Bulletin of Mathematical Biology (2023) 85:78
https://doi.org/10.1007/s11538-023-01183-y

ORIGINAL ARTICLE

Spaces of Phylogenetic Diversity Indices: Combinatorial and Geometric Properties

Kerry Manson[1] · **Mike Steel**[1]

Received: 2 March 2023 / Accepted: 20 June 2023 / Published online: 7 July 2023

Abstract

Biodiversity is a concept most naturally quantified and measured across sets of species. However, for some applications, such as prioritising species for conservation efforts, a species-by-species approach is desirable. Phylogenetic diversity indices are functions that apportion the total biodiversity value of a set of species across its constituent members. As such, they aim to measure each species' individual contribution to, and embodiment of, the diversity present in that set. However, no clear definition exists that encompasses the diversity indices in current use. This paper presents conditions that define diversity indices arising from the phylogenetic diversity measure on rooted phylogenetic trees. In this context, the diversity index 'score' given to a species represents a measure of its unique and shared evolutionary history as displayed in the underlying phylogenetic tree. Our definition generalises the diversity index notion beyond the popular Fair Proportion and Equal-Splits indices. These particular indices may now be seen as two points in a convex space of possible diversity indices, for which the boundary conditions are determined by the underlying shape of each phylogenetic tree. We calculated the dimension of the convex space associated with each tree shape and described the extremal points.

Keywords Phylogenetic tree · Diversity index · Phylogenetic diversity · Fair Proportion index · Equal-splits index · Convex space

1 Introduction

The evolutionary connections and relationships within sets of species are most commonly modelled by phylogenetic trees (Felsenstein 2004). Applying quantitative measures to such trees has proven useful in understanding these relationships and for highlighting conservation priorities in the face of the current human-induced mass

✉ Kerry Manson
kerry.manson@pg.canterbury.ac.nz

[1] School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

extinction of species (Cadotte and Jonathan Davies 2010). A large number of these quantitative approaches have been developed (see Tucker et al. (2017) for an overview).

Phylogenetic diversity (PD) is a measure that aims to quantify the biodiversity exhibited by a set of species (Faith 1992). PD does so by using a weighted phylogenetic tree that exhibits the evolutionary relationships among species in the set, where the weights represent time elapsed or genetic change along the edges of the tree. In broad terms, a PD value for a phylogenetic tree is calculated by adding the weights of all the edges in that tree together. This sum can be seen to represent the total evolutionary history shared by the species at the leaves. However, in some contexts, it is useful to ask how much of that total can be attributed to each species on an individual basis. This perspective opens up the possibility of ranking species by their 'distinctiveness' (Hartmann 2013) or 'evolutionary isolation' (Redding et al. 2014), or to quantify their 'combination of unique and shared evolutionary heritage' (Wicke and Steel 2020). Doing so provides evidence for developing conservation priorities.

The usual means of arriving at such a ranking is by way of a (phylogenetic) diversity index, viewing the weights as edge 'lengths'. These methods have been described as 'distributing edge lengths among descending leaves' (Fischer et al. 2022), a characterisation that we shall use.[1] The most commonly used diversity indices are the Fair Proportion (FP) index and the Equal-Splits (ES) index (described below). The former (under the name Evolutionary Distinctiveness) is used by the EDGE of Existence programme (EDGE 2022; Gumbs et al. 2023; Isaac et al. 2007) to help rank species by their need for conservation assistance. A recent study by Palmer and Fischer (2021) evaluated the effects of this implementation of Evolutionary Distinctiveness on conservation efforts.

Although these particular indices are built on clear principles, they are by no means the unique solutions to the problem of allocating the total PD value among the leaves. No general definition has yet appeared to encompass both the known diversity indices and further possibilities. In this paper, each particular allocation of the PD value is framed in terms of a set of coefficients that determine it. We then give a pair of conditions on these coefficients that satisfy some natural allocation requirements within a simple evolutionary model. These conditions define diversity index functions in a general sense. Some further constraints on the coefficients are derived as immediate consequences of our definition of a diversity index. We also show that, given a set of diversity indices, we can create further diversity indices by taking linear combinations of the original ones. This observation, in turn, leads to descriptions of convex spaces, within which all of the diversity indices for a given tree shape may be positioned. We show how the topology of the underlying tree (the 'tree shape') determines the dimension and boundaries of its associated diversity index space.

The FP and ES indices are thus recast as single points within these spaces. We describe the diversity indices that lie at the extreme points of the convex spaces, from which, by way of convex combination, every index in the space can be calculated. This provides a means of defining the full range of solutions to the allocation problem and the relationships among these possibilities. We illustrate these results by describing, in

[1] Alternative ranking approaches, such as the methods of Crozier (1992) or Vane-Wright et al. (1991), may be useful when the structure of a phylogeny is known but its edge lengths are uncertain. However, we do not consider these further here.

detail, the convex spaces associated with the three rooted binary tree shapes containing five leaves, and by including a case study of the phylogenetic tree of hominoids (great apes and gibbons).

With this method of determining every diversity index on a particular tree, we can investigate the properties of certain diversity indices and see whether or not these properties are unique to that index or if they are widely held. For example, we used our diversity index conditions to show that the FP index is the unique diversity index that obeys a certain ‘continuity’ condition on all tree shapes. This indicates that the FP index is the most appropriate to use with rooted trees that are not fully resolved. Another natural property, which we call ‘consistency’ and which is shared by both the FP and ES indices, is discussed. We show that given fixed edge lengths, every diversity index that satisfies our definition is able to be framed as an equivalent consistent one. This allows us to view indices as a process of re-weighting edge lengths, akin to a flow problem, and to describe the convex spaces of diversity indices in more detail. Our concluding remarks discuss other similar conditions and suggest families of diversity indices that may be of further interest.

2 Preliminaries

We begin by recalling some standard terminology of phylogenetic trees and then introduce some additional notions. Let X be a non-empty set of taxa (e.g. species), with $|X| = n$. A *rooted phylogenetic* X*-tree* is a rooted tree $T = (V, E)$, where X is the set of leaves, and all edges are directed away from a distinguished root vertex ρ, and every non-leaf vertex has out-degree at least 2. We call these non-leaf vertices *interior* vertices. In addition, when $|X| = 1$, the tree consisting of a single vertex is a rooted phylogenetic X-tree. Ignoring the labelling of the leaves by elements of X gives us the *tree shape*. If all interior vertices of T have out-degree 2, we say that T is *binary*.

All edges drawn in this paper will be directed down the page. For the directed edge $e = (u, v)$, we say that u is the *initial* vertex and that v is the *terminal* vertex. We also say that u is the *parent* of v and v is a *child* of u. An edge e from $E(T)$ is *pendant* if its terminal vertex has out-degree zero. Otherwise, e is an *interior* edge. A subtree of T is *pendant* if it can be disconnected from ρ by deleting a single edge of T. We use P_e to denote the pendant subtree formed by deleting the edge e.

We say that a vertex v is *descended from* vertex u in T if u is distinct from v and there exists a directed path in T from u to v. We also say that an edge e is descended from the distinct edge f if the terminal vertex of e is descended from the terminal vertex of f. Edges descended from vertices are defined similarly, by reference to the terminal vertex of the edge involved. However, for vertices descended from edges we do not require the vertex and the terminal vertex of the edge to be distinct. A set S of vertices and/or edges may be said to be descended from an edge e (resp. vertex v) if each member of S is descended from e (resp. v). The *cluster* of all leaves descended from edge e is denoted as $c_T(e)$.

Let e be an interior edge of T for which the terminal vertex v has out-degree d. We represent the d maximal pendant subtrees contained within P_e by

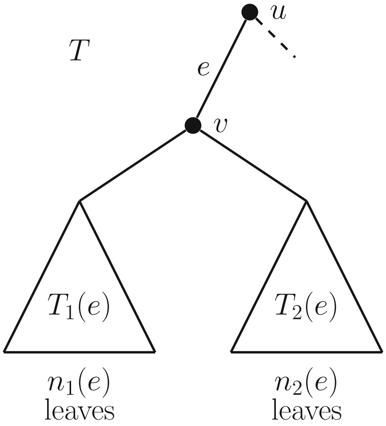


Fig. 1 Schematic diagram showing the notation for maximal pendant subtrees descended from edge e. The vertex v has out-degree 2; however, in general, we allow larger out-degrees

$T_1(e), T_2(e), \ldots, T_d(e)$. Depending on the context, it may be useful to denote these subtrees by $T_1(v), T_2(v), \ldots, T_d(v)$, and allow this notation to extend to the case where v is the root vertex. Figure 1 illustrates this notation for an edge e where $d = 2$.

The edges of every rooted phylogenetic tree considered in this paper are positively weighted. We call these weights *edge lengths*. For a rooted phylogenetic X-tree T, let $\ell : E(T) \rightarrow \mathbb{R}^{>0}$ be a function that assigns a positive real-valued length $\ell(e)$ to each edge $e \in E(T)$. We refer to ℓ as an *edge length assignment function*. We make no further restrictions on the edge lengths other than positivity (i.e., no ultrametric condition is enforced). The *phylogenetic diversity* of T given edge length assignment function ℓ, denoted $PD(T, \ell)$, is defined as the sum of the edge lengths of T. That is:

$$PD(T, \ell) = \sum_{e \in E(T)} \ell(e).$$

Two functions that we considered frequently in this paper were the Fair Proportion (FP) index (Redding 2003; Isaac et al. 2007) and the Equal-Splits (ES) index (Redding 2003; Redding et al. 2014). Let $P(T; \rho, x)$ be the path in T from the root vertex ρ to leaf vertex x. For each leaf $x \in X$, the *Fair Proportion* index score of x in T is given by:

$$FP_T(x) = \sum_{e \in P(T;\rho,x)} \frac{\ell(e)}{|c_T(e)|}.$$

For each leaf $x \in X$, the *Equal-Splits* index score of x in T is given by:

$$ES_T(x) = \sum_{e \in P(T;\rho,x)} \frac{\ell(e)}{\Pi(e, x)},$$

where $\Pi(e, x)$ is the product of the out-degrees of the interior vertices appearing on the path from the terminal vertex of e to x, when e is an interior edge, and $\Pi(e, x) = 1$, when e is incident to x.

3 Diversity Indices

In this section, we provide a set of conditions defining phylogenetic diversity indices in general, for rooted phylogenetic trees. Our definition is guided by a standard biological interpretation of rooted phylogenetic tree structures. For instance, the root vertex corresponds to the most recent common ancestor of the leaves (species) in its tree. Thus the path from the root to a leaf traces the evolutionary development of that leaf species from the common ancestor through to the present. Additionally, the length of each edge in a rooted phylogenetic tree is assumed to reflect the amount of evolutionary change that occurred along that edge. Moreover, evolutionary changes common to the cluster of species $c_T(e)$ are explained by changes occurring somewhere along the edge e. As such, the process of speciation aligns with the tree shape of a rooted phylogenetic tree, where the interior vertices correspond to speciation events.

It is against this interpretation (and the observations above) that we test the conditions that define diversity indices. The broad goal of any phylogenetic diversity index is to take the overall PD score of a rooted phylogenetic tree and distribute this value among the species in a way that is compatible with the tree shape. We aim to present this distribution in a general form, but not to the extent that species are given negative values. A negative allocation of PD value to a species would be equivalent to saying that that species was taking away from the phylogenetic diversity of other species by continuing to survive.

We begin Sect. 3.1 by describing a more general class of functions, here called allocation functions, as well as the coefficient notation that will be used. The further conditions required of diversity indices are included in Sect. 3.2. Definition 2 encompasses the FP and ES indices defined above and allows for the description of further diversity indices. The distinctions between individual diversity indices should reflect different assumptions about how species exhibit ancestral developments while respecting the observations noted above. We conclude this section with some results on the index coefficients that arise immediately from this definition.

3.1 Allocation Functions

Definition 1 Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. An *allocation function* $\varphi_\ell : X \to \mathbb{R}^{\geq 0}$ is a real-valued function on the set of leaves of T that satisfies the following equation:

$$\sum_{x \in X} \varphi_\ell(x) = \sum_{e \in E} \ell(e) = PD(T, \ell), \tag{1}$$

and moreover it may be expressed as $\varphi_\ell(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$, for every edge length assignment function ℓ, where all of the *coefficients* $\gamma(x, e)$ are non-negative.

We call the value $\varphi_\ell(x)$ the *φ-score* of leaf x (given the length assignment ℓ). Our aim is to define diversity indices that not only deliver an ordinal ranking of species according to their contribution to PD, but also to measure this contribution. With this in mind, it is sensible to require allocation functions (and, consequently, diversity indices) to partition the total PD value of a rooted phylogenetic tree among its leaf species. This idea is expressed by Eq. (1) and is reinforced by requiring each allocation function to be expressed in terms of coefficients $\gamma(x, e)$ that do not depend on the particular edge length assignment function ℓ. Edge lengths will be used to calculate the value of individual allocation function scores but do not themselves impact the method of allocation. Because of this independence, the length subscript on φ_ℓ will be omitted when ℓ is clear from the context.

The final condition of Definition 1, that all of the coefficients $\gamma(x, e)$ are non-negative, is worth considering further. Without it, the class of allocation functions would contain many biologically unreasonable functions. One such unreasonable allocation can be described on the small rooted phylogenetic tree with exactly two leaves x and y, and two edges a and b. On this tree, consider the function $\sigma : \{x, y\} \to \mathbb{R}$, defined as follows:

$$\begin{aligned} x &\mapsto 2\ell(a) + 2\ell(b) \\ y &\mapsto -\ell(a) - \ell(b) \end{aligned}$$

Then σ indeed satisfies (1), as $\sigma(x) + \sigma(y) = \ell(a) + \ell(b)$. However, claiming that $\sigma(y)$ represented the evolutionary history of y would be hard to justify. Negative coefficients and scores do not fit our intended model, where diversity indices act as a measure of (necessarily positive) evolutionary history; hence the final stipulation in Definition 1. The next result shows that, similar to PD, allocation functions are *linear* in the following sense.

Proposition 1 *Let $T = (V, E)$ be a rooted phylogenetic X-tree. Furthermore, let φ be an allocation function that may be written in the form $\varphi_\ell(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$ for every edge length assignment function ℓ. Suppose that $s, t \in \mathbb{R}$ and that ℓ, ℓ_1 and ℓ_2 are edge length assignment functions such that $\ell(e) = s\ell_1(e) + t\ell_2(e)$ for every edge $e \in E$. Then $\varphi_\ell(x) = s\varphi_{\ell_1}(x) + t\varphi_{\ell_2}(x)$ for all leaves $x \in X$.*

Proof Let $s, t \in \mathbb{R}$ and suppose that ℓ, ℓ_1 and ℓ_2 are edge length assignment functions such that $\ell(e) = s\ell_1(e) + t\ell_2(e)$ for every $e \in E$. Then:

$$\begin{aligned} \varphi_\ell(x) &= \sum_{e\in E} \gamma(x, e)\ell(e) \\ &= \sum_{e\in E} \gamma(x, e)\,(s\ell_1(e) + t\ell_2(e)) \\ &= s\sum_{e\in E} \gamma(x, e)\ell_1(e) + t\sum_{e\in E} \gamma(x, e)\ell_2(e) \\ &= s\varphi_{\ell_1}(x) + t\varphi_{\ell_2}(x). \end{aligned}$$

□

An allocation function on a phylogenetic X-tree $T = (V, E)$ may be determined by a rule or formula, or, if needed, can be completely described by listing the $|X| \times |E|$ coefficients $\gamma(x, e)$. Note that the FP and ES indices satisfy Definition 1, and hence are allocation functions. For these indices, the coefficients may be read directly from their definitions. That is, they can be read as $\frac{1}{|c_T(e)|}$ and $\frac{1}{\Pi(e,x)}$, respectively, when $x \in c_T(e)$, noting that for both indices $\gamma(x, e) = 0$ whenever x is not descended from e. In contrast, we can define the coefficients directly. Let $|X| = n$, and take $\gamma(x, e) = \frac{1}{n}$ for all $x \in X$ and all $e \in E$. Then $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ is an allocation function that shares the total PD value of T uniformly among the n leaves.

For some tree shapes, two allocation functions with different descriptions or formulae for calculating coefficients may nevertheless ultimately produce the same scores. We say that two allocation functions φ and ψ *coincide* if $\varphi_\ell(x) = \psi_\ell(x)$ for every $x \in X$ and every edge length assignment function ℓ. This concept has been used to understand the similarities and differences among some diversity indices. For instance, Wicke and Steel (2020) characterised the rooted tree shapes for which the FP and ES indices coincide.

3.2 Diversity Indices

The 'uniform' allocation function above, where each coefficient is $\frac{1}{n}$, is an allocation function that ignores the phylogenetic structure entirely. It is thus an allocation function that does not allow for relative or differing contributions to biodiversity. In contrast, diversity indices are a subclass of allocation functions that do account for the rooted phylogenetic tree structure.

We intend for $\gamma(x, e)$ to represent the proportion of evolutionary history that arises along edge e that is currently embodied by species x. This provides our first constraint, namely that the evolutionary history arising along edge e should be allocated exclusively to species descended from e.

Additionally, the set of coefficients for an edge should be entirely determined by the shape of its descendent tree structure, not by any ancestral or otherwise unconnected parts of the tree. In other words, the same pattern of descent should lead to the same pattern of coefficients. Furthermore, the particular labelling of the leaves should be inconsequential for the calculation of a diversity index. Definition 2 (below) restricts the class of allocation functions to those that obey these minimal criteria. In order to describe this definition, we first describe two further notions.

A *symmetry* of T is a permutation of the vertices that maintains precisely those relationships of descent found in T. Suppose that T and T' are two rooted phylogenetic trees with the same tree shape. Let π be a bijective map from the vertices of T to the vertices of T', such that v is descended from u if and only if $\pi(v)$ is descended from $\pi(u)$. Then v and $\pi(v)$ are said to be in *corresponding positions* of that tree shape. Moreover, recall that P_e denotes the pendant subtree formed by deleting the edge e.

Definition 2 Let $T = (V, E)$ be a rooted phylogenetic X-tree. A *diversity index* (on T) is an allocation function $\varphi_\ell : X \to \mathbb{R}^{>0}$ given by $\varphi_\ell(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ for every edge length assignment function ℓ, that additionally satisfies the conditions (DI$_1$) and (DI$_2$) below:

- (DI$_1$) *Descent condition:* $\gamma(x, e) = 0$ if x is not descended from e.
- (DI$_2$) *Neutrality condition:* The coefficients $\gamma(x, e)$ are a function of the tree shape of P_e. Moreover, suppose that P_e and P_f are pendant subtrees of T with the same

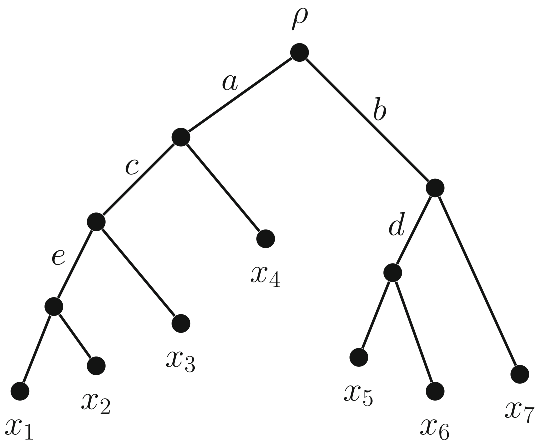


Fig. 2 Any diversity index on the above tree must allocate the length of edge c exclusively among its descendent cluster: $\{x_1, x_2, x_3\}$. The pattern of this allocation must be matched by the allocation of the length of edge b among b's own descendent cluster, namely $\{x_5, x_6, x_7\}$. This is because the maximal pendant subtrees below b and c have the same tree shape

tree shape. If leaves x in P_e and y in P_f appear in corresponding positions in their respective subtrees, then $\gamma(x, e) = \gamma(y, f)$.

Consequently, to satisfy the neutrality condition (DI$_2$), any diversity index on the tree in Fig. 2 is required to satisfy all of the following coefficient equalities: $\gamma(x_1, a) = \gamma(x_2, a)$, $\gamma(x_5, b) = \gamma(x_6, b) = \gamma(x_1, c) = \gamma(x_2, c)$, $\gamma(x_7, b) = \gamma(x_3, c)$ and $\gamma(x_5, d) = \gamma(x_6, d) = \gamma(x_1, e) = \gamma(x_2, e)$.

We now discuss some immediate consequences of our definition. Proposition 3 discusses some constraints on the allocation coefficients. Proposition 4 reframes the neutrality condition in terms of symmetries, and Proposition 5 gives bounds on the index scores of a general leaf. Wicke (2020) showed that for any function of the form $\varphi_\ell(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$ that satisfies Eq. (1), the coefficients associated with each edge sum to one. Lemma 2 reframes this result slightly so it applies to allocation functions.

Lemma 2 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment ℓ. Suppose that $\gamma(x, e) \geq 0$ for all $x \in X$ and $e \in E$. Then the function $\varphi(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$ is an allocation function if and only if $\sum_{x\in X} \gamma(x, e) = 1$ for every edge $e \in E$.*

Proposition 3 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. Let $\varphi(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$ be a diversity index. Then*

(i) $\gamma(x, e) \leq 1$ *for all $x \in X$ and $e \in E$,*
(ii) $\sum\limits_{x\in c_T(e)} \gamma(x, e) = 1$, *and*
(iii) *if e is a pendant edge, then $\gamma(x, e) = 1$ when leaf x is incident with e, and $\gamma(x, e) = 0$ otherwise.*

Proof Suppose that $\gamma(x', e) > 1$ for some edge e and leaf x'. From the definition of allocation functions, all other coefficients of φ are non-negative, so $\sum_{x\in X} \gamma(x, e) > 1$. However this contradicts Lemma 2 and hence $\gamma(x, e) \leq 1$ for all $x \in X$ and $e \in E$. Next, for a specified edge e, we can split the leaves into two sets: the set of leaves descended from e, denoted $c_T(e)$, and the rest: $X \setminus c_T(e)$. Then, starting from the Lemma 2 result:

$$1 = \sum_{x \in X} \gamma(x, e) = \sum_{x \in c_T(e)} \gamma(x, e) + \sum_{x \in X \setminus c_T(e)} \gamma(x, e) = \sum_{x \in c_T(e)} \gamma(x, e),$$

where the last equality follows from the descent condition (DI$_1$). Lastly, let e be a pendant edge. Suppose $x' \in X$ is incident with e, giving $c_T(e) = \{x'\}$. Thus by Part (ii), $1 = \sum_{x \in c_T(e)} \gamma(x, e) = \gamma(x', e)$. Now suppose that x' is not incident with e. Then x' is not descended from e, so by using (DI$_1$) we conclude that $\gamma(x', e) = 0$. □

Proposition 4 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. Let $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ be a diversity index. For distinct leaves x, x', both descended from e, if there is a symmetry in T that swaps x for x' then $\gamma(x, e) = \gamma(x', e)$.*

Proof Let T' be the resulting tree after applying the symmetry to T that swaps x for x'. Then P_e and P'_e (the pendant subtrees created by deleting edge e in T and T', respectively) have the same tree shape, and leaf x in P_e corresponds to leaf x' in P'_e. Hence, by the neutrality condition (DI$_2$) $\gamma(x, e) = \gamma(x', e)$. □

Proposition 5 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. Let e_x be the pendant edge incident with leaf $x \in X$, and let $P(T; \rho, x)$ be the path in T from the root vertex ρ to x. Let $\sigma_x(e)$ be the number of leaf vertices in $c_T(e)$ that appear in a corresponding position to x (including x). If φ is a diversity index on T, then $\ell(e_x) \leq \varphi(x) \leq \sum\limits_{e \in P(T; \rho, x)} \frac{\ell(e)}{\sigma_x(e)}$ for all $x \in X$.*

Proof Let $x \in X$ and $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$. First suppose that $\gamma(x, e) = 0$ for all $e \in E \backslash \{e_x\}$. This is the minimal possible choice, as Proposition 3(iii) ensures that $\gamma(x, e_x) = 1$. In this case, $\varphi(x) = \gamma(x, e_x)\ell(e_x) + \sum\limits_{e \in (E \backslash \{e_x\})} \ell(e) = \ell(e_x) + 0$.

Next, for some edge $e \in E$, suppose that $\gamma(x, e)$ is non-zero. By (DI$_1$), a non-zero coefficient means that x is descended from e (equivalently, $e \in P(T; \rho, x)$). The coefficients of the $\sigma_x(e)$ leaves in corresponding positions to x must all have coefficients equal to $\gamma(x, e)$. We require $\gamma(x, e) \leq \frac{1}{\sigma_x(e)}$ in order for the coefficients associated with e to sum to at most 1 and not contravene Proposition 3(ii). Thus $\varphi(x) = \sum\limits_{e \in P(T; \rho, x)} \gamma(x, e)\ell(e) \leq \sum\limits_{e \in P(T; \rho, x)} \frac{\ell(e)}{\sigma_x(e)}$. Therefore, for all $x \in X$,

$$\ell(e_x) \leq \varphi(x) \leq \sum_{e \in P(T; \rho, x)} \frac{\ell(e)}{\sigma_x(e)}.$$

□

4 Continuity

In Steel (2016)[p. 140], the following property of the FP index was noted:

> "The index $\psi = FP$ satisfies the following continuity condition: If e is an interior edge of a phylogenetic tree and T/e is the tree obtained from T by collapsing edge e, then $\lim_{l(e)\to 0} \psi_T(a) = \psi_{T/e}(a)$."

As an illustration of how Definition 2 can be used to investigate the properties of diversity indices, we show that FP is the unique diversity index that satisfies this continuity condition on every tree. We use the same notation of T/e to denote the tree obtained from T by collapsing edge e, and refer to this property as the *diversity index continuity property*.

Theorem 6 *Let ψ be a diversity index. Then $\lim_{\ell(e)\to 0} \psi_T(x) = \psi_{T/e}(x)$ for every rooted phylogenetic X-tree $T = (V, E)$, for every $x \in X$ and for every interior edge $e \in E(T)$ if and only if ψ is the Fair Proportion index.*

Proof Suppose that ψ is a diversity index that satisfies the diversity index continuity property on every rooted phylogenetic tree. Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. Since ψ is a diversity index, we have $\psi_T(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$, for coefficients $\gamma(x, e)$ that satisfy (DI$_1$) and (DI$_2$). The index ψ may be defined by its set of coefficients. Moreover, we need to specify only those coefficients that are not already determined by the definition of a diversity index.

Let $|X| = n$ and $E(T) = \{e_1, e_2, \ldots, e_{2n-2}\}$, where the pendant edges of T are indexed by $\{1, \ldots, n\}$ and the interior edges of T are indexed by $\{n+1, \ldots, 2n-2\}$. Let e_i be an interior edge of T and let $n_i = |c_T(e_i)|$, the number of leaves of T descended from e_i. First, assume that no interior edge of T is descended from e_i. That is, only pendant edges appear below e_i. By the neutrality condition (DI$_2$) and Proposition 3(ii), if x is descended from e_i, then $\gamma(x, e_i) = \frac{1}{n_i}$. Otherwise, $\gamma(x, e_i) = 0$.

Now assume that e_j is an interior edge descended from e_i. By the diversity index continuity property:

$$\lim_{\ell(e_j)\to 0} \psi_T(x) = \lim_{\ell(e_j)\to 0} \left(\gamma(x, e_j)\ell(e_j) + \sum_{e\in E\setminus\{e_j\}} \gamma(x, e)\ell(e) \right)$$
$$= \sum_{e\in E\setminus\{e_j\}} \gamma(x, e)\ell(e) = \psi_{T/e_j}(x).$$

Note that the process of contracting an interior edge does not alter the coefficients $\gamma(x, e)$ of ψ. Nor does it alter the number of leaves descended from any other edge. Let $D(e_i)$ be the set of interior edges descended from e_i. We contract each edge from $D(e_i)$ in turn. By repeated use of the diversity index continuity property,

$$\lim_{\substack{\ell(e)\to 0\\ \text{all } e\in D(e_i)}} \psi_T(x) = \sum_{e\in(E\setminus D(e_j))} \gamma(x, e)\ell(e) = \psi_{T/D(e_i)}(x).$$

In $T/D(e_i)$, no interior edge is descended from e_i. Hence, we again have $\gamma(x, e_i) = \frac{1}{n_i}$ if x is a descendant of e_i. Since the coefficients are unaffected by the contraction process, $\gamma(x, e_i) = \frac{1}{n_i}$ in T as well. By repeating this process for each interior edge $e_{n+1}, \ldots, e_{2n-2}$, we obtain every coefficient $\gamma(x, e)$ of each edge $e \in E(T)$. This argument shows that in every case $\gamma(x, e_i) = \frac{1}{n_i} = \frac{1}{|c_T(e_i)|}$, which is exactly the set of coefficients that defines the Fair Proportion index on T. Therefore, ψ is the FP index.

Conversely, suppose that ψ is the FP index. Then the coefficients $\gamma(x, e)$ depend only on the number of leaves descended from edge e, and not on the particular structure of the phylogenetic tree below e. The contraction of any interior edge e_i does not reduce the number of leaf vertices descended from any other edge e. Hence the coefficients $\gamma(x, e)$ are not altered when an interior edge distinct from e itself is contracted, and the diversity index continuity property holds. □

FP may therefore be considered the most appropriate diversity index to apply to rooted phylogenetic trees that are not fully resolved, as later refinements of the tree will not strongly impact the initial FP index scores. We note also that FP is special amongst diversity indices for a quite different reason – it is precisely the Shapley value for allocating the total phylogenetic diversity of a tree amongst the leaves (Fuchs and Jin 2015).

5 Spaces of Diversity Indices

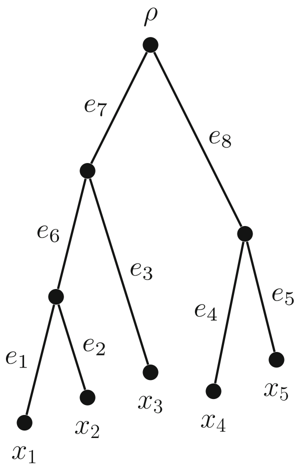
Each rooted phylogenetic tree has a restricted collection of diversity indices that may be applied to it. In this section, we discuss such collections of diversity indices. We show that these indices, when expressed as vectors of index scores, lie inside convex spaces determined by the structure of the associated tree. Examples of these spaces are presented for small rooted phylogenetic trees.

Let $T = (V, E)$ be a rooted phylogenetic X-tree with $X = \{x_1, \ldots, x_n\}$ and $E = \{e_1, \ldots, e_m\}$. Suppose that φ is a diversity index on the tree T given by $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ for the edge length assignment function ℓ. We place the coefficients $\gamma(x, e)$ in an $n \times m$ matrix $A_\varphi = (a_{ij})$, where $a_{ij} = \gamma(x_i, e_j)$. Consider the tree in Fig. 3. Its matrix of diversity index coefficients fits the pattern shown in the same figure, where the value of $0 \leq \alpha \leq 1$ is determined by the particular index. For example, FP uses $\alpha = \frac{1}{3}$ here and ES uses $\alpha = \frac{1}{2}$. We now use the matrix of coefficients to determine the φ index scores of each leaf and subsequently form a space containing all the possible scores. The next two definitions introduce these concepts.

Definition 3 Let $T = (V, E)$ be a rooted phylogenetic X-tree with the leaf set $X = \{x_1, \ldots, x_n\}$ and edges $E = \{e_1, \ldots, e_m\}$. Let φ be a diversity index on T, for which A_φ is the associated matrix of coefficients. The *index score vector* for φ is the n-component vector $\boldsymbol{v_\varphi} = A_\varphi \boldsymbol{l}$, where $\boldsymbol{l}$ is the vector of edge lengths: $[\ell(e_1), \ell(e_2), \ldots, \ell(e_m)]^T$.

Definition 4 Let $T = (V, E)$ be a rooted phylogenetic tree. The *space of index score vectors* on T, denoted $S(T, \ell)$, contains all the index score vectors of T.

Note that $S(T, \ell)$ is not a vector space, since $\mathbf{0}$ is not in $S(T, \ell)$ for any edge length assignment function ℓ. This is because the index score of any leaf is always at least the (strictly positive) length of its incident pendant edge. However, $S(T, \ell)$ is compact by Proposition 5 and convex, as we now show.



$$A_\varphi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1-\alpha}{2} & 0 \\ 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1-\alpha}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

Fig. 3 Left: A rooted phylogenetic tree on five leaves. Right: the associated matrix of diversity index coefficients for this tree. The value of $0 \leq \alpha \leq 1$ determines the diversity index (up to coincident indices)

Proposition 7 *Let $T = (V, E)$ be a rooted phylogenetic X-tree. For any edge length assignment function ℓ, the space of index score vectors $S(T, \ell)$ is convex.*

Proof Let $X = \{x_1, \ldots, x_n\}$ and $E = \{e_1, \ldots, e_m\}$, and suppose that φ and ψ are diversity indices on T, where $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ and $\psi(x) = \sum_{e \in E} \gamma'(x, e)\ell(e)$. Let $A_\varphi = (a_{ij}) = (\gamma(x_i, e_j))$ and $B_\psi = (b_{ij}) = (\gamma'(x_i, e_j))$ be the respective $n \times m$ coefficient matrices of φ and ψ. Finally, let $\boldsymbol{u} = A_\varphi \boldsymbol{l}$ and $\boldsymbol{v} = B_\psi \boldsymbol{l}$ be the respective vectors of index scores for φ and ψ. That is, $\boldsymbol{u}, \boldsymbol{v} \in S(T, \ell)$.

We first prove that the linear combination $\boldsymbol{w} = t\boldsymbol{u} + (1 - t)\boldsymbol{v}$ is an allocation function for all real values of $t \in [0, 1]$. Let $\delta(x_i, e_j) = t\gamma(x_i, e_j) + (1 - t)\gamma'(x_i, e_j)$ for all $1 \leq i \leq n$ and $1 \leq j \leq m$. We then have:

$$\begin{aligned} t\boldsymbol{u} + (1 - t)\boldsymbol{v} &= tA_\varphi \boldsymbol{l} + (1 - t)B_\psi \boldsymbol{l} \\ &= [tA_\varphi + (1 - t)B_\psi]\boldsymbol{l} \\ &= (ta_{ij} + (1 - t)b_{ij})\boldsymbol{l} \\ &= (t\gamma(x_i, e_j) + (1 - t)\gamma'(x_i, e_j))\boldsymbol{l} \\ &= (\delta(x_i, e_j))\boldsymbol{l}. \end{aligned}$$

We form a new matrix $C = (c_{ij}) = (\delta(x_i, e_j))$ from these values. The coefficients $\gamma(x_i, e_j)$ and $\gamma'(x_i, e_j)$ are non-negative for all $1 \leq i \leq n$ and $1 \leq j \leq m$, and for every $t \in [0, 1]$ we have $t \geq 0$ and $(1 - t) \geq 0$. Thus the coefficients $\delta(x_i, e_j)$ are also non-negative. Then C is the matrix of coefficients of an allocation function because, for every $1 \leq j \leq m$, the characterisation from Lemma 2 is satisfied:

$$\begin{aligned}\sum_{i=1}^{n}\delta(x_i,e_j)=\sum_{x\in X}\delta(x,e_j)&=\sum_{x\in X}t\gamma(x,e_j)+(1-t)\gamma'(x,e_j)\\&=t\sum_{x\in X}\gamma(x,e_j)+\sum_{x\in X}\gamma'(x,e_j)-t\sum_{x\in X}\gamma'(x,e_j)\\&=t+1-t=1.\end{aligned}$$

It remains to show that the coefficients $\delta(x_i,e_j)$ satisfy the conditions of a diversity index, namely (DI$_1$) and (DI$_2$). If x_i is not descended from e_j, then $\gamma(x_i,e_j)=0$ and $\gamma'(x_i,e_j)=0$. Therefore, $\delta(x_i,e_j)=t\cdot 0+(1-t)\cdot 0=0$, and the coefficient $\delta(x_i,e_j)$ satisfies (DI$_1$).

Next assume that, for some $x,y\in X$ and $e,f\in E$, to satisfy (DI$_2$), we require $\gamma(x,e)=\gamma(y,f)$. Then $\gamma'(x,e)=\gamma'(y,f)$ as well, and hence

$$\delta(x,e)=t\gamma(x,e)+(1-t)\gamma'(x,e)=t\gamma(y,f)+(1-t)\gamma'(y,f)=\delta(y,f),$$

as required. So any set of coefficients that are all equal in A_φ and are also all equal in B_ψ will all be equal in C. Thus C is the matrix of coefficients of a diversity index, and $\boldsymbol{w}=C\boldsymbol{l}$ is contained in $S(T,\ell)$. Therefore, $S(T,\ell)$ is closed under convex combinations and is a convex space. □

Suppose that we fix a particular rooted phylogenetic X-tree T with edge length assignment function ℓ. Each possible diversity index for T may be viewed as a point inside the convex space $S(T,\ell)$, and the Euclidean distances between distinct diversity index vectors indicate their degree of difference. If this distance is zero in $S(T,\ell)$, the diversity indices in question coincide on T. The space $S(T,\ell)$ consists of $|X|$-dimensional vectors. However, it will be shown (Proposition 13) that the vectors in $S(T,\ell)$ do not span all of $\mathbb{R}^n$, but rather $S(T,\ell)\subset\mathbb{R}^k$ for some $k<n$. The smallest such value of k is called the *dimension* of $S(T,\ell)$. Diversity indices are completely described by their coefficients rather than their edge lengths, so the dimension of $S(T,\ell)$ is determined by the tree shape of T alone. Although the dimension relies only on the tree shape, the particular boundaries are determined by the edge lengths, as described in Proposition 5.

5.1 Examples and the Special Case When $S(T,\ell)$ has Dimension Zero

To illustrate the effect of tree shape on the dimension of the diversity index space, we examine the spaces of diversity indices for some rooted phylogenetic trees with five leaves. The connection between tree shape and dimension will be formalised in the next section.

Consider again the five-leaf tree in Fig. 3. By Proposition 3, the pendant edge lengths are entirely allocated to their incident leaves. By (DI$_2$), the length of edge e_6 must be shared equally between x_1 and x_2 in every diversity index. Similarly, the length of e_8 must be shared equally between x_4 and x_5 in every diversity index. Thus, for this tree, only the allocation of edge e_7 between x_1, x_2 and x_3 changes. Suppose

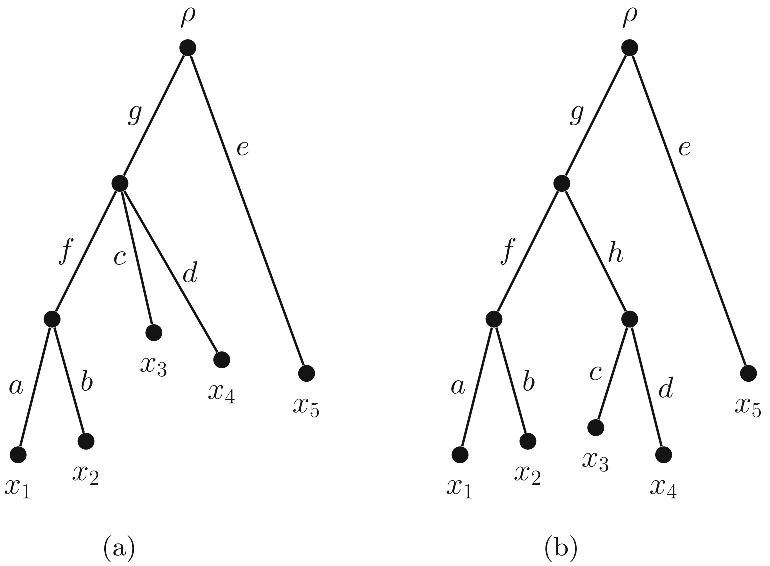


Fig. 4 Two rooted phylogenetic tree shapes on five leaves

Table 1 Diversity index scores for the trees in Fig. 4. Each value of α, for $0 \leq \alpha \leq 1$, determines a diversity index for the tree in (a)

x	$\varphi(x)$
(a)	
x_1	$\frac{1}{2}(1-2\alpha)\ell(g) + \frac{\ell(f)}{2} + \ell(a)$
x_2	$\frac{1}{2}(1-2\alpha)\ell(g) + \frac{\ell(f)}{2} + \ell(b)$
x_3	$\alpha\ell(g) + \ell(c)$
x_4	$\alpha\ell(g) + \ell(d)$
x_5	$\ell(e)$
(b)	
x_1	$\frac{\ell(g)}{4} + \frac{\ell(f)}{2} + \ell(a)$
x_2	$\frac{\ell(g)}{4} + \frac{\ell(f)}{2} + \ell(b)$
x_3	$\frac{\ell(g)}{4} + \frac{\ell(h)}{2} + \ell(c)$
x_4	$\frac{\ell(g)}{4} + \frac{\ell(h)}{2} + \ell(d)$
x_5	$\ell(e)$

that $\gamma(x_3, e_7) = \alpha$ is the share of the length of e_7 allocated to x_3. Accordingly, $\gamma(x_1, e_7) + \gamma(x_2, e_7) + \alpha = 1$. The share of the length of e_7 allocated to x_1 must equal the share of the length of e_7 allocated to x_2, so $\gamma(x_1, e_7) = \gamma(x_2, e_7) = \frac{1}{2}(1-\alpha)$. We require $0 \leq \alpha \leq 1$, but α is free to be chosen within this range and the stated coefficients will satisfy the diversity index definition. Thus the set of diversity indices on this five-leaf tree form a one-dimensional space, parametrised by the value of α.

An analogous parametrisation defines the diversity indices on the non-binary tree appearing in Fig. 4a.

For some tree shapes, however, (DI$_2$) provides enough of a constraint that all diversity indices coincide. One such shape is given by the tree in Fig. 4b, with the necessary index scores appearing in Table 1c. A *balanced* tree is a rooted phylogenetic X-tree T in which, for every vertex v, the pendant subtrees $T_1(v), \dots, T_d(v)$ all have the same tree shape as each other (where d is the out-degree of v). We call a pendant subtree P of T *maximal* if it is not a subtree of any other pendant subtree, that is, $P \in \{T_1(\rho), \dots, T_d(\rho)\}$, where d is the out-degree of the root vertex ρ. A *semi-balanced* tree is a rooted phylogenetic X-tree T where the maximal pendant subtrees of T are all balanced trees.

Proposition 8 *Let T be a rooted phylogenetic X-tree. The space of diversity indices on T consists of a single point if and only if T is a semi-balanced tree.*

Proof Let $\varphi(x) = \sum_{e \in E(T)} \gamma(x, e)\ell(e)$ be a diversity index on T. Suppose that T is a semi-balanced tree. For every leaf vertex in $T_1(\rho)$, there is a symmetry in T that swaps that vertex with any other leaf vertex in $T_1(\rho)$. A similar argument holds for leaves sharing any other maximal pendant subtree of T. Let e be an interior edge of T. All leaves in $c_T(e)$ must belong to the same maximal pendant subtree of T. Thus, by Proposition 4, $\gamma(x, e) = \gamma(x', e)$ for any x, x' in $c_T(e)$. Therefore, for each leaf x descended from e, we can set $\gamma(x, e) = \frac{1}{|c_T(e)|}$. The coefficients associated with every interior edge of T can be determined in this way, and those of the pendant edges are fixed by Proposition 3 (iii). Hence, there is only one possible set of index scores, i.e., $S(T, \ell)$ has dimension zero for any edge length assignment function ℓ.

Conversely, suppose that T is not a semi-balanced tree. Then there exists some edge $f = (u, v)$ (with positive length) such that $T_1(v)$ does not have the same tree shape as, say, $T_2(v)$. Without loss of generality, assume that there are a_1 maximal pendant subtrees below v with the same tree shape as $T_1(v)$, and a_2 maximal pendant subtrees below v with the same tree shape as $T_2(v)$. Let n_1 be the number of leaves in $T_1(v)$ and let n_2 be the number of leaves in $T_2(v)$. We define the following two distinct diversity indices on T.

Firstly, take $\gamma(x, f) = \frac{1}{a_1 n_1}$ whenever x lies in a maximal pendant subtree below v that has the same tree shape as $T_1(v)$. Then $\gamma(x, f) = 0$ whenever x lies in a maximal pendant subtree below v that has a different tree shape from $T_1(v)$. For a second set of coefficients, take $\gamma'(x, f) = \frac{1}{a_2 n_2}$ whenever x lies in a maximal pendant subtree below v that has the same tree shape as $T_2(v)$. Then $\gamma'(x, f) = 0$ whenever x lies in a maximal pendant subtree below v that has a different tree shape from $T_2(v)$. In both cases, we use the ES coefficients for every other edge (unless an edge has the same descendent structure as f, in which case, we copy the pattern of f's coefficients in order to satisfy (DI$_2$)). Let $\varphi(x) = \sum_{e \in E} \gamma(x, e)$ and $\psi(x) = \sum_{e \in E} \gamma'(x, e)$. The index scores of φ and ψ are never coincident, as $\ell(f) > 0$. Thus $S(T, \ell)$ contains more than one index and cannot consist of a single point. □

Corollary 9 *The Fair Proportion and Equal-Splits indices coincide on a rooted binary phylogenetic tree T if and only if all diversity indices coincide on that tree as well.*

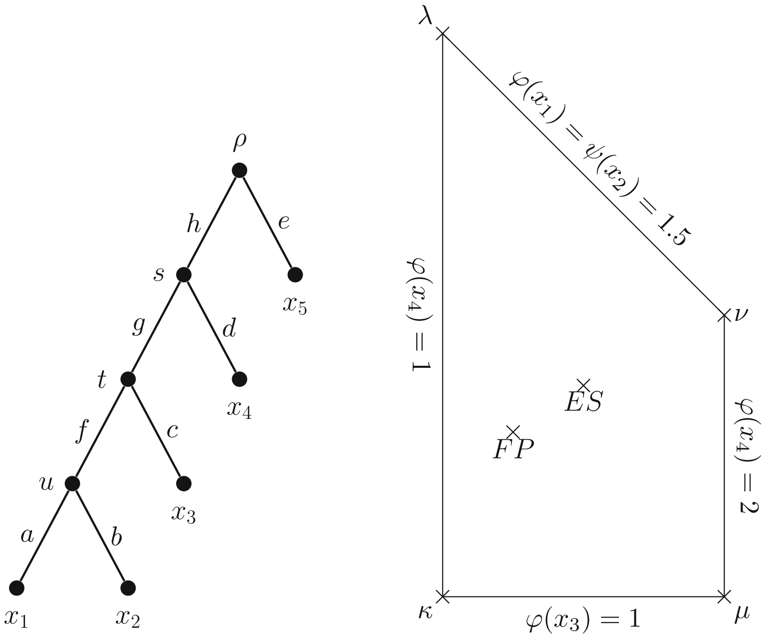


Fig. 5 Left: The five-leaf rooted caterpillar tree Cat_5, where each edge has unit length. Right: The diversity index space $S(Cat_5, \mathbf{1})$, projected onto the $\varphi(x_3)\varphi(x_4)$-plane, with boundary conditions and extremal indices κ, λ, μ, ν at the corner points. See Fig. 7 for illustrations of these corner indices. Notice that the FP and ES diversity indices appear as interior points of this convex space

Proof Wicke and Steel (2020, Theorem. 4) showed that FP and ES coincide on a rooted binary phylogenetic tree T if and only if T is semi-balanced. Their proof may be extended directly to the non-binary case. By Proposition 8, T being semi-balanced is equivalent to $S(T, \ell)$ consisting of a single point. That is, all diversity indices on T are coincident. □

6 Consistent Diversity Indices

The coefficients of the FP and ES diversity indices share a property that allows us to view their calculation as a type of flow problem. This is a useful means of understanding the allocation of PD for these indices. For both FP and ES, the allocation of an edge length among the maximal pendant subtrees of a clade is the same for every edge ancestral to that clade, not in the value of the coefficients but in their ratios. For example, consider the tree in Fig. 5. The FP allocation of edge g to $\{x_1, x_2\}$ compared with $\{x_3\}$ is $\frac{2}{3} : \frac{1}{3}$, whereas the allocation of edge h among these same three species is $\frac{2}{4} : \frac{1}{4}$. In both cases, a 2:1 ratio of allocations holds. When we use the ES index, the allocation between these same sets follows a $1 : 1$ ratio for both g and h. A diversity index that has the same ratio of allocation at vertex v, for every edge that v is descended from, we call *consistent at* v. We introduce some notation to aid our discussion of this property. Let $\Gamma_i(v, e) := \sum_{x \in T_i(v)} \gamma(x, e)$, that is, the sum of all coefficients associated with both edge e and some leaf in the ith pendant subtree below vertex v.

Definition 5 *(Consistency condition)* Let $T = (V, E)$ be a rooted phylogenetic X-tree. Choose $e, f \in E$, with $e = (u, v)$ descended from f, and let $T_1(e), \ldots, T_d(e)$ be the maximal pendant subtrees descended from the terminal vertex of e. A diversity index $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ is *consistent at* v if there exists a constant $k \in \mathbb{R}^{\geq 0}$ such that $\Gamma_i(v, f) = k\Gamma_i(v, e)$. If φ is consistent at v for every $v \in V$, then we say that φ itself is *consistent*.

At vertex v, descended from $e \in E$, the ratio $\Gamma_1(v, e) : \cdots : \Gamma_d(v, e)$ is called the *ratio of allocations* at v. A ratio of allocations is *normalised* if it has been scaled by some $t \in \mathbb{R}$ such that $t \sum_{i=1}^{d} \Gamma_i(v, e) = 1$. Reiterating our earlier example, a consistent diversity index on the rooted phylogenetic tree in Fig. 5 requires the ratio of allocation $\left[\gamma(x_1, g) + \gamma(x_2, g)\right] : \gamma(x_3, g)$ to be the same as $\left[\gamma(x_1, h) + \gamma(x_2, h)\right] : \gamma(x_3, h)$.

Consistent diversity indices have the convenient property that their index scores are able to be determined by a simple flow-based algorithm. The idea is to view each index as a rule for re-weighting edges, moving weight from each interior edge to its immediate descendent edges. The algorithm begins with the edges incident to the root vertex and continues to re-weight edges until all interior edges have a weight of zero. The diversity index score of each leaf is then given by the final length of its incident pendant edge. This approach is reminiscent of the transformations of edge lengths in (Haake et al. 2008) that maintain the Shapley values of each leaf.

Consider a particular rooted phylogenetic tree T with fixed edge lengths ℓ. We now show that any diversity index on T may be framed as a consistent index. This reframing, in turn, will allow us to describe the dimension and extreme points of $S(T, \ell)$.

Lemma 10 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with a fixed edge length assignment function ℓ. Let $\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ be a consistent diversity index on T. Given the ratios of allocation at each vertex, we can reconstruct the diversity index coefficients.*

Proof For each $v \in V$, we normalise the ratios of allocation, writing the scaled ratio as $r_1(v) : \cdots : r_d(v)$. Let $e = (u_0, u_1)$ be an interior edge of T and let $x \in X$ be descended from e. Let $P(T; u_1, x)$ consist of the vertices $\{u_1, u_2, \ldots, u_k, x\}$ and without loss of generality assume that $x \in X$ is in $T_1(v)$ for all $v \in P(T; u_1, u_k)$.

Firstly, the proportion of $\ell(e)$ allocated among vertices in $T_1(u_1)$ is $r_1(u_1)$. Next note that the proportion of $\ell(e)$ allocated among vertices in $T_1(u_2)$ is $r_1(u_2)$ of the $\ell(e)$ allocation coming into u_2. That is, $r_1(u_1) \cdot r_1(u_2)$ overall. Similarly, the leaves of $T_1(u_3)$ are allocated a total of $r_1(u_1) \cdot r_1(u_2) \cdot r_1(u_3)$ of $\ell(e)$. Continuing in this manner until the parent of x, we find that $\gamma(x, e)$ is given by the product $\prod_{v \in P(T; u_2, u_k)} r_1(v)$. For an arbitrary leaf x' descended from edge e', the coefficient $\gamma(x', e')$ can be expressed as a similar product, taking the appropriate ratio terms at vertices along the path connecting edge e' to leaf x'. □

We give a short example of the calculation above in order to clarify the final sentence of this proof. Suppose that x is a leaf of a phylogenetic tree rooted at ρ, and that the path from ρ to x in T passes through vertices s, t, u and v in turn. Let e be the edge (s, t). Suppose further that $x \in T_2(t)$, $x \in T_5(u)$ and $x \in T_1(v)$. Then the coefficient $\gamma(x, e)$ in this case is given by the product $r_2(t) \cdot r_5(u) \cdot r_1(v)$.

Proposition 11 *Let $T = (V, E)$ be a rooted phylogenetic X-tree, rooted at ρ, with a fixed edge length assignment function ℓ. Let $\varphi(x) = \sum_{e\in E} \gamma(x, e)\ell(e)$ be a diversity index on T. Then there exists a consistent diversity index $\psi_\ell(x) = \sum_{e\in E} \gamma'(x, e)\ell(e)$ such that φ and ψ_ℓ coincide.*

Proof We construct a consistent diversity index $\psi_\ell(x) = \sum_{e\in E} \gamma'(x, e)\ell(e)$ in the following manner. Let v be a vertex of T with out-degree d. We take the ratio of allocations for ψ_ℓ at v, using the coefficients from φ and the edge lengths given by ℓ, as

$$\sum_{e\in P(T;\rho,v)} \Gamma_1(v, e)\ell(e) : \cdots : \sum_{e\in P(T;\rho,v)} \Gamma_d(v, e)\ell(e). \tag{2}$$

Let $r_1(v) : \cdots : r_d(v)$ be the normalised ratio equivalent to ratio 2. We construct the normalised ratio of allocations at each interior non-root vertex of V in a similar way. Then applying the method of Lemma 10 gives the $\gamma'(x, e)$ coefficients of ψ_ℓ.

We now show that ψ_ℓ coincides with φ for an arbitrary leaf $x \in X$. By the descent condition (DI$_1$), we need only consider edges from the path $\mathcal{P} = P(T; \rho, x)$. Specifically, let $\mathcal{P}$ consist of the vertices $\{\rho, u_1, u_2, \ldots, u_k, x\}$ and edges $e_1 = (\rho, u_1)$, $e_2 = (u_1, u_2), \ldots, e_{k+1} = (u_k, x)$. Assume, without loss of generality, that x is in $T_1(v)$ for all $v \in \mathcal{P}\backslash\{\rho, x\}$.

Using ratio 2 and normalising gives values of $r_1(u_1) = \sum_{x'\in T_1(u_1)} \gamma(x', e_1)$ and $r_1(u_2) = \frac{\sum_{x'\in T_1(u_2)}(\gamma(x',e_1)\ell(e_1)+\gamma(x',e_2)\ell(e_2))}{\ell(e_2)+r_1(u_1)\ell(e_1)}$ and so on until $T_1(u_k)$ contains only the leaf x, where

$$r_1(u_k) = \frac{\sum_{e\in\mathcal{P}\backslash\{e_{k+1}\}} \gamma(x, e)\ell(e)}{\ell(e_k) + r_1(u_{k-1})\left(\ell(e_{k-1}) + r_1(u_{k-2})\left(\ldots\left(\ell(e_2) + r_1(u_1)\ell(e_1)\right)\ldots\right)\right)}.$$

By Lemma 10, the coefficients of ψ_ℓ are given by $\gamma'(x, e_i) = \prod_{v\in\{u_i,\ldots,u_k\}} r_1(v)$. So

$$\begin{aligned}
\psi_\ell(x) &= \sum_{e\in\mathcal{P}} \gamma'(x, e)\ell(e)\\
&= \ell(e_{k+1}) + r_1(u_k)\ell(e_k) + r_1(u_k)r_1(u_{k-1})\ell(e_{k-1}) + \ldots + \prod_{v\in\mathcal{P}\backslash\{\rho,x\}} r_1(v)\ell(e_1)\\
&= \ell(e_{k+1}) + r_1(u_k)[\ell(e_k) + r_1(u_{k-1})(\ell(e_{k-1})\\
&\quad + r_1(u_{k-2})\left(\ldots\left(\ell(e_2) + r_1(u_1)\ell(e_1)\right)\ldots\right))]\\
&= \sum_{e\in\mathcal{P}} \gamma(x, e)\ell(e) = \varphi(x).
\end{aligned}$$

Therefore, as x was arbitrary, the consistent diversity index ψ_ℓ coincides with φ. □

Let $T = (V, E)$ be a rooted phylogenetic $|X|$-tree with fixed edge length assignment function ℓ. To each interior non-root vertex v we associate an integer: the *degrees of freedom* of v. This value is calculated as one less than the number of distinct tree shapes across the set of maximal pendant subtrees below v (i.e. in $\{T_1(v), \dots, T_d(v)\}$). In a rooted binary phylogenetic tree, each interior non-root vertex therefore has zero or one degree of freedom.

We construct an equivalence relation $\sim$ on the set of interior non-root vertices of T. We write $u \sim v$ if and only if u and v both have out-degree d, and the multiset of tree shapes of $\{T_1(u), \dots, T_d(u)\}$ equals the multiset of tree shapes of $\{T_1(v), \dots, T_d(v)\}$. In other words, $u \sim v$ if and only if the structure of the subtree descended from u is exactly the same as that descended from v. Theorem 12 allows us to use the $\sim$-equivalence classes to determine the dimension of each diversity index space exactly.

Theorem 12 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with a fixed edge length assignment function ℓ. Let V' be a set that contains precisely one vertex from each of the $\sim$-equivalence classes of T. The dimension of the convex space $S(T, \ell)$ of diversity indices is the sum of the degrees of freedom of the vertices in V'.*

Proof For each $v \in V'$, consider the normalised ratio of allocations $r_1(v) : \cdots : r_d(v)$ (where d is the out-degree of v). Assume that v is an interior non-root vertex with zero degrees of freedom. Then to satisfy condition (DI$_2$) requires that $r_i(v) = \frac{1}{d}$ for all $i \in \{1, \dots, d\}$. Next, assume that v is an interior non-root vertex with non-zero degrees of freedom. Suppose that there are k_1 maximal pendant subtrees with the same tree shape as $T_1(v)$. Any value in the interval $\left[0, \frac{1}{k_1}\right]$ is possible for $r_1(v)$. If $T_i(v)$ has the same tree shape as $T_1(v)$, then set $r_i(v)$ equal to $r_1(v)$.

As v has non-zero degrees of freedom, there is a maximal pendant subtree, say $T_2(v)$, with a tree shape that is different from $T_1(v)$. Suppose that there are k_2 maximal pendant subtrees with the same tree shape as $T_2(v)$. If v has exactly one degree of freedom, then $r_2(v) = \frac{1-k_1 r_1(v)}{k_2}$. Otherwise, any value in the interval $\left[0, \frac{1-k_1 r_1(v)}{k_2}\right]$ is possible for $r_2(v)$, and those other terms of the ratio corresponding to a tree shape matching $T_2(v)$. This process continues until the number of values selected matches the degrees of freedom of v. Then the ratio term(s) corresponding to the last tree shape (among the maximal pendant subtrees of v) is the value which ensures that the ratio of allocations is normalised. Finally, if $u \sim v$ and $T_i(u)$ has the same tree shape as $T_j(v)$, we set $r_i(u) = r_j(v)$.

For each $\sim$-equivalence class of interior non-root vertices with non-zero degrees of freedom, we can choose the ratio terms independently. We now form a vector where each component corresponds to a choice of a ratio term by the process above. There is thus one component per degree of freedom (in total, across all equivalence classes). By Lemma 10, for each possible set of ratios there is a unique set of coefficients and thus a unique corresponding diversity index. Hence, the dimension of $S(T, \ell)$ is at least the sum of the degrees of freedom of the vertices in V'.

On the other hand, Proposition 11 shows that any diversity index in $S(T, \ell)$, say φ, coincides with a consistent diversity index, say $\bar{\varphi}$. At each vertex, the terms in the normalised ratio of $\bar{\varphi}$ must lie within the intervals described in our construction above. Otherwise, the sum of the ratio terms would add to more than one, or else some of the ratio terms would be negative. Both possibilities are excluded by the definition of a normalised ratio of allocations. Thus it is possible to construct $\bar{\varphi}$ in the manner described above. Hence, the dimension of the convex space $S(T, \ell)$ is precisely the sum of the degrees of freedom of the vertices in V'. □

We note that the sum of degrees of freedom will always be less than the number of leaves in a given rooted phylogenetic tree and express this inequality through the following result.

Proposition 13 *Let T be a rooted phylogenetic X-tree with fixed edge lengths ℓ. The dimension of $S(T, \ell)$ is at most $|X| - 2$.*

Proof The dimension of $S(T, \ell)$ is calculated by adding up the degrees of freedom in each $\sim$-equivalence class. We show that even when adding together degrees of freedom from every interior non-root vertex (as opposed to just one per $\sim$-equivalence class), that the sum of degrees of freedom always remains at most $|X| - 2$.

Let $n = |X|$ and let $df(T)$ be the total degrees of freedom summed across every interior non-root vertex in T. Then $df(T)$ is at least the dimension of $S(T, \ell)$. Furthermore, let $\mathcal{I}(T)$ be the number of interior non-root vertices of T and let $i(d)$ be the number of interior vertices of T that have out-degree d. We consider every combination of interior vertices, in terms of their out-degrees. The total $df(T)$ value is comprised of, at most, one degree of freedom per interior non-root vertex, plus an extra degree of freedom for each third and subsequent edge originating from an interior non-root vertex. In symbols, $df(T) \leq \mathcal{I}(T) + \sum_{d=3}^{n} i(d)(d-2)$.

The number of interior non-root vertices of T can be determined by starting with an appropriate binary X-tree T' and contracting edges to form T. The number of such contractions required is $d - 2$ for each interior vertex with degree $d \geq 3$ in T and each reduces the number of interior vertices by one. As every binary X-tree has $n - 2$ interior non-root vertices, $\mathcal{I}(T) = n - 2 - \sum_{d=3}^{n} i(d)(d-2)$.

Finally, this gives $df(T) \leq n-2-\sum_{d=3}^{n} i(d)(d-2)+\sum_{d=3}^{n} i(d)(d-2) = n-2$. Therefore the total number of degrees of freedom of T, and hence $S(T, \ell)$, is at most $|X| - 2$. □

6.1 Example: Application to a Tree of Hominoids

A brief illustration of this is provided by the rooted phylogenetic tree of Hominoids appearing in Fig. 6. For this tree, each of the nine labelled interior vertices is a representative of a $\sim$-equivalence class that has one degree of freedom. There are also two equivalence classes among the unfilled circle vertices, each with zero degrees of freedom. Hence, there are nine degrees of freedom overall for this tree, and the diversity index space for the tree of hominoids has nine dimensions. In other words, any diversity index on this tree is required to specify the ratio of allocation at these nine labelled representative vertices in order to determine its full set of coefficients.

The presence of a nine-dimensional diversity index space here does not particularly illuminate anything about hominoid-specific diversity. However it does illustrate the point that for even quite a modestly-sized phylogenetic tree as this, a large number of

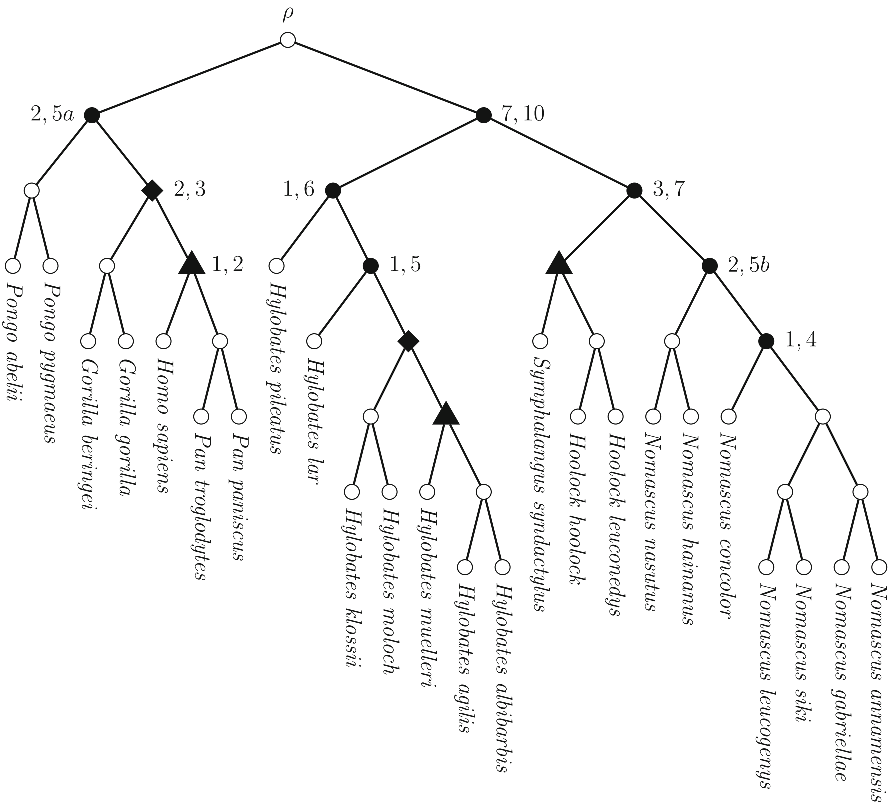


Fig. 6 The phylogenetic tree of the superfamily of Hominoids (great apes and gibbons). This tree was constructed using data from Springer et al. (2012); Carbone et al. (2014), via OneZoom Core Team (2021). Filled vertices represent interior vertices that have one degree of freedom. Each numerical label indicates a representative of each $\sim$-equivalence class. The two diamond-shaped vertices both belong to the 2,3 class, and the three triangular vertices belong to the 1,2 class. Note that the vertices labelled 2,5a and 2,5b lie in distinct $\sim$-equivalence classes because of the difference in tree shape between their maximal pendant subtrees with five leaves. Thus, there are nine equivalence classes, contributing one degree of freedom each

potential diversity indices exist. Moreover, most potential diversity indices that could be applied to the hominoid tree are not directly related to the known FP and ES indices.

The implication for conservation is that more investigation of the diversity index space is required to ensure optimal use of the diversity index concept. This would involve testing various properties of candidate diversity indices against the assumption that FP or ES is the most suitable index (from among the infinitely many possibilities). Doing so could help establish whether FP, ES or some new diversity index (or indices) best give a meaningful and robust measurement of a taxon's phylogenetic isolation and embodiment of shared and unique evolutionary history.

6.2 Boundaries of Index Spaces

We give an example of a two-dimensional diversity index space on the rooted caterpillar tree on five leaves (Fig. 5), which we denote here as Cat_5. Let φ be a diversity index

Table 2 The corners of $S(Cat_5, \mathbf{1})$ are determined by taking the four possible combinations of extreme ratios at s and t

Index	Ratio at s	Ratio at t
κ	1:0	1:0
λ	1:0	0:1
μ	0:1	1:0
ν	0:1	0:1

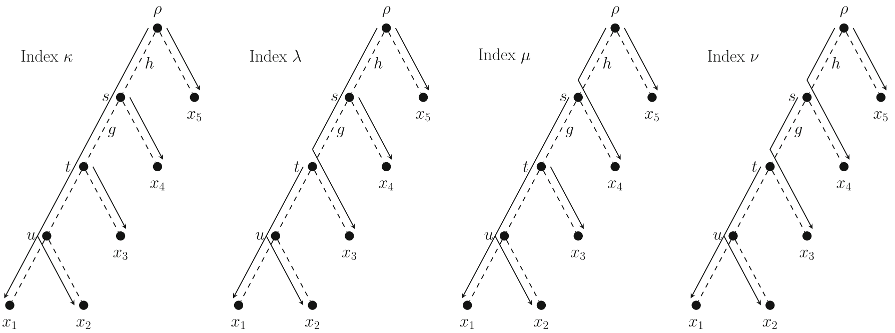


Fig. 7 Arrows show the allocations for the four diversity indices on Cat_5 (indicated by dashed edges) that lie at the corners of $S(Cat_5, \mathbf{1})$

on Cat_5, and suppose that every edge in Cat_5 has unit length; we denote this as $\boldsymbol{\ell} = \mathbf{1}$. By taking the extreme ratios of allocation at vertices s and t (see Table 2), we find the diversity indices κ, λ, μ and ν (Fig. 7, index score vectors listed below) that lie at the boundary points of $S(Cat_5, \mathbf{1})$:

$$\kappa = [2.5, 2.5, 1, 1, 1], \qquad \lambda = [1.5, 1.5, 3, 1, 1],$$
$$\mu = [2, 2, 1, 2, 1], \qquad \nu = [1.5, 1.5, 2, 2, 1].$$

Similarly, the 'corner' indices of a diversity index space may be found by taking the extreme ratios of allocation for each $\sim$-equivalence class in the corresponding phylogenetic tree. We can then use these corner indices to obtain further diversity indices by way of linear combination. Carathéodory's Theorem (this version was drawn from Steinitz (1914)) describes how each point of a convex space may be described as a combination of a limited number of points.

Theorem 14 (Carathéodory's Theorem) *Let A be a non-empty subset of $\mathbb{R}^d$. Every vector from the convex hull of A can be represented as a convex combination of, at most, $d + 1$ vectors from A.*

Specifically, for a rooted phylogenetic tree T, we can construct the required $d + 1$ vectors as follows. If T is semi-balanced, then only one vector is required, so we can choose that belonging to FP. Otherwise, for each $\sim$-equivalence class, we choose an extreme ratio of allocation and let the resulting diversity index be called B_1. Next, we construct a new diversity index B_2 by matching the ratios of allocation from B_1, except that for precisely one of the $\sim$-equivalence classes, an alternative extreme ratio of allocation is chosen. Exactly d such diversity indices may be constructed in addition to B_1, one for each degree of freedom in the tree. The resulting diversity indices $B_1, \ldots, B_{d+1}$ cover the total degrees of freedom of T. Hence $B_1, \ldots, B_{d+1}$ are the required $d+1$ diversity indices, from which all others in $S(T, \ell)$ may be constructed.

Continuing our example, $S(Cat_5, \mathbf{1}) \subset \mathbb{R}^2$ is itself convex. Therefore the vector of index scores of any diversity index in this space may be expressed as a convex combination of at most three points of $S(Cat_5, \mathbf{1})$. Specifically, we can express any diversity index φ for Cat_5 as

$$\begin{aligned}\varphi &= (1-p-q)\kappa + p\lambda + q\mu \\ &= \kappa + p(\lambda - \kappa) + q(\mu - \kappa) \\ &= [2.5, 2.5, 1, 1, 1] + p[-1, -1, 2, 0, 0] + q[-0.5, -0.5, 0, 1, 0],\end{aligned}$$

where $0 \leq p, q \leq 1$, provided that $p + \frac{q}{2} \geq 1$. FP is given by $(p, q) = \left(\frac{7}{24}, \frac{1}{4}\right)$, and ES is given by $(p, q) = \left(\frac{9}{24}, \frac{1}{2}\right)$. These positions are noted in Fig. 5.

7 Concluding Remarks

In this paper we have investigated the combinatorial and geometric properties of the space of phylogenetic diversity indices, having defined these in a very general way. The benefit of doing so is that one can more readily investigate which properties of known diversity indices hold generally and which are specific to a restricted class of diversity indices. Understanding the properties of diversity indices may be useful when deciding which index to use in a particular setting. For example, we have discussed the continuity property of the Fair Proportion index and shown that it is unique to that index.

We briefly present two further properties that may be of interest in this regard. Each are held by both the FP and ES indices. Let $T = (V, E)$ be a rooted phylogenetic X-tree, and let $e = (u, v) \in E$.

- Property 1: the ratio of allocations $\Gamma_1(v, e) : \cdots : \Gamma_d(v, e)$ depends only on the number of leaves in each of the d maximal pendant subtrees below e, and not their structure.
- Property 2: $\Gamma_i(v, e) \geq \Gamma_j(v, e)$ whenever $T_i(v)$ contains at least as many leaves as $T_j(v)$.

We have focussed on a particular structure, namely that of rooted phylogenetic trees. For consistent diversity indices, the ability to view their calculation as a flow problem allows a straightforward extension of this framework to rooted phylogenetic networks. In the more general setting of a phylogenetic network, we need only to add the stipulation that the flow into a so-called reticulation vertex is matched by the flow out from that vertex. A second more general approach could be developed by applying allocation functions to unrooted phylogenetic trees. It is for this reason that we have presented the definition of a diversity index as a subclass of allocation functions, although investigation of allocating PD among leaves of unrooted trees has been left for further work.

Acknowledgements The authors were supported by the New Zealand Marsden Fund (MFP-UOC2005). We thank Martin Frohn for useful suggestions regarding both Proposition 5 and specifying a 'basis' for diversity index spaces.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. The authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

References

Cadotte MW, Jonathan Davies T (2010) Rarest of the rare: advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. Divers Distrib 16(3):376–385

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B et al (2014) Gibbon genome and the fast karyotype evolution of small apes. Nature 513(7517):195–201

Crozier RH (1992) Genetic diversity and the agony of choice. Biol Cons 61(1):11–15

EDGE of existence programme: edge of existence: evolutionarily distinct and globally endangered (2022). www.edgeofexistence.org Accessed 3 July 2022

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Cons 61:1–10

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA, USA

Fischer M, Francis A, Wicke K (2022) Phylogenetic diversity rankings in the face of extinctions: the robustness of the fair proportion index. Syst Biol 72(3):606–615

Fuchs M, Jin EY (2015) Equality of shapley value and fair proportion index in phylogenetic trees. J Math Biol 71:1133–1147

Gumbs R, Gray CL, Böhm M, Burfield IJ, Couchman OR, Faith DP, Forest F, Hoffmann M, Isaac NJB, Jetz W et al (2023) The EDGE2 protocol: advancing the prioritisation of evolutionarily distinct and globally endangered species for practical conservation action. PLoS Biol 21(2):e3001991

Haake C-J, Kashiwada A, Su FE (2008) The Shapley value of phylogenetic trees. J Math Biol 56(4):479–497

Hartmann K (2013) The equivalence of two phylogenetic biodiversity measures: the Shapley value and fair proportion index. J Math Biol 67(5):1163–1170

Isaac NJ, Turvey ST, Collen B, Waterman C, Baillie JE (2007) Mammals on the edge: conservation priorities based on threat and phylogeny. PLoS ONE 2(3):296

OneZoom core team: OneZoom tree of life explorer version 3.5 (2021). www.onezoom.org Accessed 3 July 2022

Palmer C, Fischer B (2021) Should global conservation initiatives prioritize phylogenetic diversity? Philosophia, 1–20

Redding D (2003) Incorporating genetic distinctness and reserve occupancy into a conservation priorisation approach. Master's thesis: University of East Anglia (2003)

Redding DW, Mazel F, Mooers AØ (2014) Measuring evolutionary isolation for conservation. PLoS ONE 9(12):113490

Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C, Ryder OA, Janečka JE et al (2012) Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7(11):49521

Steel M (2016) Phylogeny: discrete and random processes in evolution. SIAM, Philadelphia, PA, USA

Steinitz E (1914) Bedingt konvergente Reihen und konvexe Systeme. J für die reine und angewandte Mathematik 144:1–40

Tucker CM, Cadotte MW, Carvalho SB, Davies TJ, Ferrier S, Fritz SA, Grenyer R, Helmus MR, Jin LS, Mooers AO et al (2017) A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol Rev 92(2):698–715

Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect?-systematics and the agony of choice. Biol Cons 55(3):235–254

Wicke K (2020) Novel aspects of mathematical phylogenetics. PhD thesis, Universität Greifswald

Wicke K, Steel M (2020) Combinatorial properties of phylogenetic diversity indices. J Math Biol 80(3):687–715