Spaces of phylogenetic diversity indices: combinatorial and geometric properties

Kerry Manson and Mike Steel

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand.

Contributing authors: kerry.manson@pg.canterbury.ac.nz;

Abstract

Biodiversity is a concept most naturally quantified and measured across sets of species. Yet 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 popular diversity indices in current use. This paper presents conditions that define diversity indices arising from the phylogenetic diversity (PD) 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, whose boundary conditions are determined by the underlying shape of each phylogenetic tree. We calculate the dimension of the convex space associated with each tree shape, and describe 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 [1]. Applying quantitative measures to such trees has proved useful in understanding these relationships, and for highlighting conservation priorities in the face of the current human-induced mass extinction of species [2]. A large number of these quantitative approaches have been developed (see [3] for an overview).

Phylogenetic diversity (PD) is a measure that aims to quantify the biodiversity exhibited by a set of species [4]. PD does so using a weighted phylogenetic tree that exhibits the evolutionary relationships between 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 together the weights of all the edges in that tree. 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 to rank species by their 'distinctiveness' [5], 'evolutionary isolation' [6], or to quantify their 'combination of unique and shared evolutionary heritage' [7]. Doing so provides evidence towards developing conservation priorities.

The usual means for 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' [8], a characterisation that we shall utilise.[1] Common diversity indices (highlighted below) are the Fair Proportion (FP) index and the Equal-Splits (ES) index. The former (under the name Evolutionary Distinctiveness) is used by the EDGE of Existence programme [11–13] to help rank species by their need for conservation assistance. A recent paper by Palmer and Fischer [14] evaluates the effects of this implementation of Evolutionary Distinctiveness on conservation efforts.

While 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 known diversity indices and further possibilities. In this paper we frame an allocation of the PD value in terms of a set of coefficients that determine it. We then give a pair of conditions on these coefficients that satisfy natural allocation requirements within a simple evolutionary model. These conditions define *phylogenetic diversity index* functions in a general sense. Some further constraints on the coefficients are derived as immediate consequences of our diversity index definition. 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

[1] Alternative ranking approaches, such as the methods of [9] or [10], 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.

show how the topology of the underlying tree (the 'tree shape') determines the dimension and boundaries of its associated diversity index space.

The Fair Proportion and Equal-Splits indices are thus re-cast 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 between these possibilities. We illustrate these results by describing in detail the convex spaces associated with the three rooted binary tree shapes containing five leaves, and by including a case study on the phylogenetic tree of Hominoids (great apes and gibbons).

With this method of determining every diversity index on a particular tree, we can investigate properties of certain diversity indices and see whether or not these properties are unique to that index or widely held. For example, we use our diversity index conditions to show that the Fair Proportion 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, that we call 'consistency', 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 for 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 shall 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 write P_e for the pendant subtree formed by deleting the edge e.

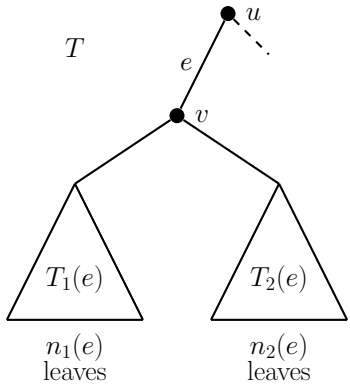


Fig. 1: Schematic diagram showing notation for maximal pendant subtrees descended from edge e. Vertex v has out-degree 2, but in general we allow larger out-degrees.

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 $c_T(e)$.

Let e be an interior edge of T whose terminal vertex v has out-degree d. We represent the d maximal pendant subtrees contained within P_e by $T_1(e), T_2(e), \ldots, T_d(e)$. Depending on 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 $\ell : E(T) \to \mathbb{R}^{\geq 0}$ 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 restriction on the edge lengths besides 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 frequently consider in this paper are the Fair Proportion (FP) index [13, 15] and the Equal-Splits (ES) index [6, 15]. 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* (FP) 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* (ES) 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 to define 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 change common to the cluster of species $c_T(e)$ is explained by a change occurring somewhere along edge e. As such, the process of speciation aligns with the tree shape of a rooted phylogenetic tree, where interior vertices correspond to speciation events.

It is against this interpretation (and the observations above) that we test the conditions that will define diversity indices. The broad goal of any phylogenetic diversity index is to take the PD score of a rooted phylogenetic tree overall and distribute this value among the species in a way that is compatible with the tree shape. The distinctions between individual diversity indices should reflect different assumptions about how species exhibit ancestral developments, while respecting the above noted observations.

We begin in Section 3.1 by describing a more general class of function, here called allocation functions, as well as the coefficient notation that will be used. The further conditions required of diversity indices are then added in Section 3.2. Our main definition on page 8 encompasses the Fair Proportion and Equal-Splits indices defined above, and allows for the description of further diversity indices. 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 equation

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

and moreover may be expressed as $\varphi_\ell(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$, for every edge length assignment function ℓ, and where all of the *coefficients* $\gamma(x, e)$ are greater than or equal to zero.

We call the value $\varphi_\ell(x)$ the *φ-score* of leaf x (given length assignment ℓ). Our aim is to define diversity indices that not only deliver an ordinal ranking of species according to their contribution to phylogenetic diversity, but also 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 Equation (1), and is reinforced by requiring allocation functions 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 particular 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 clear from context.

Equation (1) is a weak enough constraint that defining allocation functions by this property alone would include 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 following 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, like 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 (for real-valued coefficients $\gamma(x, e)$) 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 $s, t \in \mathbb{R}$ 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 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

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

□

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 $\gamma(x, e)$. We note that the Fair Proportion and Equal-Splits indices satisfy Definition 1, and hence are allocation functions. For these indices, the coefficients may be read directly from their definitions. That is, as $\frac{1}{|c_T(e)|}$ and $\frac{1}{\Pi(e,x)}$, respectively, when $x \in c_T(e)$, and noting that for both indices $\gamma(x, e) = 0$ whenever x is not descended from e. In contrast, we can define 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.

We note that on 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 known diversity indices. For instance, Wicke and Steel [7] characterise the rooted tree shapes for which FP and ES 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 will be a subclass of allocation functions that does take into account 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 to species descended from e exclusively.

Additionally, the set of coefficients for an edge should be entirely determined by the shape of its descendent tree structure, and 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. Also, recall that P_e denotes the pendant subtree formed by deleting the edge e.

Definition 2 Let $T = (V, E)$ and 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 tree shape. If leaves x in P_e and y in P_f appear in corresponding positions in their respective subtrees, then $\gamma(x, e) = \gamma(y, f)$.

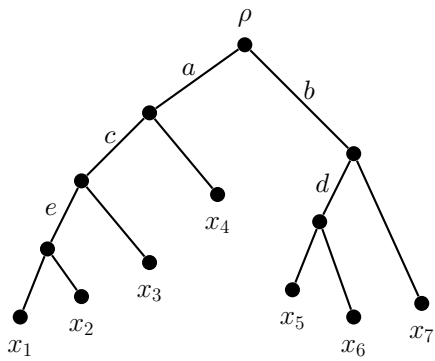


Fig. 2: Any diversity index on the above tree must allocate the length of edge c exclusively among its descendent cluster: $\{x_1, x_2, x_3\}$. The pattern of this allocation must be matched by the allocation of the length of edge b among b's own descendant cluster, namely $\{x_5, x_6, x_7\}$. This is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

Consequently, to satisfy the neutrality condition (DI_2), any diversity index on the tree in Figure 2 is required to satisfy all of the following coefficient equalities: $\gamma(x_1, a) = \gamma(x_5, a)$; $\gamma(x_5, b) = \gamma(x_6, b) = \gamma(x_1, c) = \gamma(x_2, c)$; $\gamma(x_7, b) = \gamma(x_3, c)$; $\gamma(x_5, d) = \gamma(x_6, d)$; and $\gamma(x_1, e) = \gamma(x_2, e)$.

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 [16] showed that for any function $\varphi_\ell(x) = \sum_{e\in E} \gamma(x,e)\ell(e)$ that satisfies Equation 1, the coefficients belonging to each edge sum to one. Lemma 2 reframes this result slightly to apply 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 $\gamma(x', e) > 1$ for some edge e and leaf x'. From the definition of allocation functions, all other coefficients of φ are non-negative, and so $\sum_{x\in X} \gamma(x,e) > 1$. But 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 condition (DI_1). Lastly, let e be a pendant edge. Suppose $x' \in X$ is incident with e, giving $c_T(e) = \{x'\}$. So, by part (ii), $1 = \sum_{x\in c_T(e)} \gamma(x,e) = \gamma(x', e)$. Now suppose x' is not incident with e. Then x' is not descended from e, and 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 resultant 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 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. If φ is a diversity index on T, then $\ell(e_x) \leq \varphi(x) \leq \sum_{e \in P(T;\ \rho,x)} \ell(e)$ 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 $e \in E - e_x$. This is the minimal possible choice, as Prop 3(iii) ensures that $\gamma(x, e_x) = 1$. In this case, $\varphi(x) = \gamma(x, e_x)\ell(e_x) + \sum_{e \in (E \setminus \{e_x\})} \ell(e) = \ell(e_x) + 0$.

Next, for some edge $e \in E$, suppose that $\gamma(x, e)$ is nonzero. By condition (DI_1), a non-zero coefficient means x is descended from e. Equivalently, $e \in P(T;\ \rho, x)$. Then

$$\varphi(x) = \sum_{e \in P(T;\ \rho,x)} \gamma(x, e)\ell(e) + \sum_{e \in (E \setminus P(T;\ \rho,x))} \gamma(x, e)\ell(e) \leq \sum_{e \in P(T;\ \rho,x)} 1 \cdot \ell(e) + 0.$$

Therefore $\ell(e_x) \leq \varphi(x) \leq \sum_{e \in P(T;\ \rho,x)} \ell_e$ for all $x \in X$. □

These bounds may be sharp for some diversity indices (an example is provided later by the indices λ and μ in Figure 7). Yet for some leaves, such as x_1 in the same figure, the bounds are never attained due to the constraints imposed by condition (DI_2).

4 Continuity

In [17][p. 140] the following property of the Fair Proportion 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 properties of diversity indices, we show that the Fair Proportion index 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 the above condition as the *diversity index continuity condition.*

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 $T = (V, E)$, every $x \in X$ and every interior edge $e \in E(T)$ if and only if ψ is the Fair Proportion index.*

Proof Suppose ψ is a diversity index that satisfies the diversity index continuity condition 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 conditions (DI_1) and (DI_2). The index ψ may be defined by its set of coefficients, and moreover we need only to specify those coefficients that are not already determined by the diversity index definition.

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 condition

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

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 condition,

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

In $T/D(e_i)$ no interior edge is descended from e_i. Hence, we again have $\gamma(x, e_i) = \frac{1}{n_i}$. Since the coefficients are unaffected by the contraction process, $\gamma(x, e_i) = \frac{1}{n_i}$ in T too. 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 $e \in E(T)$. The above argument shows that in every case $\gamma(x, e_i) = \frac{1}{n_i} = \frac{1}{|c_T(e_i)|}$, exactly the set of coefficients that defines the Fair Proportion index on T. Therefore ψ is the Fair Proportion index.

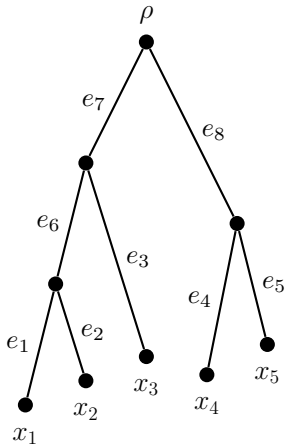
Conversely, suppose ψ is the Fair Proportion index. Then the coefficients $\gamma(x, e)$ depend only on the number of leaves descended from edge e, and not on the particular structure of the phylogenetic tree below e. The contraction of any interior edge e_i does not reduce the number of leaf vertices 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 condition holds. □

Fair Proportion therefore may be considered the most appropriate diversity index to apply to rooted phylogenetic trees that are not fully resolved, because later refinements of the tree will not impact the initial FP index scores.

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 T given by



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

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).

$\varphi(x) = \sum_{e \in E} \gamma(x, e)\ell(e)$ 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 $a_{ij} = \gamma(e_j, x_i)$. Consider the tree in Figure 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, Fair Proportion uses $\alpha = \frac{1}{3}$ here and Equal-Splits 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 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 phylogenetic 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 by $S(T)$, contains all index score vectors of T. In the case where T has a fixed edge length assignment function ℓ, we use the notation $S(T, \ell)$.

Note that $S(T)$ is not a vector space since $\mathbf{0}$ is not in $S(T)$, however $S(T)$ is convex.

Proposition 7 *Let $T = (V, E)$ be a rooted phylogenetic X-tree with edge length assignment function ℓ. Then $S(T)$, the space of index score vectors on T, 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, $\boldsymbol{u}, \boldsymbol{v} \in S(T)$.

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 $1 \leq j \leq m$. Then

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

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 Prop 1 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 both conditions (DI_1) and (DI_2). If x_i is not descended from e_j, then $\gamma(x_i, e_j) = 0$ and $\gamma'(x_i, e_j) = 0$. So $\delta(x_i, e_j) = t \cdot 0 + (1-t) \cdot 0 = 0$, and the coefficient $\delta(x_i, e_j)$ satisfies (DI_1).

Next assume that, for some $x, y \in X$ and $e, f \in E$, satisfying condition (DI_2) requires $\gamma(x, e) = \gamma(y, f)$. 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 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 $S(T)$. Therefore $S(T)$ 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, but it will be shown (Proposition 13) that $S(T, \ell) \subset \mathbb{R}^k$, for some $k < n$. The smallest such k we call the *dimension* of $S(T)$. Diversity indices are completely described by their coefficients rather than edge lengths, so the dimension of $S(T)$ 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.)

5.1 Examples, and the special case when $S(T)$ 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 Figure 3. By Proposition 3, the pendant edge lengths are entirely allocated to their incident leaves. By condition (DI_2), the length of edge f must be shared equally between x_1 and x_2 in every diversity index. Thus, for this tree only the allocation of edge g between x_1, x_2 and x_3 changes. Suppose that $\gamma(x_3, g) = \alpha$ is the share of the length of g allocated to x_3. Accordingly, $\gamma(x_1, g) + \gamma(x_2, g) + \alpha = 1$. The share of the length of g allocated to x_1 must equal the share of the length of g allocated to x_2, so $\gamma(x_1, g) = \gamma(x_2, g) = \frac{1}{2}(1 - \alpha)$. 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 diversity indices on the non-binary tree appearing in Figure 4a.

For some tree shapes however, condition (DI_2) provides enough of a constraint that all diversity indices coincide. One such shape is given by the tree in Figure 4b, with the necessary index scores appearing in Table 1c. A *balanced* tree is a rooted phylogenetic X-tree in which, for every vertex v, the pendant subtrees $T_1(v), \ldots, T_d(v)$ all have the same tree shape as each other (where d is the out-degree of v). 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 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. So by Proposition 4, $\gamma(x, e) = \gamma(x', e)$ for any x, x' in $c_T(e)$. So for each leaf x descended from e, we can set $\gamma(x, e) = \frac{1}{|c_T(e)|}$. The coefficients belonging to every interior edge of T can be determined in this way, and those of pendant edges are fixed by Proposition 3 (iii). Hence there is only one possible set of index scores, that is $S(T)$ has dimension zero.

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

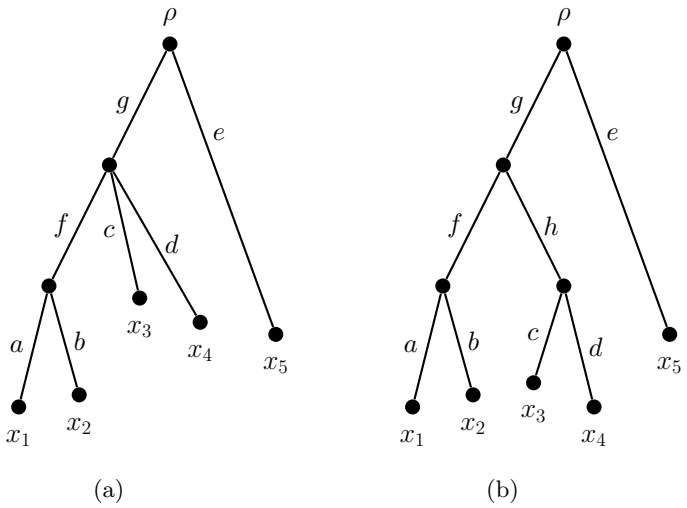


Fig. 4: Two rooted phylogenetic tree shapes on five leaves.

x	$\varphi(x)$
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	$\ell(e)$

(a)

x	$\varphi(x)$
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	$\ell(e)$

(b)

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

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 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, use the Equal-Splits 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 condition (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 $S(T)$ contains more than one index, and cannot consist of a single point. $\square$

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

Proof Wicke and Steel [7, Theorem. 4] show 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 to the non-binary case directly. By Proposition 8, T being semi-balanced is equivalent to $S(T)$ consisting of a single point. That is, all diversity indices on T are coincident. □

6 Consistent diversity indices

The coefficients of the Fair Proportion and Equal-Splits 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 rather in their ratios. For example, consider the tree in Figure 5. The FP allocation of edge g to $\{x_1, x_2\}$ compared to $\{x_3\}$ is $\frac{2}{3} : \frac{1}{3}$, while 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 using the ES index, the allocation between these same sets follows a $1 : 1$ ratio for both g and h. We say that an index that has this property is *consistent*. 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 belonging to 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 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 $\Gamma_1(v, e) : \cdots : \Gamma_d(v, e)$ shall be 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 Figure 5 requires the ratio of allocation $[\gamma(x_1, g) + \gamma(x_2, g)] : \gamma(x_3, g)$ to be the same as $[\gamma(x_1, h) + \gamma(x_2, h)] : \gamma(x_3, h)$.

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, to move weight from each interior edge to its immediate descendant edges. The algorithm begins with edges incident to the root vertex, and continues to re-weight edges until

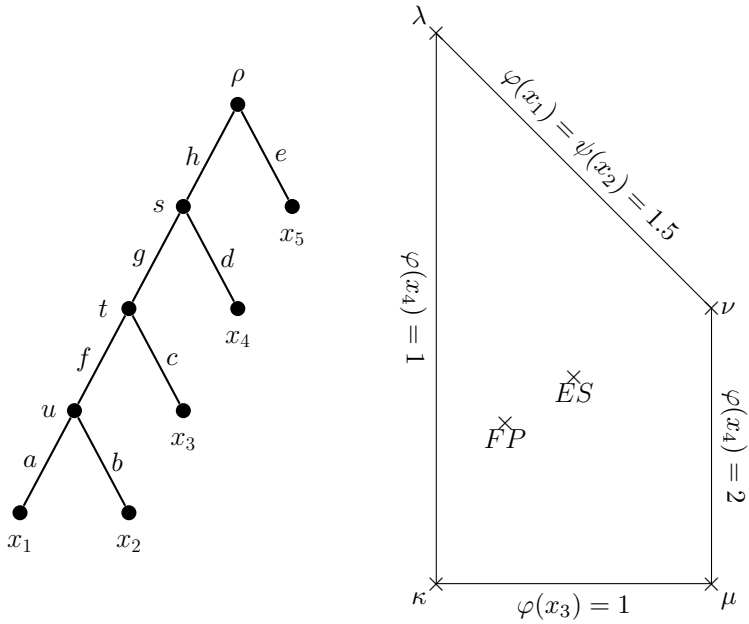


Fig. 5: Left: The five-leaf rooted caterpillar tree, Cat_5, where each edge has unit length. Right: The diversity index space $S(Cat_5, \mathbf{1})$, projected onto the $\varphi(x_3)\varphi(x_4)$-plane, with boundary conditions and extremal indices $\kappa, \lambda, \mu, \nu$ at the corner points. Notice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

all interior edges have zero weight. 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 [18] 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 $e = (u_1, u_2)$ be an edge of T. Without loss of generality assume $x \in X$ is in $T_1(v)$ for all $v \in P(T;\ u_2, x)$. Then $\gamma(x, e)$ is given by the product $\prod_{v \in P(T;\ u_2, x)} r_1(v)$. 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 above calculation in order to clarify the final sentence of the above 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). Further suppose 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 Let v be a vertex of T with out-degree d. Then the ratio of allocations at v is

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

After determining the ratio of allocations at each vertex $v \in V$, we can apply the method of Lemma 10 to calculate the $\gamma'(x, e)$ coefficients of ψ_ℓ, that by construction 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 $\{T_1(v), \ldots, T_d(v)\}$). 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 $\{T_1(u), \ldots, T_d(u)\}$ equals the multiset of tree shapes of $\{T_1(v), \ldots, T_d(v)\}$. 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 exactly the dimension of each diversity index space.

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 $\sim$-equivalence class 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 its 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 $i \in \{1, \ldots, d\}$. 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 some maximal pendant subtree, say $T_2(v)$, with a tree shape different from $T_1(v)$. Suppose 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 the 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 maximal pendant subtrees of v) is that which ensures the ratio of allocations is normalised. Finally, if $u \sim v$ and $T_i(u)$ has the same tree shape as $T_j(v)$, set $r_i(u) = r_j(v)$.

For each $\sim$-equivalence class of interior non-root vertex with non-zero degrees of freedom, we can make the above choices of ratio terms independently. We now form a vector where each component corresponds to a choice of a ratio term from 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 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 $S(T, \ell)$ of diversity indices 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 less than $|X|$.*

Proof Suppose, with a view to contradiction, that the dimension of $S(T)$ is greater than $|X|$. We are free to assume the strongest case, where each interior non-root vertex of T lies in its own $\sim$-equivalence class. The dimension of $S(T)$ is then found by adding together the degrees of freedom of every interior non-root vertex of T. There are at most $|X| - 2$ such vertices of T. Hence there must exist some interior non-root vertex v such that the degrees of freedom of v is greater than $|c_T(v)|$.

The degrees of freedom at v is one less than the number of distinct tree shapes across the set of maximal pendant subtrees below v (i.e. in $T_1(v), \ldots, T_d(v)$). However, $d > 1$ distinct tree shapes contain at least $1 + 2 + 3(d-2) = 3(d-1)$ leaves,

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

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.

which is clearly greater than the $d-1$ degrees of freedom arising from these tree shapes. So no such vertex v exists where the degrees of freedom of v is greater than the size of $c_T(v)$, and therefore the dimension of $S(T)$ is less than $|X|$. □

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 Figure 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.

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 (Figure 5), that we denote here by Cat_5. Let φ be a diversity index on Cat_5, and suppose that every edge in Cat_5 has unit length; we denote this by $\boldsymbol{\ell} = \mathbf{1}$. By taking the extreme ratios of allocation at vertices s and t (see Table 2), we find the indices listed below 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. Then we may use these corner indices to obtain further diversity indices, by way of linear combination. Carathéodory's Theorem (this version due to Steinitz [22]) 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 nonempty subset of $\mathbb{R}^n$. Every vector from the convex hull of A can be represented as a convex combination of at most $n+1$ vectors from A.*

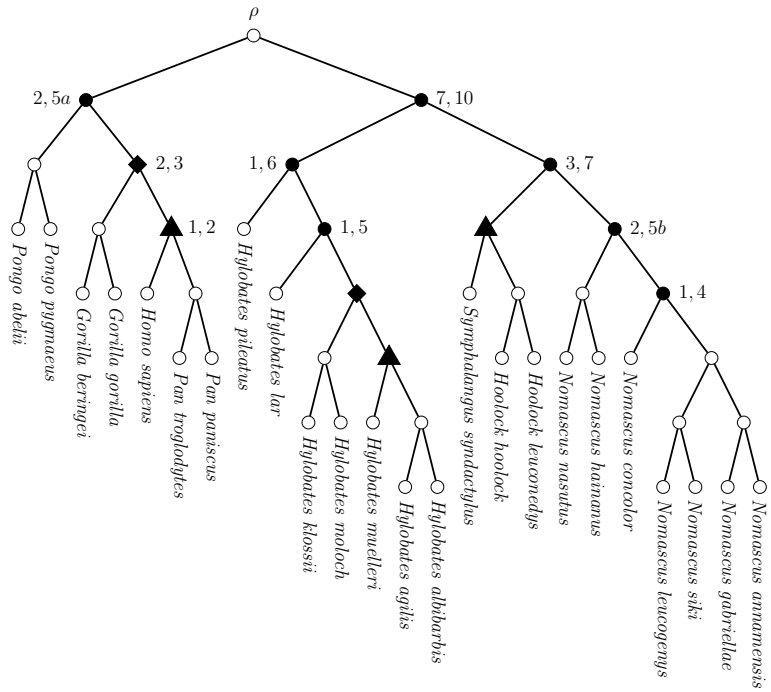


Fig. 6: The phylogenetic tree of the superfamily of Hominoids (great apes and gibbons). This tree was constructed using data from [19, 20], via [21]. Filled-in vertices represent interior vertices that have one degree of freedom. Each numerical label indicates a representative of each $\sim$-equivalence class. The two diamond vertices both belong to the 2,3 class, and the three triangle 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 a total of nine equivalence classes contributing one degree of freedom each.

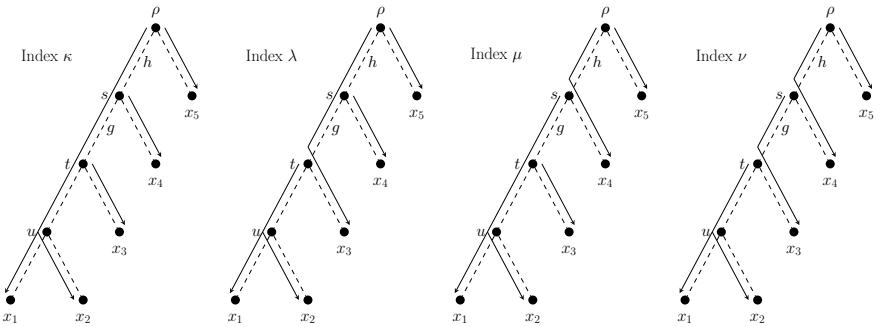


Fig. 7: Arrows show the allocations for the four diversity indices on Cat_5 (indicated by dashed edges) that lie at the corners of $S(Cat_5, \mathbf{1})$.

Continuing our example, $S(Cat_5, \mathbf{1}) \subset \mathbb{R}^2$ is itself convex. So 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 $p + \frac{q}{2} \geq 1$. Fair Proportion is given by $(p, q) = \left(\frac{7}{24}, \frac{1}{4}\right)$, and Equal-Splits is given by $(p, q) = \left(\frac{9}{24}, \frac{1}{2}\right)$. These positions are noted in Figure 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 we may 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 Fair Proportion and Equal-Splits 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, 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 of 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 presented the definition of diversity index as a subclass of allocation functions, although investigation of allocating PD among leaves of unrooted trees is left for further work.

Acknowledgments. The authors were supported by the New Zealand Marsden Fund (MFP-UOC2005).

References

[1] Felsenstein, J.: Inferring Phylogenies. Sinauer associates, Sunderland, MA, USA (2004)

[2] Cadotte, M.W., Jonathan Davies, T.: Rarest of the rare: advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. Diversity and Distributions **16**(3), 376–385 (2010)

[3] Tucker, C.M., Cadotte, M.W., Carvalho, S.B., Davies, T.J., Ferrier, S., Fritz, S.A., Grenyer, R., Helmus, M.R., Jin, L.S., Mooers, A.O., *et al.*: A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biological Reviews **92**(2), 698–715 (2017)

[4] Faith, D.P.: Conservation evaluation and phylogenetic diversity. Biological Conservation **61**, 1–10 (1992)

[5] Hartmann, K.: The equivalence of two phylogenetic biodiversity measures: the shapley value and fair proportion index. Journal of Mathematical Biology **67**(5), 1163–1170 (2013)

[6] Redding, D.W., Mazel, F., Mooers, A.Ø.: Measuring evolutionary isolation for conservation. PLoS One **9**(12), 113490 (2014)

[7] Wicke, K., Steel, M.: Combinatorial properties of phylogenetic diversity indices. Journal of Mathematical Biology **80**(3), 687–715 (2020)

[8] Fischer, M., Francis, A., Wicke, K.: Phylogenetic Diversity Rankings in the Face of Extinctions: The Robustness of the Fair Proportion Index. Systematic Biology (2022)

[9] Crozier, R.H.: Genetic diversity and the agony of choice. Biological conservation **61**(1), 11–15 (1992)

[10] Vane-Wright, R.I., Humphries, C.J., Williams, P.H.: What to protect?—systematics and the agony of choice. Biological conservation **55**(3), 235–254 (1991)

[11] EDGE of Existence Programme: www.edgeofexistence.org (2022). Accessed 2022-07-03

[12] Gumbs, R., Gray, C.L., Böhm, M., Burfield, I.J., Couchman, O.R., Faith, D.P., Forest, F., Hoffmann, M., Isaac, N.J., Jetz, W., et al.: The EDGE2 protocol: advancing the prioritisation of evolutionarily distinct and globally endangered species for practical conservation action. PLoS Biology (In Press)

[13] Isaac, N.J., Turvey, S.T., Collen, B., Waterman, C., Baillie, J.E.: Mammals on the edge: conservation priorities based on threat and phylogeny. PloS one **2**(3), 296 (2007)

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

[15] Redding, D.: Incorporating genetic distinctness and reserve occupancy into a conservation priorisation approach. University of East Anglia (2003)

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

[17] Steel, M.: Phylogeny: Discrete and Random Processes in Evolution. SIAM, Philadelphia, PA, USA (2016)

[18] Haake, C.-J., Kashiwada, A., Su, F.E.: The shapley value of phylogenetic trees. Journal of mathematical biology **56**(4), 479–497 (2008)

[19] Springer, M.S., Meredith, R.W., Gatesy, J., Emerling, C.A., Park, J., Rabosky, D.L., Stadler, T., Steiner, C., Ryder, O.A., Janečka, J.E., *et al.*: Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PloS one **7**(11), 49521 (2012)

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

[21] OneZoom Core Team: OneZoom Tree of Life Explorer Version 3.5. http://www.onezoom.org (2021)

[22] Steinitz, E.: Bedingt konvergente reihen und konvexe systeme. (1914)