profiles for two arrays with the same evenness: 
\( X = (1,3,5,10) = (10,5,3,1) \) and \( X_1 = (10,10,5,5,3,3,1,1) \); \( X \ll X_1 \)

A measure respecting the partial order for intrinsic diversity profiles (\( \ll \)) is obtained by taking twice the area between the intrinsic diversity profile, the line \( x = 0 \) (the ordinate axis) and the line \( y = 1 \). We call this measure the adapted Gini coefficient (AG). It is calculated as:

\[
AG(X) = \frac{2}{S} \left( \sum_{j=1}^{N} j x_j \right) - 1
\]

where \( X = (x_1, x_2, \ldots, x_N) \) and the \( x_j \)'s are ordered from high to low; \( S \) is the sum of the \( x_j \)'s as before. Applying an atan-transformation yields a normalised function (a function taking values between 0 and 1).

Recall that we required that no class has a species abundance of zero otherwise problems arise with the construction of intrinsic diversity profiles. It is indeed, known that, unless zero-classes are accepted, problems are to be expected (Routledge, 1983; Ricotta, 2004). We, moreover, claim that zero-classes may occur in practice.
Indeed, when studying a specified number of plant species in a fixed number of (experimental or even natural) sites, it may happen that some species disappear during the experiment, or are simply not present. Consider now, for instance, $X = (4, 2, 1, 0)$ and $Y = (4, 2, 1)$. Their adapted Gini coefficients are the same and the larger species richness of $X$ is only apparent (mathematically it is a 4-array, but in reality it is only a 3-array). The evenness of $X$ is strictly smaller than that of $Y$; hence one would expect that the intrinsic diversity of $Y$ is larger than that of $X$. This does not happen, showing that intrinsic diversity profiles cannot cope with null classes. Another variation on the Lorenz curve theme, namely shifted Lorenz curves does not have this drawback. Fig. 3 illustrates this concept. Here the Lorenz evenness curve is essentially shifted 1/2N to the right. For more details about this construction, the reader is referred to (Rousseau, 1992).

Fig.3 Shifted Lorenz curves of the diagonal (denoted as DS) and $X = (1, 3, 5, 10)$, denoted as XS.

4. Shorrocks' generalised Lorenz curves take abundances into account

In ecological studies not only the number of species and the relative abundances among species are important, but also the abundances themselves. The previous approaches do not take this aspect into account. Economists studying household incomes face the same problem. Besides the number of households and the inequality between their incomes, also the absolute income plays a role.
In order to cope with this problem Shorrocks (1983) proposed the use of generalised Lorenz curves - in (Rousseau et al., 1999) we referred to these curves as intrinsic stability profiles. A generalised Lorenz curve is simply a convex Lorenz curve where each ordinate value is multiplied by the average abundance, \( \mu \) (average income in Shorrocks’ case). This yields a new partial order where situations with a higher average number of specimen (abundances) are considered to be more diverse than those with a lower average number of specimens. This approach clearly has the disadvantage that it is not anymore scale invariant, but in cases where the counting measure is used this is of no importance. Also here twice the area under the convex generalised Lorenz curve is an acceptable diversity measure of Gini-type. This Gini measure, however, is not anymore bounded by one.

Fig. 4 Generalised Lorenz curves: B corresponds to the array (1, 3, 5, 10) with an average of 19/4, while A corresponds to the array (2, 6, 10, 20) with an average of 38/4=19/2.
5. Comparison with an average or hypothesized situation: weighted Lorenz curves

Sometimes one is not interested in the diversity of abundances over sources, but it is more interesting to study how different the diversity is with respect to an average or a hypothesized situation. An average can be considered as an internal standard, a hypothesized situation as an external standard. Often this external standard is a factor hypothesized to be the main factor in explaining observed differences in diversity.

Consider the following hypothetical table of cross-classified data (Table 1): a row classification consisting of M rows and a column classification consisting of N columns, with M, N ≥ 2, cf. (Rousseau, 2001). Let us assume that columns are abundance arrays of key species, while rows refer to different regions.

Table 1 Hypothetical cross tabulation (3 x 4) of species and regions

<table>
<thead>
<tr>
<th>Species</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>Average: R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>20</td>
<td>140</td>
<td>200</td>
<td>140</td>
<td>125</td>
</tr>
<tr>
<td>R2</td>
<td>25</td>
<td>10</td>
<td>20</td>
<td>95</td>
<td>37.5</td>
</tr>
<tr>
<td>R3</td>
<td>15</td>
<td>90</td>
<td>80</td>
<td>65</td>
<td>62.5</td>
</tr>
<tr>
<td>Average</td>
<td>S</td>
<td>20</td>
<td>80</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Given the data of Table 1 one may be interested in the question of how the evenness of the species abundance array of a particular region compares with the average species abundance array. This average array plays then the role of an internal standard. On the other hand, one might assume that average rainfall is the main factor explaining the observed difference in occurrence between regions of species 3. In that case one may use the 3-array of average rainfall and compare it to the S3-column. Here the rainfall array acts as an external standard. In both cases (external or internal standard) one uses weighted Lorenz curves and measures respecting the partial order determined by these weighted Lorenz curves.

Weighted Lorenz curves are constructed as follows (Theil, 1967; Patil and Taillie, 1982). Let \( S = (s_1, s_2, \ldots, s_N) \) denote the standard array and let \( X = (x_1, x_2, \ldots, x_N) \) denote the studied array. This array is the one we want to compare with the standard. Note that indices must correspond and that the studied array and the standard one have in a natural way the same number of cells (or sources). If, e.g., \( X \) denotes the abundance of a particular species in different regions and \( S \) denotes rainfall, then \( x_j \) and \( s_j \) must refer to the same region. We assume, moreover, that none of the components of the standard array \( S \) is zero. In order to construct the weighted Lorenz curve for comparison with a standard the components of both arrays are first ordered in such a way that
\[
\frac{x_1}{s_1} \geq \frac{x_2}{s_2} \geq \cdots \geq \frac{x_N}{s_N} \tag{5}
\]

This ranking leads to a concave curve. Using the other one leads to a convex representation. Next we normalise the arrays \( S \) and \( X \), leading to arrays \( W \) and \( A_X \), where components are determined as:

\[
a_j = \frac{x_j}{\sum_{k=1}^{N} x_k} \quad \text{and} \quad w_j = \frac{s_j}{\sum_{k=1}^{N} s_k} \tag{6}
\]

Note that normalising does not change the order. Finally, the weighted Lorenz curve is defined as the broken line connecting the origin (0, 0) to the points with components

\[
\left( \sum_{k=1}^{N} w_k, \sum_{k=1}^{N} a_k \right)_{j=1, \ldots, N} \tag{7}
\]

For a fixed standard, these weighted Lorenz curves again introduce a partial order in the set of N-arrays.

Fig. 5 Weighted Lorenz curve (concave form) for the array \( X = (1, 3, 5, 10) \) compared with \( S = (1, 1, 7, 1) \); after re-ranking and normalising we have drawn the curve based on \( A_X = (10/19, 3/19, 1/19, 5/19) \) and \( W = (1/10, 1/10, 1/10, 7/10) \).
Functions that respect this partial order are sometimes referred to as measures of asymmetric relative diversity (Egghe & Rousseau, 2001). The term ‘relative’ refers to the fact that one compares with a standard. The term ‘asymmetric’ stresses the fact that the roles of the standard and the array under study cannot be interchanged. Examples of such measures are:

a) The weighted Gini diversity index:

\[ G_w'(X) = 1 - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_i a_j - w_j a_i \] (8)

The interpretation of this index is the same as that for the (unweighted) Gini index, namely twice the area above the weighted Lorenz curve. When the weighted Lorenz curve coincides with the diagonal, the weighted Gini diversity index is 1.

b) The asymmetric (or weighted) inverse squared coefficient of variation

\[ V_w^2(X) = \frac{1}{\sum_{j=1}^{N} (a_j - w_j)^2} - 1 \] (9)

This measure takes values between 0 and infinity.

6. Source per source comparison of abundances: absolute differences

In this approach, introduced by Egghe and Rousseau (2001) one directly compares relative arrays. Relative contributions of the same sources (species), but e.g. at different times are compared by taking differences. These differences may be positive or negative, and one compares with the zero-array.

A Lorenz curve adapted to this situation has been constructed as follows. Let \( X = (x_j)_{j=1,...,N} \) and \( Y = (y_j)_{j=1,...,N} \) be two \( N \)-arrays and let \( A = (a_j)_{j=1,...,N} \) and \( B = (b_j)_{j=1,...,N} \) denote their relative arrays (sum of all components equal to one). Then the components of the difference array \( D = (d_j)_{j=1,...,N} \) with \( d_j = a_j - b_j \) are ranked from largest to smallest. Finally, one puts

\[ t_j = \sum_{k=1}^{j} d_k = \sum_{k=1}^{j} (a_k - b_k) \] (10)
The corresponding Lorenz curve is obtained by joining the origin \((0, 0)\) with the points with coordinates

\[
\left( \frac{j}{N}, t_j \right)_{j=1,\ldots,N}
\]

Note that this curve always ends in the point \((1, 0)\). For this reason one may call it the Lorenz bridge (my own proposal). See Fig. 6 for an example.

![Lorenz bridge](image)

**Fig. 6** Lorenz bridge for the arrays \(B = (0.6, 0.2, 0.1, 0.1); A = (0.1, 0.2, 0.2, 0.5)\)

Similar to the other Lorenz curves also the Lorenz bridge leads to a partial order and functions respecting this partial order are the ones we are interested in. Such functions are sometimes referred to as measures of symmetric relative diversity (or concentration). Here the term ‘relative’ again refers to the fact that one compares with a standard (the zero line). Examples of such measures are:

- the Gini diversity (bridge) function:
where the $d_j$ are ranked in decreasing order. This is nothing but the area of the unit square (namely 1) minus the area under the Lorenz bridge.

- another acceptable measure of symmetric relative diversity is the adapted Simpson index of relative difference:

$$V_r^2 = \frac{1}{N} \sum_{j=1}^{N} d_j^2$$  \hspace{1cm} (13)

It is related to the squared coefficient of variation, hence the notation $V_r^2$. We end these sections on Lorenz curves and diversity measures by observing that the exact relation between different forms of Lorenz curves and the appropriate measures follows from a general mathematical theory put forward by Egghe (2002).

7. Diversity for absence-presence data: the Jaccard index and its relation to the Lorenz curve

Using presence-absence data is a basic approach for representing categorical data. As presence is often represented by 1 and absence by 0, such data are also known as 1 - 0 data. Similarity between sites may be determined by comparing absence-presence site representations. In ecological studies (as in information retrieval) it is often appropriate not to consider common zeros. Indeed, species that do not occur in at least one of the sites under investigation do not make the sites more similar. Sites situated in Western Europe are not more similar if they both do not contain baobabs. For this reason we will refer to this approach as the zero insensitive case. This will be the only case explained here.

**Lorenz similarity curves**

Let $r$ and $s$ be two presence-absence arrays of length $N$ (in short: N-arrays) corresponding to two sites $r$ and $s$. The similarity of $D = \{r, s\}$ must not depend on the order in which we consider $r$ and $s$. It must, moreover, not depend on the order in which the species present in $r$ and/or $s$ are enumerated. Of course, in practice one uses a particular order (always the same for the two items involved) but the point is that this must not influence their similarity. These requirements
are also imposed in diversity studies, hence it seems natural to consider Lorenz curves (Lorenz, 1905) in a similar vain as for studies in diversity. In order to emphasize the fact that it is irrelevant in which order document representations for similarity studies are considered we refer to \( D = \{r, s\} \) as a duo, a word that has no “rank” connotations.

**Construction of Lorenz curves for duo similarity**

Let \( r = (x_i)_{i=1, \ldots, N} \) and \( s = (y_i)_{i=1, \ldots, N} \) be two presence-absence N-arrays of the sites \( r \) and \( s \) and let \( (a_i)_{i=1, \ldots, N} \) and \( (b_i)_{i=1, \ldots, N} \) denote their relative arrays. This means that each value is divided by the total sum, and hence the new sum of all components becomes equal to one. Then we constrict the Lorenz bridge of these arrays. These Lorenz bridges again introduce a partial order. If the length of an array is fixed, say N, then the ordinate of the top is at most one; its abscissa may be any of the values \( \{1/N, 2/N, \ldots, (N-1)/N\} \). These abscissas lead to \( N-1 \) different intersecting curves. Consequently there is no minimum intrinsic similarity curve, even when \( N \) is fixed (unless \( N = 2 \)), and certainly not for variable \( N \). All these curves are called (in this context) minimal Lorenz bridges.

In the context of this article we define the Gini similarity measure, denoted as \( G_S \), as one minus twice the area under the Lorenz similarity bridge. This normalises the Gini similarity measure in such a way that all minimal Lorenz similarity curves correspond to a Gini-value of zero, and the equality line has a similarity value of one. Using this normalisation, the Gini similarity measure coincides with the Jaccard index as shown in (Egghe & Rousseau, 2005). Recall that the Jaccard index is equal to the number of species that the two sites have in common, divided by the number of species present in at least one of the two sites (Magurran, 1991).

Another example of an acceptable similarity measure is

\[
V^2_S(D) = V^2_S(r, s) = \frac{1}{N \sum_{i=1}^N d_i^2}
\]  

(14)

This is the adapted Simpson index of the relative difference array, see (13). The factor \( \sum_{i=1}^N d_i^2 \) can be interpreted as the number of unique species (species that are either unique to \( r \) or unique to \( s \)) divided by the product of the number of species in \( r \) and the number of species in \( s \). In this form the adapted Simpson index has the drawback that if \( r = s \) the similarity value becomes \(+\infty\). The Simpson similarity of minimal Lorenz curves tends to zero as \( N \) increases.
We further observe that replacing an element unique to \( r \) by an element unique to \( s \) leaves the Jaccard measure invariant, but not Simpson’s measure. In this sense the latter is more sensitive.

8. Conclusion

We conclude that the Lorenz curve and Gini index are universal tools for studying many forms of diversity. It is important though to be precise in stating the aim of diversity or evenness measurements, and carefully choose the corresponding Lorenz curve and Gini index (Rousseau, 2001).

Acknowledgement

The author thanks Jan Bogaert (ULB, Belgium) for useful advice during the preparation of this article.

References


