Which Bio-Diversity Indices Are Most Adequate

Olga Kosheleva\textsuperscript{1}, Craig Tweedie\textsuperscript{2},
and Vladik Kreinovich\textsuperscript{3}

\textsuperscript{1}Department of Teacher Education
\textsuperscript{2}Environmental Science and Engineering Program
\textsuperscript{3}Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
olgak@utep.edu, ctweedie@utep.edu
vladik@utep.edu
1. Gauging Bio-Diversity Is Important

- One of the main objectives of ecology is to study and preserve bio-diversity.
- Most existing measures of diversity are based on the relative frequencies $p_i$ of different species.
- The most widely used measures are:
  - the Shannon index $H = -\sum_{i=1}^{n} p_i \cdot \ln(p_i)$, and
  - the Simpson index $D = \sum_{i=1}^{n} p_i^2$.
- Ecologists also use indices related to $D$: $\frac{1}{D}$ and $1 - D$.
- They also use indices related to the sum $\sum_{i=1}^{n} p_i^q$, such as
  Rényi entropy $H_q = \frac{1}{1 - q} \cdot \ln \left( \sum_{i=1}^{n} p_i^q \right)$. 
2. Why These Measures?

- The above measures of diversity are, empirically, in good accordance with the ecologists’ intuition.

- However, from the theoretical viewpoint, the success of these measures of diversity is somewhat puzzling.

- Why these expressions and not other possible expressions?

- In this talk, we provide possible justification for the above empirically successful measures.

- We provide two possible justification:
  - we start with a simple fuzzy logic-based justification which explains Simpson index, and then
  - we provide a more elaborate justification that explains all the above diversity measures.
3. An Intuitive Meaning of Bio-Diversity

• An ecosystem is perfectly diverse if all its species are reasonably frequent but not dominant.

• In other words, the ecosystem is healthy if:
  – the first species is reasonably frequent but not dominant, and
  – the second species is reasonably frequent but not dominant,
  – etc.

• This statement uses an imprecise (“fuzzy”) natural-language terms like “reasonably frequent”.

• We need to translate this statement into precise terms.

• We will use fuzzy logic, since fuzzy logic was invented exactly for such a translation.
4. Let Us Use Fuzzy Logic

- For each $p_i$, let $\mu(p_i)$ be the degree to which the species is reasonably frequent and not dominant.
- To compute bio-diversity, we need to combine use “and”-operation to combine these degrees.
- The general strategy in applications of fuzzy techniques is to select the simplest possible “and”-operation.
- The two simplest (and most frequently used) “and”-operations are the product and the minimum.
- Our objective is to optimize bio-diversity, and most efficient optimization techniques use differentiation.
- From this viewpoint, it is desirable to come up with the differentiable measure of diversity.
- This eliminates min (since $\min(a, b)$ is not differentiable when $a = b$), so we use the product $\prod_{i=1}^{n} \mu(p_i)$.
5. Let Us Use Fuzzy Logic (cont-d)

- Maximizing \( \prod_{i=1}^{n} \mu(p_i) \) is equivalent to maximizing its logarithm \( L = \sum_{i=1}^{n} f(p_i) \), where \( f(p_i) \equiv \ln(\mu(p_i)) \).

- In a diverse ecosystem all the frequencies \( p_i \) are rather small.

- Indeed, if one of the values is large, this means that we have a dominant species, not a diversity.

- For small \( p_i \), we can replace each value \( f(p_i) \) with the sum of the few first terms in its Taylor expansion.

- In the first approximation, \( f(p_i) = a_0 + a_1 \cdot p_i \), so

\[
L = a_0 \cdot n + a_1.
\]

- This expression does not depend on the frequencies \( p_i \) and thus, cannot serve as a measure of diversity.
6. Fuzzy Logic Justifies the Simpson Index

- We want to maximize \( L = \sum_{i=1}^{n} f(p_i) \).

- We have shown that linear terms in \( f(p_i) \) are not sufficient.

- So, to adequately describe diversity, we need to take into account quadratic terms

\[
f(p_i) = a_0 + a_1 \cdot p_i + a_2 \cdot p_i^2.
\]

- In this approximation,

\[
L = a_0 \cdot n + a_1 + a_2 \cdot \sum_{i=1}^{n} p_i^2.
\]

- Maximizing this expression is equivalent to maximizing the Simpson index \( D = \sum_{i=1}^{n} p_i^2 \).
7. The Ultimate Purpose of Diversity Estimation Is Decision Making

- We want to describe which combinations of frequencies \( p = (p_1, \ldots, p_n) \) are preferred and which are not.

- Most plans succeed only with a certain probability.

- So, we need to consider “lotteries”, in which different combinations \( A_i \) appear with different probabilities \( P_i \).

- The main result of utility theory states that:
  - if we have a consistent ordering relation \( L \succeq L' \) (“\( L \) is preferable to \( L' \)”) between such lotteries,
  - then there exists a function \( u \) (called utility) s.t.
    \[
    L \succeq L' \text{ if and only if } u(L) \geq u(L'),
    \]
    where
    \[
    u(L) = P_1 \cdot u(A_1) + \ldots + P_n \cdot u(A_n).
    \]

- In our case, we need a utility function \( u(p_1, \ldots, p_n) \).
8. Bio-Diversity of Subsystems

- An important intuitive feature of bio-diversity is the *localness* property:
  - that, in addition to the bio-diversity of the whole ecosystem,
  - we may be interested in the bio-diversity of its subsystems.

- For the whole ecosystem, the sum of frequencies is 1.

- When we analyze a subsystem, we only take into account some of the species.

- So the sum of the frequencies can be smaller than 1.

- Thus, we need to consider the values $u(p)$ for tuples for which $\sum_i p_i < 1$.

- It makes sense to compare possible arrangements within a subsystem.
9. Localness Property

- We only compare tuples \( p = (p_1, \ldots, p_n) \) and \( p' = (p'_1, \ldots, p'_n) \) for which \( \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} p'_i \).

- Let us assume that for all species \( i \) from some set \( I \), the frequencies are the same: \( p_i = p'_i \).

- Suppose also that, from the point of bio-diversity, the tuple \( p \) is preferable to tuple \( p' \): \( p \succeq p' \); so:
  - while in the two tuples, the level of diversity is the same for species from the set \( I \),
  - species from the complement set \(-I\) have a higher degree of bio-diversity.

- Thus, if we replace the values \( p_i = p'_i \) for \( i \in I \) with some other values \( q_i = q'_i \), we will still have \( q \succeq q' \).
10. Localness Property in Precise Terms

- **Localness property:**
  - Let $I \subseteq \{1, \ldots, n\}$ be a set of indices.
  - Let $p \succeq p'$ be two tuples s.t. $p_i = p'_i$ for all $i \in I$.
  - Let $q$ and $q'$ be another two tuples for which:
    - $q_i = p_i$ and $q'_i = p'_i$ for all $i \notin I$; and
    - $q_i = q'_i$ for all $i \in I$.
  - Then, $q \succeq q'$.

- **It is known:** in this case, utility has the form
  
  $u(p_1, \ldots, p_n) = \sum_{i=1}^{n} u_i(p_i)$ or $U(p_1, \ldots, p_n) = \prod_{i=1}^{n} U_i(p_i)$.

- Maximizing the product is equivalent to maximizing its logarithm $\sum \ln(U_i(p_i))$.

- So, w.l.o.g., we can assume that $u = \sum_{i=1}^{n} u_i(p_i)$. 
11. The Degree of Bio-Diversity Should Not Change If We Rename the Species

- Numbers assigned to species – which species is number 1, which is number 2, etc. – are arbitrary.

- So, if we simply change these arbitrarily selected numbers, the degree of bio-diversity should not change.

- Thus, the dependence of \( u_i \) on \( p_i \) should not depend on \( i \).

- So, we should have \( u_i(p_i) = d(p_i) \) for one and the same function \( d(p) \).

- In this case, the desired degree of bio-diversity is equal to \( u(p) = \sum_{i=1}^{n} d(p_i) \).

- So, the question is which functions \( d(p) \) are appropriate for describing bio-diversity.
12. Without Losing Generality, We Can Assume That the Function $d(p)$ is Twice Differentiable

- Our ultimate goal is optimization.
- Many useful optimization techniques use second derivatives.
- So, it is desirable to consider only *twice* differentiable functions.
- Every continuous function can be:
  - with an arbitrary accuracy,
  - approximated by twice differentiable functions (even by polynomials).
- So, we can assume that $d(p)$ is twice differentiable without losing generality.
13. Possibility of Scaling

- Relative bio-diversity of a region should not depend on:
  - whether we consider it as a separate ecosystem,
  - or we consider it as a part of a larger ecosystem.

- When we consider an ecosystem by itself, the frequencies add up to 1: \( \sum_{i=1}^{n} p_i = 1 \).

- When we consider it as a part of a larger ecosystem with \( N > n \) species, we get \( p'_i = \lambda \cdot p_i \), where \( \lambda = \frac{n}{N} \).

- Thus, if we have \( p \succeq p' \), we should also have \( \lambda \cdot p \succeq \lambda \cdot p' \).

- We say that a twice differentiable function \( d(p) \) is scale-invariant if \( \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} p'_i \) and \( \sum_{i=1}^{n} d(p_i) = \sum_{i=1}^{n} d(p'_i) \) imply
  \[
  \sum_{i=1}^{n} d(\lambda \cdot p_i) = \sum_{i=1}^{n} d(\lambda \cdot p'_i).
  \]
14. Main Result

- **Definition.** We say that functions $d_i(p)$ are equivalent if $d_2(p) = a + b \cdot p + c \cdot d_1(p)$.

- **Motivation.** In this case, optimizing $\sum d_2(p_i)$ is equivalent to optimizing $\sum d_1(p_i)$.

- **Theorem.** Every scale-invariant function $d(p)$ is equivalent:
  - either to $d(p) = \pm \ln(p)$,
  - or to $d(p) = \pm p^q$ for some $q$, or
  - or to $d(p) = \pm p \cdot \ln(p)$.

- **Observation.** The corresponding sums are exactly Shannon, Simpson, and Rényi indices.

- **Conclusion.** We have explained why only these indices adequately describe bio-diversity.
15. Conclusions

- One of the main goals of ecology is to maintain bio-diversity.
- To properly maintain bio-diversity, it is important to adequately gauge it.
- Several semi-heuristic measures have been proposed for measuring bio-diversity.
- Their successful use confirms that these measures adequately reflect our ideas of bio-diversity.
- In this talk, we provide a fuzzy-motivated theoretical explanation for the existing bio-diversity indices.
16. Acknowledgment

This work was supported in part by the National Science Foundation grants:

- HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
- DUE-0926721.
17. **Proof: Differential Equation**

- For small deviations $p'_i = p_i + \varepsilon \cdot \Delta p_i$, localness means that $\sum_{i=1}^{n} \Delta p_i = 0$ and $\sum_{i=1}^{n} d'(p_i) \cdot \Delta p_i = 0$ imply

  $$\sum_{i=1}^{n} d'(\lambda \cdot p_i) \cdot \Delta p_i = 0.$$

- Let $e = (1, \ldots, 1)$, $d' = (d'(p_1), \ldots)$, and $d'_\lambda = (d'(\lambda \cdot p_1), \ldots)$.

- Localness means that if $e \cdot \Delta p = 0$ and $d' \cdot \Delta p = 0$, then $d'_\lambda \cdot \Delta p = 0$.

- One can see that in this case, $d'_\lambda$ is in the linear space spanned by $e$ and $d'$: $d'(\lambda \cdot p_i) = \alpha(\lambda, p) + \beta(\lambda, p) \cdot d'(p_i)$.

- Let us show that the values $\alpha$ and $\beta$ depend only on $\lambda$ and do not depend on the tuple $p$. 


18. \textbf{Proof: $\beta$ Does Not Depend on $p$}

- We have proven that $d'(\lambda \cdot p_i) = \alpha(\lambda, p) + \beta(\lambda, p) \cdot d'(p_i)$.

- If we subtract the equations corresponding to two different indices $i$ and $j$, we conclude that

$$d'(\lambda \cdot p_i) - d'(\lambda \cdot p_j) = \beta(\lambda, p) \cdot (d'(p_i) - d'(p_j)).$$

- Thus, $\beta(\lambda, p) = \frac{d'(\lambda \cdot p_i) - d'(\lambda \cdot p_j)}{d'(p_i) - d'(p_j)}$.

- The right-hand side of this equality only depends on $p_i$ and $p_j$ and does not depend on any other $p_k$.

- Thus, the coefficient $\beta(\lambda, p)$ only depends on $p_i$ and $p_j$ and does not depend on any other frequencies $p_k$.

- For a different pair $(i', j')$, we will conclude that $\beta(\lambda, p)$ does not depend on the frequencies $p_i$ and $p_j$ either.

- Thus, $\beta$ does not depend on the tuple $p$ at all.
19. **Proof: \( \beta \) Is Differentiable**

- \( \beta(\lambda, p) = \beta(\lambda) \), so \( d'(\lambda \cdot p_i) = \alpha(\lambda, p) + \beta(\lambda) \cdot d'(p_i) \).
- Hence, \( \alpha(\lambda, p) = d'(\lambda \cdot p_i) - \beta(\lambda) \cdot d'(p_i) \).
- The right-hand side of this formula only depends on \( p_i \) and does not depend on any other frequency \( p_j \).
- Thus, the coefficient \( \alpha(\lambda, p) \) only depends on \( p_i \) and does not depend on any other frequency \( p_j \).
- For a different index \( i' \), we will conclude that \( \alpha(\lambda, p) \) does not depend on the frequency \( p_i \) either.
- Thus, \( \alpha \) does not depend on the tuple \( p \) at all, it only depends on \( \lambda \): \( d'(\lambda \cdot p_i) = \alpha(\lambda) + \beta(\lambda) \cdot d'(p_i) \).
- For \( D(p) \overset{\text{def}}{=} d'(p) \), we get \( D(\lambda \cdot p_i) = \alpha(\lambda) + \beta(\lambda) \cdot D(p_i) \).
- Since \( \beta(\lambda) = \frac{d'(\lambda \cdot p_i) - d'(\lambda \cdot p_j)}{d'(p_i) - d'(p_j)} \), the function \( \beta(\lambda) \) is differentiable.
20. Proof: Final Part

- Since \( \alpha(\lambda) = d'(\lambda \cdot p_i) - \beta(\lambda) \cdot d'(p_i) \) and \( \beta(\lambda) \) is differentiable, the function \( \alpha(\lambda) \) is also differentiable.
- Differentiating \( D(\lambda \cdot p_i) = \alpha(\lambda) + \beta(\lambda) \cdot D(p_i) \) w.r.t. \( \lambda \) and taking \( \lambda = 1 \), we get
  \[
p \cdot \frac{dD}{dp} = A + B \cdot D.
  \]
- Separating variables, we get \( \frac{dD}{A + B \cdot D} = \frac{dp}{p} \).
- For \( B = 0 \), we get \( D(p) = d'(p) = A \cdot \ln(p) + C \).
- In this case, \( d(p) \) is equivalent to \( p \cdot \log(p) \).
- For \( B \neq 0 \), we get \( d'(p) = D(p) = C \cdot p^A + C' \).
- In this case, \( d(p) \) is equivalent to \( p^q \) or to \( \ln(p) \).