Ernst Moritz Arndt Universität Greifswald
Faculty of Mathematics and Natural Sciences

The Shapley Value and the Fair Proportion Index as measures of Biodiversity - Analysis, Comparison and Computation

BACHELOR THESIS

submitted in partial fulfilment of the requirements
for the degree of Bachelor of Science (B.Sc.) in Biomathematics by

Kristina Wicke

First supervisor: Prof. Dr. Mareike Fischer
Second supervisor: Dr. Katharina Hoff

Greifswald, 9th October 2014

Abstract

Due to limited financial means, biodiversity conservation programs often need to prioritize the species to conserve. Two indices used in this matter are the *Shapley Value* and the *Fair Proportion Index*. Both are based on phylogenetic trees and aim at quantifying the importance of a taxon to overall biodiversity. While the *Shapley Value* reflects the average biodiversity contribution of a species, the *Fair Proportion Index* lacks a biological link to conservation, but is significantly easier to calculate and has been preferred in practice. Depending on the definition of the *Shapley Value*, the two indices are either identical or highly correlated, in which case the *(modified) Shapley Value* can be derived from the *Fair Proportion Index*. This allows for a less complex calculation and makes computation feasible even for large trees. Despite the strong link between the two measures, the ranking order of taxa provided by the *modified Shapley Value* and the *Fair Proportion Index* can differ for non-ultrametric trees, calling for further analysis before choosing the taxa to conserve.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ED ..... Evolutionary Distinctiveness
EDGE ..... Evolutionary Distinct and Globally Endangered
FP ..... Fair Proportion Index
PD ..... Phylogenetic Diversity
SV ..... Shapley Value
$\widetilde{\mathbf{SV}}$ ..... modified Shapley Value

1 Introduction

Due to limited resources, biological conservation has to prioritize the species to conserve. Instead of conserving as many taxa as possible, it has been argued to consider overall biodiversity and aim at minimizing its future loss. Based on phylogenetic trees, several indices have been developed in order to indicate a taxon's importance to biodiversity and thus to provide a prioritization criterion.

In this thesis we focus on the *Shapley Value* and the *Fair Proportion Index*, two distinctiveness indices recently studied in the literature (Haake et al. [6], Hartmann [7], Fuchs and Jin [5]). Our aim is to analyse these indices, study their relationship and present a way of computing them.

The *Shapley Value*, which originates in cooperative game theory, is complex to understand and calculate. Still, as it reflects the average contribution a species makes to total diversity, it is biologically highly justified to use.

The *Fair Proportion Index*, on the other hand, lacks a biological motivation, but is far more easy to understand and calculate. Therefore it has been used in practical applications, e.g. the EDGE project.

The two indices are, however, closely related. Depending on the definition of the *Shapley Value*, they are either identical (*original Shapley Value*) or strongly correlated, becoming equivalent in the limit (*modified Shapley Value*). Proof for this was given by (Fuchs and Jin [5]) and (Hartmann [7]) and will be illustrated in this work.

In case of the *modified Shapley Value* we additionally ask the question, whether it will always find the same ranking order of taxa as the *Fair Proportion Index* and how this is influenced by the shape of the phylogenetic tree.

Finally, we focus on the computation of both indices and include two programs, `shapley.pl` and `diversity_indices.pl`, which can be used to calculate the *Shapley Value* and/or the *Fair Proportion Index* for the taxa of a given phylogenetic tree.

2 Biodiversity and Conservation

2.1 Definition of Biodiversity

According to the Convention on Biological Diversity (CBD)

> ‘ “Biological diversity” means the variability among living organisms from all sources, including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; ’ (on Biological Diversity [12])

This definition includes three levels of diversity: diversity within species, between species and of ecosystems (on Biological Diversity [12]). Moreover, biodiversity can be interpreted in reference to genetic or molecular differences. Depending on the point of view, different aspects might be of interest, thus leading to different definitions of biodiversity. Whereas biodiversity in Ecology could mean species richness and abundance in certain habitats, the same term could refer to allelic diversity in Molecular Biology. Therefore

> ‘ A definition of biodiversity that is altogether simple, comprehensive, and fully operational (...) is unlikely to be found.’ (Noss [11])

Not only does the understanding of biodiversity vary, but also the methods used to measure the variability among species. An intuitive approach would be to simply describe the differences among species by measuring and comparing traits such as body size, respiration type (anaerobic vs. aerobic) or reproduction process. The general problem with this approach is its tendency to be biased as the ‘difference among species will strongly depend on the choice of traits measured’ (Vellend et al. [15]). Furthermore, not all criteria are applicable to all species, e.g. one can only compare the type of respiration among species, which are actually using respiration (and not fermentation). This motivates a more general approach based on DNA-sequence data. Phylogenetic biodiversity aims at quantifying differences among species by evaluating their evolutionary history and relationships.

2.2 Phylogenetic Approach to Conservation Biology

Phylogenetic trees, which can be reconstructed from DNA-sequence data, reflect the evolutionary relationships among its leaves, i.e. different species. The phylogenetic distance between two species (measured by the number of different nucleotides in the underlying DNA-sequence) can be interpreted as an estimate of the amount of time since their divergence, i.e. the amount of time they have evolved independently (Vellend et al. [15]). In order to quantify the evolutionary distinctness of single species or the phylogenetic diversity of a group of species within a phylogenetic tree, several metrics have been introduced. An overview of these can be found in (Vellend et al. [15]). Based on these metrics,

simple indices and algorithms have been developed in order to prioritize the taxa to conserve, given limited resources. The idea behind this is, not only to conserve the greatest number of taxa possible, but to take taxon distinctness into account and minimize the future loss of biodiversity (Vellend et al. [15], Hartmann and Steel [8]).

In the following, two of these indices, the *Shapley Value* and the *Fair Proportion Index*, will be further described and analysed.

3 The Shapley Value

3.1 The Shapley Value in Game Theory

In game theory the *Shapley Value* is an important concept for cooperative games, where a *cooperative game* consists of a set of *players* $N = \{1, 2, \ldots, n\}$ and a *characteristic function* $\nu : 2^N \to \mathbb{R}$, that assigns a value to every subset of N. The subsets of N are called *coalitions* (the subset consisting of all players is called the *grand coalition*) and the function value of ν is called the *worth* of the coalition. As some players might be more important to the coalition than others, the question arises, how the total worth of the coalition should be distributed among them.

The *Shapley Value* provides a "fair" solution to this problem. Given a cooperative game (N, ν) the *Shapley Value* for player i can be calculated as follows:

$$\varphi_i(N, \nu) = \frac{1}{n!} \sum_{\substack{S \subseteq N \\ i \in S}} (s-1)!(n-s)!(\nu(S) - \nu(S-i)) \tag{3.1}$$

where $s = |S|$ is the size of the coalition S and $n = |N|$ is the total number of players. This value can be interpreted as the *expected marginal contribution* of player i (see Haake et al. [6]).

The *Shapley Value* fulfils four axioms (see Haake et al. [6]):

1. **Efficiency**. The sum of the individual *Shapley Values* equals the worth of the *grand coalition*:
$$\sum_{i \in N} \varphi_i(N, \nu) = \nu(N)$$

2. **Symmetry.** If two players i and j play the same role in the game, meaning
$$\nu(S \cup i) = \nu(S \cup j)$$
for every subset $S \subseteq N$ neither containing i nor j, then $\varphi_i(\nu) = \varphi_j(\nu)$.

3. **Dummy Axiom.** The *Shapley Value* of a player i, who is not adding worth to any coalition it joins, i.e. $\nu(S \cup \{i\}) = \nu(S)$ for all coalitions S, is zero.

4. **Additivity.** Given to cooperative games (N, ν) and (N, ω) with the same set of players N and characteristic functions ν and ω, the *Shapley Value* for the game $(N, \nu + \omega)$ is the sum of the *Shapley Values* of the individual games:
$$\varphi_i(\nu + \omega) = \varphi_i(\nu) + \varphi_i(\omega) \text{ for all } i \in N$$

3.2 The Shapley Value in Phylogenetics

3.2.1 Phylogenetic Diversity

One of the metrics introduced to measure the diversity present in a tree (see 2.2), is called *Phylogenetic Diversity (PD)*.

Definition 1. Let T be a rooted phylogenetic tree with leaf set N. For a subset $S \subseteq N$ of taxa the PD is calculated by summing up the branch lengths of the phylogenetic tree containing S and the root (i.e. the sum of branch lengths in the smallest spanning tree containing S and the root).

Example 1. Consider the following phylogenetic tree:

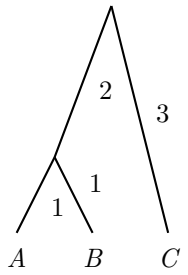

Fig. 1: PD - rooted tree

Here the *phylogenetic diversity* is given by:

$$\begin{aligned} PD(\emptyset) &= 0 \\ PD(A) &= 1 + 2 = 3 \\ PD(B) &= 1 + 2 = 3 \\ PD(C) &= 3 \\ PD(AB) &= 1 + 1 + 2 = 4 \\ PD(AC) &= 1 + 2 + 3 = 6 \\ PD(BC) &= 1 + 2 + 3 = 6 \\ PD(ABC) &= 1 + 1 + 2 + 3 = 7 \end{aligned}$$

Remark. *Definition 1* holds for rooted phylogenetic trees. In case of unrooted phylogenetic trees, the PD of a subset of taxa S is the sum of branch lengths in the minimal spanning tree connecting those leaves. The PD of single taxa is defined as zero.

Example 2. Considering the tree above as unrooted the PD changes to:

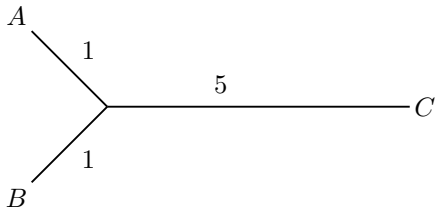

Fig. 2: PD - unrooted tree

$$PD(\emptyset) = PD(A) = PD(B) = PD(C) = 0$$
$$PD(AB) = 2$$
$$PD(AC) = 6$$
$$PD(BC) = 6$$
$$PD(ABC) = 7$$

3.2.2 Definition and Interpretation of the Shapley Value

As introduced by (Haake et al. [6]), an unrooted phylogenetic tree T can be associated with a cooperative game (N, ν_T), called the *phylogenetic tree game*. In this context N represents the set of leaves of the tree, i.e. the species and ν_T maps every subset $S \subseteq N$ of taxa to the sum of the edge weights of the spanning tree connecting the members in S.

This concept can be generalized to be applicable to both rooted and unrooted phylogenetic trees by claiming $\nu_T(S) = PD(S)$ for any subset $S \subseteq N$. Plugging this into (3.1) leads to the following definition:

Definition 2. Let T be a phylogenetic tree with leaf set N and let *PD(S)* denote the *phylogenetic diversity* of $S \subseteq N$. Then the *Shapley Value* for a taxon a is defined as

$$\phi(a) = \frac{1}{n!} \sum_{\substack{S \subseteq N \\ a \in S}} (|S| - 1)!(n - |S|)!(PD(S) - PD(S \setminus \{a\})), \tag{3.2}$$

where $n = |N|$ and S denotes a coalition of species including taxon a and the sum runs over all such coalitions possible. [1]

The *Shapley Value* of a taxon can be seen as the 'average diversity the species can be expected to add to a group that it joins' (Haake et al. [6]), i.e. the expected contribution to total *PD*. So if one species, say a, obtains a higher *Shapley Value* than another species, b, i.e. $\phi(a) > \phi(b)$, then species a may contribute more diversity to a group of species than b (Haake et al. [6]). Therefore the *Shapley Value* provides a sensible ranking criterion to be used in conservation prioritization.

[1] A "biological" derivation of this formula, independent from game theory, can be found in (Hartmann [7][2.2]).

Besides this general interpretation of the value itself, it is also possible to motivate the Shapley axioms in a biological context (see 3.1).

Efficiency simply implies that the total diversity in the tree will be distributed among its leaves. Thus the *Shapley Value* of a taxon describes its contribution to total PD and therefore answers the question of how important this species is for diversity (Haake et al. [6]).

Symmetry means that two species playing the same role in a tree and thus contributing equally to total diversity, should also be ranked equally, what seems like a plausible feature.

The *Dummy Axiom* does not have a meaning in case of phylogenetic trees as every species i adds at least the worth of its pending edge to a coalition it joins, since we can assume that edge weights are non-negative and non-zero. The former holds, because edge weights either represent the passage of time or some kind of evolutionary distance between species, both of which are non-negative. If edge weights were allowed to be zero, the two species on either side of these edges would be the same, which is not possible. Therefore every species must add to the PD of a coalition it joins, meaning that there are no dummy species. For the hypothetical case of species not adding diversity to any coalition, this axiom is still reasonable.

For *Additivity* suppose you are given a set of nucleotide sequences of length 200 for a set of species N. The observed distance (also p-Distance) between two species i and j is then given by the relative frequency of nucleotide differences. This distance can be used to reconstruct a phylogenetic tree (note that in practice it is better to use corrected distances). Assuming the sequence is split by half, both the first and the last 100 positions can be used to construct trees and thus tree games (N, ν_1) and (N, ν_2), respectively. Then the *Shapley Value* of the game $(N, \nu_1 + \nu_2)$ is the sum of the individual games. This seems appropriate, since if the observed distances in both halves of the sequence 'arise from a tree metrics on the same topological tree, then the sum game will arise from the tree reconstructed'(Haake et al. [6]) from the whole sequence. This scenario serves an illustration of the additivity of the *Shapley Value* in case both sequence parts originate in the same tree topology. It is, however, possible that reconstructing trees from both halves of the sequences leads to two different topologies. In this case the *Additivity Axiom* lacks the interpretation indicated above (example taken from (Haake et al. [6])).

3.2.3 Two versions of the Shapley Value

The *Shapley Value* is calculated in different ways in the literature, depending on whether the singleton set $S = \{a\}$ is included in (3.2) or not. To avoid confusion let the following

notation be introduced:

$$\phi(a) = \begin{cases} \widetilde{SV}(a) \text{ if claimed } |S| \geq 2 \\ SV(a) \text{ else} \end{cases}$$

and refer to $\widetilde{SV}(a)$ as the *modified Shapley Value* in contrast to the ordinary *Shapley Value* $SV(a)$.

Remark. $SV(a)$ can easily be calculated from $\widetilde{SV}(a)$ and the other way round:

$$SV(a) = \widetilde{SV}(a) + \frac{PD(a)}{n} \tag{3.3}$$

Proof. Consider the contribution of the singleton set $\{a\}$ to $SV(a)$:
By definition this is

$$\begin{aligned} \frac{1}{n!}\Big((1-1)!(n-1)!(PD(a) - PD(\emptyset))\Big) &= \frac{(n-1)!}{n!} PD(a) \\ &= \frac{PD(a)}{n}. \end{aligned}$$

As $\widetilde{SV}(a)$ "lacks" this contribution it is

$$\widetilde{SV}(a) = SV(a) - \frac{PD(a)}{n}$$

□

Remark. (Fuchs and Jin [5]) provide a slightly different formula for calculating $SV(a)$ from $\widetilde{SV}(a)$. They argue that it can be calculated by using the depth of a in T, where the depth is defined as the number of edges on the path from a to the root:

$$SV(a) = \widetilde{SV}(a) + \frac{depth(a)}{n}$$

This only holds if the sum of all edge lengths on the path from a to the root is the same as the number of edges on this path, because then and only then $PD(a) \equiv depth(a)$. One may consider this a special case, but an easy example would be a tree, where all edges are defined to be of length one. Nonetheless this formula will not be used in the following.

Note that the *modified Shapley Value* does not fulfil the *Efficiency Axiom* (3.1) as

$$\begin{aligned}\sum_{i=1}^{n}\widetilde{SV}(i) &= \sum_{i=1}^{n}\Big(SV(i)-\frac{PD(i)}{n}\Big)\\ &= \sum_{i=1}^{n}SV(i)-\sum_{i=1}^{n}\frac{PD(i)}{n}\\ &= PD(N)-\sum_{i=1}^{n}\frac{PD(i)}{n}\\ &\neq PD(N).\end{aligned}$$

The impact of the different versions of the *Shapley Value* will be discussed in more detail later, when compared with the *Fair Proportion Index*.

3.2.4 Example: Five-leaf-tree

Example 3. The tree in Fig.3 is used in (Hartmann [7]) to demonstrate the calculation of *PD* and the *Fair Proportion Index*. Here it is also used as an example for calculating the *Shapley Value*.
The *Shapley Value* for taxon b is calculated as follows:

$$\begin{aligned}SV(b) &= \frac{1}{5!}\sum_{S,b\in S}(|S|-1)!(5-|S|)!(PD(S)-PD(S\setminus\{b\}))\\ &= \frac{1}{5!}\Big[(1-1)!(5-1)!(4-0)\\ &\quad+(2-1)!(5-2)!\big((6-4)+(8-4)+(8-4)+(8-4)\big)\\ &\quad+(3-1)!(5-3)!\big((10-8)+(10-8)+(10-8)\\ &\qquad+(9-5)+(10-6)+(10-6)\big)\\ &\quad+(4-1)!(5-4)!\big((11-7)+(12-10)+(12-10)+(11-9)\big)\\ &\quad+(5-1)!(5-5)!(13-11)\Big]\\ &= \frac{1}{120}\big[96+84+72+60+48\big]\\ &= \frac{360}{120}\\ &= 3\end{aligned}$$

Analogously one calculates:
$SV(c)=3\ ,\quad SV(d)=2\frac{1}{6}\ ,\quad SV(e)=2\frac{1}{6}\ ,\quad SV(f)=2\frac{2}{3}$

For the *modified Shapley Value* we have:
$\widetilde{SV}(b)=2\frac{1}{5}\ ,\quad \widetilde{SV}(c)=2\frac{1}{5}\ ,\quad \widetilde{SV}(d)=1\frac{11}{30}\ ,\quad \widetilde{SV}(e)=1\frac{11}{30}\ ,\quad \widetilde{SV}(f)=1\frac{13}{15}$

This small example shows that the calculation of the *Shapley Value* takes some effort as 2^{n-1} (or $2^{n-1} - 1$ for the modified *Shapley Value*) subsets $S \subseteq N$ containing the taxon in question have to be considered.

This motivates the introduction of the *Fair Proportion Index*, which is significantly easier to calculate and yet a suitable prioritization criterion.

4 The Fair Proportion Index

4.1 Definition and Example

The idea of the *Fair Proportion Index (FP)* is to divide the phylogenetic diversity of a rooted tree among its leaves. For each edge the edge length is distributed equally among the taxa descending from that edge or in other words 'each taxon descendant from an edge is allocated an equal proportion of that edge length' (Hartmann [7]).

Definition 3. Let λ_e denote the edge weight of edge e. Then the *Fair Proportion Index* for a taxon a can be calculated as follows:

$$FP(a) = \sum_e \frac{\lambda_e}{D_e}, \tag{4.1}$$

where the sum runs over all edges e on the path from a to the root and D_e denotes the number of leaves descendent from that edge.

This method may seem a sensible way to divide the evolutionary history of a tree, but in contrast to the *Shapley Value* it lacks a biological interpretation (Hartmann [7]). However, the great advantage over the *Shapley Value* is its easy calculation. And as will be discussed later, the *Fair Proportion Index* provides a measure equal or strongly related to the *Shapley Value* (depending on which version of the *Shapley Value* is used), which justifies its use in prioritization.

Example 4. Calculation of the *Fair Proportion Indices* for "Hartmann's Example Tree" (see Fig.3) yields:

$$FP(b) = FP(c) = \frac{2}{1} + \frac{2}{2} = 3$$
$$FP(d) = FP(e) = \frac{1}{1} + \frac{1}{2} + \frac{2}{3} = 2\tfrac{1}{6}$$
$$FP(f) = \frac{2}{1} + \frac{2}{3} = 2\tfrac{2}{3}$$

Note that the sum of these values $2 \cdot 3 + 2 \cdot 2\frac{1}{6} + 2\frac{2}{3} = 13$ equals the sum of edge lengths in the tree, i.e. the total PD.
More generally, the *Fair Proportion Index* fulfils the *Efficiency Axiom* introduced in (3.1), since each edge weight is distributed equally among the descending taxa. Thus, when summing over the FP for all taxa, the terms $1/D_e$ cancel out and we obtain the sum of all edge weights or in other words the total PD.

The attentive reader may also have observed the equivalence of the FP-indices to the SV-values calculated above (see 3.2.4). Table 1 summarises these results.

taxon	FP	SV	$\widetilde{SV}$
b	3	3	$2\frac{1}{5}$
c	3	3	$2\frac{1}{5}$
d	$2\frac{1}{6}$	$2\frac{1}{6}$	$1\frac{11}{30}$
e	$2\frac{1}{6}$	$2\frac{1}{6}$	$1\frac{11}{30}$
f	$2\frac{2}{3}$	$2\frac{2}{3}$	$1\frac{13}{15}$

Table 1: Summary Example 5-leaf tree

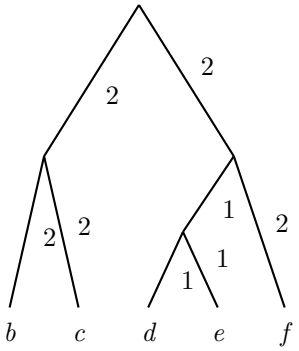

Fig. 3: Hartmann's Example Tree

4.2 The EDGE Project

An eminent example for the utilisation of the *Fair Proportion Index* is the "EDGE of Existence" Project, established by the *Zoological Society of London (ZLS)* in 2007. This programme aims at conserving species, 'that are both evolutionary distinct and globally endangered' (Isaac et al. [9]) and therefore called "EDGE-species". To identify those species, "EDGE"-scores, consisting of a value for both the evolutionary distinctiveness and the extinction risk are calculated as

$$\begin{aligned} EDGE &= \ln(1 + ED) + GE \cdot \ln(2) \\ &= \ln(1 + FP) + GE \cdot \ln(2), \end{aligned}$$

where the ED-score (*Evolutionary Distinctiveness*) corresponds to the *Fair Proportion Index* and 'GE is the Red List category weight [Least Concern = 0, Near Threatened and Conservation Dependent = 1, Vulnerable = 2, Endangered = 3, Critically Endangered = 4]' (Isaac et al. [9]).

This approach, which combines the species' importance to phylogenetic diversity (as indicated by the *Fair Proportion Index*) with their risk of extinction, has so far produced priority lists for amphibians, mammals, corals and birds, which can be downloaded from `http://www.edgeofexistence.org/about/edge_science.php` .

In case of the mammalians, (Isaac et al. [9]) state, that 'the 100 highest-ranking species represent a high proportion of total mammalian diversity and include many species not usually recognised as conservation priorities'.

This shows, that although at first sight the *Fair Proportion Index* lacks a biological motivation or interpretation, it is an important tool to use. Additionally, given the strong link between it and the *Shapley Value*, which will be discussed in the following, it can be argued, that the biological interpretation of the latter also holds for the *Fair Proportion Index*.

5 Relationship between the Fair Proportion Index and the Shapley Value

As mentioned in (3.2.3), there are different versions of the *Shapley Value* in the literature, depending on whether the singleton set $\{a\}$ is included in its calculation or not.
The *Shapley Value* was first introduced for unrooted phylogenetic trees by (Haake et al. [6]). For unrooted trees, however, the *modified Shapley Value* and the ordinary *Shapley Value* are the same as the contribution of the single set $\{a\}$ to $\phi(a)$ equals zero

$$\begin{aligned}\frac{1}{n!}\Big((1-1)!(n-1)!(PD(a)-PD(\emptyset))\Big) &= \frac{1}{n!}\Big(0!\cdot(n-1)!\cdot(0-0)\Big)\\ &= 0\end{aligned}$$

and it therefore doesn't make a difference if one claims $|S| \geq 2$ or not. All the same, at some stage (Haake et al. [6][p.490]) states $|S| \geq 2$.
This might be the reason, why (Hartmann [7]), who further analysed the *Shapley Value* for rooted phylogenetic trees, also claims $|S| \geq 2$ in a proof. As discussed before, this does make a difference here, thus leading to the *modified Shapley Value* $\widetilde{SV}$.
The definition of the *Shapley Value* in (Hartmann [7]), however, does not make it clear if this modification was intended.

When examining the relationship between the *Fair Proportion Index*, which is only defined for rooted trees, and the *Shapley Value*, one has to differentiate between the two versions of the latter. In the following we will first compare the *Fair Proportion Index* to the *original* and afterwards to the *modified Shapley Value*.

5.1 Equivalence of the original Shapley Value and the Fair Proportion Index

As Table (1) suggests the *original Shapley Value* and the *Fair Proportion Index* for a rooted phylogenetic tree T are equivalent. Mathematical proof for this equivalence is given by (Fuchs and Jin [5]). We have

Theorem 1 (2013, Fuchs and Yin)**.** *The Fair Proportion Index and the Shapley Value are identical, i.e.*

$$FP_T(a) = SV_T(a).$$

Proof. In order to prove this statement, (Fuchs and Jin [5]) show that the *Fair Proportion Index* and the *Shapely Value* can be computed by the same recursion. On this purpose let T_l and T_r denote the left and the right subtree of the root of T and assume their size

to be j and $n-j$, respectively. Furthermore, assume that a is in T_l and denote the left edge incident to the root by e.

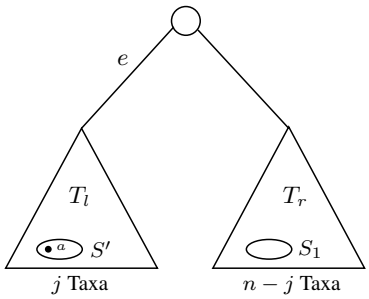

Fig. 4: Rooted phylogenetic tree with subtrees T_l and T_r

1. The *Fair Proportion Index* can be calculated recursively by

$$FP_T(a) = \frac{\lambda_e}{j} + FP_{T_l}(a), \tag{5.1}$$

which is directly implied by its definition.

2. Likewise, for the *Shapley Value* we have

$$SV_T(a) = \frac{\lambda_e}{j} + SV_{T_l}(a). \tag{5.2}$$

To see this, fix a set S' of taxa of T_l with $a \in S'$.
First note that the PD of coalitions containing a and a set of taxa S_1 of the right subtree T_r can be expressed as a function of S', that is

$$PD_T(S') - PD_T(S' \setminus \{a\}) = PD_T(S' \cup S_1) - PD_T((S' \cup S_1) \setminus \{a\}),$$

which holds because of the linearity of PD:

$$\begin{aligned} &PD_T(S' \cup S_1) - PD_T((S' \cup S_1) \setminus \{a\}) \\ &\quad = PD_T(S') + PD_T(S_1) - PD_T(S' \setminus \{a\}) - \underbrace{PD_T(S_1 \setminus \{a\})}_{PD_T(S_1)} \\ &\quad = PD_T(S') - PD_T(S' \setminus \{a\}) \end{aligned}$$

Remembering the definition of the *Shapley Value*

$$SV_T(a) = \frac{1}{n!} \sum_{S, a \in S} (|S|-1)!(n-|S|)!(PD_T(S) - PD_T(S \setminus \{a\}),$$

this means that only terms with $S = S'$ appear in the sum.
As there are $\binom{n-j}{l}$ possibilities to choose l Taxa of T_r with $l = 0, 1, \ldots, n-j$, the

coefficient in front of $PD_T(S') - PD_T(S' \setminus \{a\})$ becomes

$$\begin{aligned}
&\frac{1}{n!}\sum_{l=0}^{n-j}\binom{n-j}{l}(((|S'|+l)-1)!(n-(|S'|+l))!\\
&=\frac{1}{n!}\sum_{l=0}^{n-j}\binom{n-j}{l}(|S'|+l-1)!(n-|S'|-l)!\\
&=\frac{1}{n!}\sum_{l=0}^{n-j}\frac{(n-j)!}{l!(n-j-l)!}(|S'|+l-1)!(n-|S'|-l)!\\
&=\frac{(n-j)!}{n!}\sum_{l=0}^{n-j}\frac{(|S'|+l-1)!}{l!}\frac{(n-|S'|-l)!}{(n-j-l)!}\\
&=\frac{(n-j)!}{n!}(|S'|-1)!(j-|S'|)!\sum_{l=0}^{n-j}\frac{(|S'|+l-1)!}{l!(|S'|-1)!}\frac{(n-|S'|-l)!}{(n-j-l)!(j-|S'|)!}\\
&=\frac{(n-j)!}{n!}(|S'|-1)!(j-|S'|)!\sum_{l=0}^{n-j}\binom{|S'|-1+l}{l}\binom{n-|S'|-l}{n-j-l}\\
&\overset{(*)}{=}\frac{(n-j)!}{n!}(|S'|-1)!(j-|S'|)!\binom{n}{j}\\
&=\frac{(n-j)!}{n!}(|S'|-1)!(j-|S'|)!\frac{n!}{j!(n-j)!}\\
&=\frac{(|S'|-1)!(j-|S'|)!}{j!}
\end{aligned}$$

Thus, we have

$$SV_T(a) = \frac{1}{j!}\sum_{S',a\in S'}(|S'|-1)!(j-|S'|)!(PD_T(S') - PD_T(S' \setminus \{a\})), \tag{5.3}$$

where the sum runs over all taxa from T_l with $a \in S'$. Now consider two cases:

- For $S' \neq \{a\}$ we have

$$\underbrace{PD_T(S')}_{\text{containing edge } e} - \underbrace{PD_T(S' \setminus \{a\})}_{\text{containing edge } e} = PD_{T_l}(S') - PD_{T_l}(S' \setminus \{a\})$$

  as the edge weight λ_e of e (see Fig. 4) is counted in both terms on the left-hand side of the equation and therefore adjusts to zero.

- For $S' = \{a\}$ we have $PD_T(\{a\}) = \lambda_e + PD_{T_l}(\{a\})$.

Plugging this into (5.3) yields

$$\begin{aligned} SV_T(a) &= \frac{\lambda_e}{j} + \frac{1}{j!} \sum_{S', a \in S'} (|S'| - 1)!(j - |S'|)!(PD_{T_l}(S') - PD_{T_l}(S' \setminus \{a\})) \\ &= \frac{\lambda_e}{j} + SV_{T_l}(a). \end{aligned}$$

Consequently, the *Fair Proportion Index* and the *Shapley Value* follow the same recursion. To make the proof complete, consider the start of the recursion T_0, where T_0 is a tree of size two.
We have $FP_{T_0}(a) = \frac{\lambda_e}{1} = \lambda_e = SV_{T_0}(a)$.
Thus, the *Fair Proportion Index* and the *Shapley Value* are the same. □

Remark. The prove the identity

$$\sum_{l=0}^{n-j} \binom{|S'| - 1 + l}{l} \binom{n - |S'| - l}{n - j - l} = \binom{n}{j} \tag{5.4}$$

used in $(*)$, first recall the following generating functions:

$$\sum_{n=0}^{\infty} \binom{n + c - 1}{c - 1} x^n = \frac{1}{(1 - x)^c} \tag{5.5}$$

$$\sum_{n=0}^{\infty} \binom{n + c}{c} x^n = \frac{1}{(1 - x)^{c+1}} \tag{5.6}$$

Let $N = n - j$ and $s = |S'|$. Then the generating function of the sum in (5.4) is

$$\begin{aligned} &\sum_{N=0}^{\infty} \sum_{l=0}^{N} \binom{s - 1 + l}{l} \binom{N + j - s - l}{n - j - l} x^N \\ &= \sum_{N=0}^{\infty} \sum_{l=0}^{N} \binom{s - 1 + l}{s - 1} \binom{N + j - s - l}{j - s} x^N \\ &= \sum_{l=0}^{\infty} \binom{s - 1 + l}{s - 1} x^l \sum_{N=0}^{\infty} \binom{N + j - s - l}{j - s} x^N \\ &\overset{(5.5)}{=}_{(5.6)} \frac{1}{(1 - x)^s} \frac{1}{(1 - x)^{j - s + 1}} \\ &= \frac{1}{(1 - x)^{j+1}} \end{aligned}$$

Thus, we have

$$\sum_{N=0}^{\infty}\sum_{l=0}^{N}\binom{s-1+l}{l}\binom{N+j-s-l}{n-j-l}x^N = \frac{1}{(1-x)^{j+1}} \underset{(5.6)}{=} \sum_{N=0}^{\infty}\binom{N+j}{j}x^N$$

Coefficient comparison yields

$$\sum_{l=0}^{N}\binom{s-1+l}{l}\binom{N+j-s-l}{n-j-l} = \binom{N+j}{j}$$

and by replacing s with $|S'|$ and N with $n-j$, we finally have

$$\sum_{l=0}^{n-j}\binom{|S'|-1+l}{l}\binom{n-|S'|-l}{n-j-l} = \binom{n}{j}.$$

5.2 Equivalence in the limit for the modified Shapley Value and the Fair Proportion Index

As we have shown $SV = FP$, we can use (3.2.3) to express the relationship between the *Fair Proportion Index* and the *modified Shapley Value*. This yields

$$FP(a) = \widetilde{SV}(a) + \frac{PD(a)}{n}. \tag{5.7}$$

One can already guess that these two indices will become equivalent in the limit $n \to \infty$, since we have $\frac{PD(a)}{n} \underset{n\to\infty}{\longrightarrow} 0$.

(Hartmann [7]) chose a different approach to prove this. He considered the contribution an edge e makes to the *modified Shapley Value* and the *Fair Proportion Index*, respectively. In the limit this contribution becomes equivalent for both indices.
Let $\theta_{FP}(a,e)$ and $\theta_{\widetilde{SV}}(a,e)$ denote the contribution of edge e to the *Fair Proportion Index* and *modified Shapley Value* of taxon a, respectively. Then we have

Theorem 2 (Hartmann, 2012)**.** *In the limit $n \to \infty$ we have $\theta_{FP}(a,e) = \theta_{\widetilde{SV}}(a,e)$. In other words the contribution of edge e to the Shapley Value and FP value of taxon a becomes equivalent.*

Proof. Let λ_e be the edge weight of edge e and D_e the set of descendants. Furthermore let J denote the set of taxa separated from a by the edge e, i.e. $J = N - D_e$ (see Fig. 5)

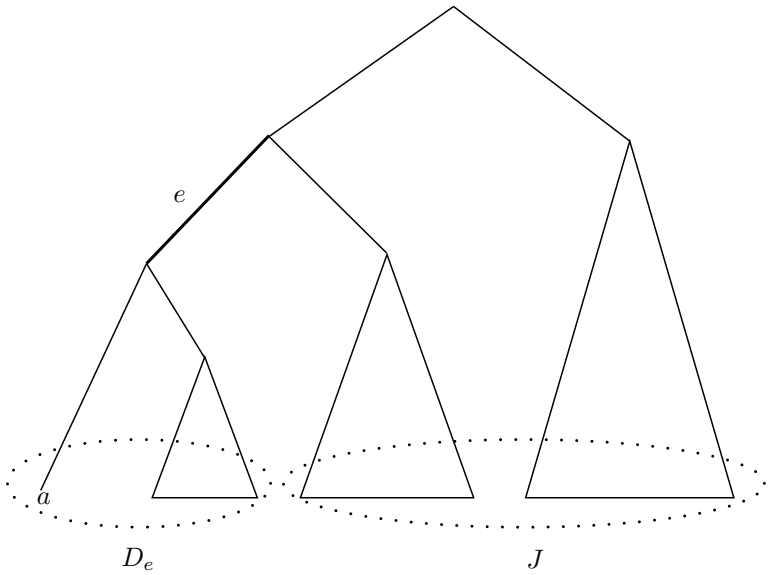

Fig. 5: Contribution of edge e to FP and $\widetilde{SV}$

1. For the *Fair Proportion Index* the contribution of edge e is

$$\theta_{FP}(a,e) = \begin{cases} \frac{\lambda_e}{|D_e|} & \text{for } a \in D_e \\ 0 & \text{for } a \notin D_e \end{cases} \tag{5.8}$$

2. For the *modified Shapley Value* of a taxon a note that an edge e only contributes to it, if all taxa in S (i.e. all taxa a forms a coalition with) are separated from a by this edge. That is, if $S - J = \{a\}$ or in other words $S = J \cup \{a\}$. If that is not the case, e.g. if a forms a coalition with taxa from D_e, the edge e will not add extra worth to the coalition and therefore does not contribute to $\widetilde{SV}(a)$.
   As we are now considering the *modified Shapley Value*, we also have to claim $|S| \geq 2$. Thus, we have $2 \leq |S| \leq |J| + 1$. For a fixed subset size $|S|$ there are $\binom{|J|}{|S|-1}$ sets, which fulfil this condition.
   Recall the definition of the *modified Shapley Value*:

$$\widetilde{SV}(a) = \frac{1}{n!} \sum_{\substack{|S|\geq 2, \\ a \in S}} (|S|-1)!(n-|S|)!(PD(S) - PD(S \setminus \{a\}))$$

   The total contribution the edge e makes to this value, is its coefficient times the

number of coalitions a adds worth to. Thus, we have

$$\begin{aligned}\theta_{\widetilde{SV}}(a,e) &= \frac{1}{n!}\sum_{|S|=2}^{|J|+1}(|S|-1)!(n-|S|)!\lambda_e \times \binom{|J|}{|S|-1} \\ &= \lambda_e \sum_{|S|=2}^{|J|+1}\frac{(|S|-1)!(n-|S|)!}{n!}\frac{|J|!}{(|S|-1)!(|J|-|S|+1)!} \\ &= \lambda_e \sum_{|S|=2}^{|J|+1}\frac{|J|!(n-|S|)!}{(|J|-|S|+1)!n!}. \end{aligned} \tag{5.9}$$

As this term is independent of a, $\theta_{\widetilde{SV}}(a,e)$ is the same for all taxa on the same side of edge e.

Based on this, (Hartmann [7]) now refers to (Haake et al. [6]) and states that 'each edge length is shared out in its entirety amongst the taxa', meaning that 'λ_e will be divided amongst the taxa in D_e'.

This, together with the fact that $\theta_{\widetilde{SV}}(a,e)$ is the same for all taxa in D_e, yields

$$\theta_{\widetilde{SV}}(a,e) = \frac{\lambda_e}{|D_e|} \text{ for } a \in D_e. \tag{5.10}$$

Remark. This reasoning seems plausible, but actually calculating the sum in (5.9) makes it somehow less clear. While extending the sum to include $|S| = 1$ (in other words: using the *original Shapley Value*) yields equality right away

$$\begin{aligned}\lambda_e \sum_{|S|=1}^{|J|+1}\frac{|J|!(n-|S|)!}{(|J|-|S|+1)!n!} &= \lambda_e \frac{|J|!}{n!}\sum_{|S|=1}^{|J|+1}\frac{(n-|S|)!}{(|J|-|S|+1)!} \\ &= \lambda_e \frac{|J|!}{n!}\sum_{|S|=1}^{|J|+1}\binom{n-|S|}{n-|J|-1}(n-|J|-1)! \\ &= \lambda_e \frac{|J|!}{n!}(n-|J|-1)!\sum_{|S|=1}^{|J|+1}\binom{n-|S|}{n-|J|-1} \\ &= \lambda_e \frac{|J|!}{n!}(n-|J|-1)!\binom{n}{n-|J|} \\ &= \lambda_e \frac{|J|!}{n!}(n-|J|-1)!\frac{n!}{(n-|J|)!|J|!} \\ &= \lambda_e \frac{1}{n-|J|} \\ &= \frac{\lambda_e}{|D_e|}, \end{aligned}$$

calculation of the sum as defined in (5.9) leads to

$$\begin{aligned}\lambda_e \sum_{|S|=2}^{|J|+1} \frac{|J|!(n-|S|)!}{(|J|-|S|+1)!n!} &= \lambda_e\Big(\sum_{|S|=1}^{|J|+1} \frac{|J|!(n-|S|)!}{(|J|-|S|+1)!n!} - \frac{|J|!(n-1)!}{(|J|-1+1)!n!}\Big) \\ &= \lambda_e\Big(\frac{1}{n-|J|} - \frac{1}{n}\Big) \\ &= \lambda_e \frac{|J|}{n(n-|J|)}.\end{aligned}$$

However, (Hartmann [7]) now considers the case $a \notin D_e$ or in other words $a \in J$. The contribution of edge e to $\widetilde{SV}$ of a taxon in J can be calculated by substituting $n-|J|$ for $|J|$ in (5.9):

$$\theta_{\widetilde{SV}}(a,e) = \lambda_e \sum_{|S|=2}^{n-|J|+1} \frac{(n-|J|)!(n-|S|)!}{(n-|J|-|S|+1)!n!} \text{ with } a \notin D_e$$

Considering all taxa in J, the contribution of e can be calculated by making this substitution and summing the result over all taxa in J:

$$\begin{aligned}\sum_{a\in J} \theta_{\widetilde{SV}}(a,e) &= \lambda_e |J| \sum_{|S|=2}^{n-|J|+1} \frac{(n-|J|)!(n-|S|)!}{(n-|J|-|S|+1)!n!} \\ &\overset{n-|J|=|D_e|}{=} \lambda_e \sum_{|S|=2}^{|D_e|+1} \frac{|J||D_e|!(n-|S|)!}{(|D_e|-|S|+1)!n!}\end{aligned}$$

Finally, taking the limit, we get:

$$\begin{aligned}\lim_{n\to\infty} \sum_{a\in J} \theta_{\widetilde{SV}}(a,e) &= \lambda_e \lim_{n\to\infty} \sum_{|S|=2}^{|D_e|+1} \frac{|J||D_e|!(n-|S|)!}{(|D_e|-|S|+1)!n!} \\ &= 0 \qquad (5.11)\end{aligned}$$

As all terms in the sum are positive, the sum can only converge to zero, if the summands themselves converge to zero. Therefore, summarizing (5.10) and (5.11), we obtain

$$\lim_{n\to\infty} \theta_{\widetilde{SV}}(a,e) = \begin{cases} \frac{\lambda_e}{|D_e|} & \text{for } a \in D_e \\ 0 & \text{for } a \notin D_e \end{cases} \qquad (5.12)$$

This shows that in the limit the contribution of edge e to the *modified Shapley Value* equals its contribution to the *Fair Proportion Index* (compare (5.8)), which we wanted to prove. □

6 Influence of the Tree Shape

As shown above, the *Fair Proportion Index* and the *modified Shapley Value* become equivalent in the limit $n \to \infty$. The following section will deal with smaller trees and try to explore, whether the two measures always result in the same ranking order and how this is influenced by the shape of the tree.

6.1 Ultrametric Trees

If the branch lengths of a phylogenetic tree represent the passage of time and the taxa are all present-day species, the tree has to be *ultrametric*. This means that all paths from the root ρ to a leaf are of the same length, because ρ lived at a certain time and since then a certain amount of time has passed for all present-day species.
Recall that the modified *Shapley Value* of a leaf a differs from the *Fair Proportion Index* by the term $PD(a)/n$ (see 5.7). In case of an ultrametric tree, this term is the same for all leaves, because $PD(a)$ reflects the length of the path from a to the root.
Thus, the *Fair Proportion Index* and the *modified Shapley Value* will always result in the same ranking order of taxa, although the numerical values for the individual taxa differ. As an example reconsider the tree used by (Hartmann [7]) and see Table 1 and Fig. 3.

6.2 Non-ultrametric Trees

Branch lengths in a phylogenetic tree can also be interpreted as representations of evolutionary change instead of passage of time. In this case, the tree is not necessarily ultrametric, since the rate of evolutionary change can vary among different branches.

Under these circumstances it is possible to construct trees, where the ranking order provided by the *Fair Proportion Index* and the *modified Shapley Value* differ. The tree in Fig. 6 even causes a situation, where FP "fails" as a ranking criterion as it assigns the same value to every leaf, whereas $\widetilde{SV}$ does find a ranking.

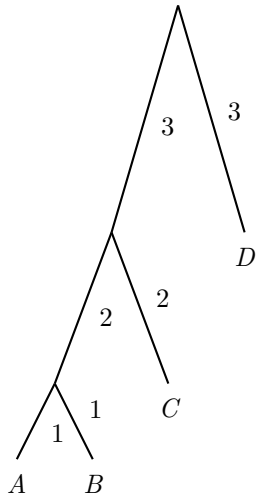

Fig. 6: Non-ultrametric imbalanced tree (i)

taxon	FP	$\widetilde{SV}$
A	3	1.5
B	3	1.5
C	3	1.75
D	3	2.25

Table 2: FP and $\widetilde{SV}$ for a non-ultrametric imbalanced tree (i)

Note that this tree is highly *imbalanced*, which means that the leaves are not separated from the root by the same number of nodes, e.g. the lineage from the root to A or B bifurcates two times, to C it bifurcates one time and D descends directly from the root.

However, the *Fair Proportion Index* and the *modified Shapley Value* do not necessarily result in a different ranking order for highly *imbalanced* non-ultrametric trees:

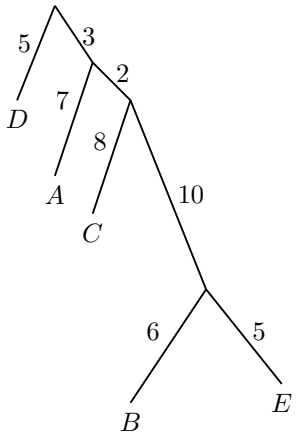

Fig. 7: Non-ultrametric imbalanced tree (ii)

taxon	FP	$\widetilde{SV}$
A	$7\frac{3}{4}$	$5\frac{3}{4}$
B	$12\frac{5}{12}$	$8\frac{13}{60}$
C	$9\frac{5}{12}$	$6\frac{49}{60}$
D	5	4
E	$11\frac{5}{12}$	$7\frac{5}{12}$

Table 3: FP and $\widetilde{SV}$ for a non-ultrametric imbalanced tree (ii)

When looking at more *balanced* trees, i.e. trees, in which most lineages bifurcate the same number of times, the ranking order of the two measures can still differ. As an example consider the two 5-leaf trees in Fig. 8 and Fig.9.
Both trees have the same topology. Still, in the first case the ranking orders between the *Fair Proportion Index* and the *modified Shapley Value* differ (FP regards A as more important than D, $\widetilde{SV}$ does it the other way round), whereas the two measures lead to the same ranking in the second tree. These examples suggest that it is hard to predict, whether the two measures will find the same ranking order for non-ultrametric trees or not.

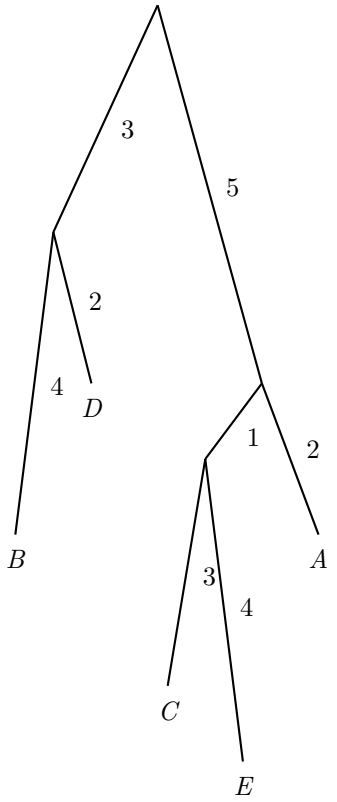

Fig. 8: Non-ultrametric balanced tree (i)

taxon	FP	$\widetilde{SV}$
A	$3\frac{2}{3}$	$2\frac{4}{15}$
B	$5\frac{1}{2}$	$4\frac{1}{10}$
C	$5\frac{1}{6}$	$3\frac{11}{30}$
D	$3\frac{1}{2}$	$2\frac{1}{2}$
E	$6\frac{1}{6}$	$4\frac{1}{6}$

Table 4: FP and $\widetilde{SV}$ for a non-ultrametric balanced tree (i)

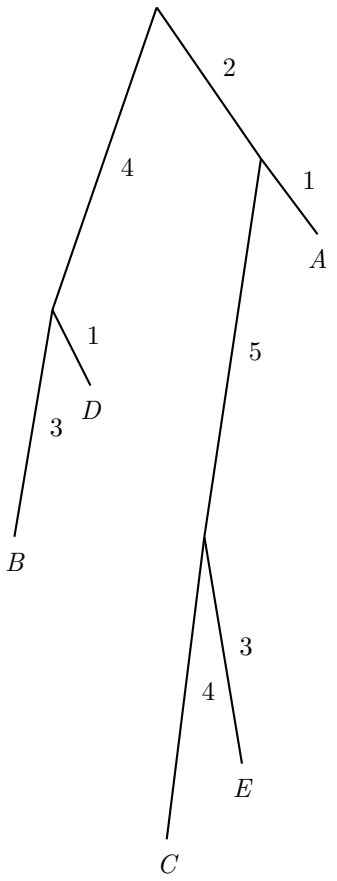

Fig. 9: Non-ultrametric balanced tree (ii)

taxon	FP	$\widetilde{SV}$
A	$1\frac{2}{3}$	$1\frac{1}{15}$
B	5	$3\frac{3}{5}$
C	$7\frac{1}{6}$	$4\frac{29}{30}$
D	3	2
E	$6\frac{1}{6}$	$4\frac{1}{6}$

Table 5: FP and $\widetilde{SV}$ for a non-ultrametric balanced tree (ii)

In order to analyse this question further, random trees of different sizes were sampled, using the programming language R (R Development Core Team [13]). In a first analysis the branch lengths were chosen to be random integers $\in [0, 10]$, in a second one integers $\in [0, 100]$. For each tree size 100 trees were generated, respectively. Then the *Fair Proportion Index* and the *modified Shapley Value* were computed for all trees and their ranking order was compared. Furthermore, the mean of the correlation coefficient r between the two indices was calculated for every sample of size 100. The results can be found in Table 6 and Table 7.

Both for branch lengths in $[0, 10]$ and in $[0, 100]$ the outcomes are roughly the same, which suggests that the degree of "non-ultrametricity" (see Table 8) does not influence the relationship between the two indices. In both cases a high correlation between the *Fair Proportion Index* and the *modified Shapley Value* can be observed, converging to one with growing tree size. This observation coincides with the strong correlation between the two indices for Yule trees described in (Hartmann [7]). More surprising is the fact that at the same time the number of cases, in which both measures find the same ranking order of taxa, decreases. So even though the indices grow more alike with increasing tree size, they tend to rank taxa differently. So far, we could not find an explanation for this phenomenon.

Number of taxa	Same ranking order	Different ranking order	mean correlation coefficient
4	66	34	0.947022
5	58	42	0.960145
6	47	53	0.975664
7	41	59	0.981419
8	33	67	0.982440
10	19	81	0.984582
15	4	96	0.990727
20	1	99	0.994788
100	0	100	0.999488
200	0	100	0.999841

Table 6: Comparison of FP and $\widetilde{SV}$ for random non-ultrametric trees with branch lengths in $[0, 10]$

Number of taxa	Same ranking order	Different ranking order	mean correlation coefficient
4	68	32	0.948596
5	59	41	0.975446
6	44	56	0.978429
7	40	60	0.981647
8	20	80	0.981752
10	17	83	0.989007
15	6	94	0.993010
20	3	97	0.994094
100	0	100	0.999513
200	0	100	0.999832

Table 7: Comparison of FP and $\widetilde{SV}$ for random non-ultrametric trees with branch lengths in $[0, 100]$

Number of taxa	Mean variance in leaf PD for branch lengths in $[0, 10]$	Mean variance in leaf PD for branch lengths in $[0, 100]$
5	36.43	3916.21
10	66.68	6031.95
20	107.42	9951.78
100	218.78	19812.12

Table 8: Mean variance in leaf PD as a measure for the degree of non-ultrametricity

7 Computation of the Fair Proportion Index and the Shapley Value

After discussing the properties of the *Fair Proportion Index* and the *Shapley Value* as well as their relationship, the following section will present a way of computing the indices using the programming language Perl. Before going into the details of the implementation, we introduce some general aspects concerning the computation.

7.1 Representing trees - The Newick Format

One way of representing trees in a computer-readable format, is the so called *Newick tree format*. It represents trees with or without edge lengths using parentheses and commas. To make the idea precise, consider the following example and its Newick representation:

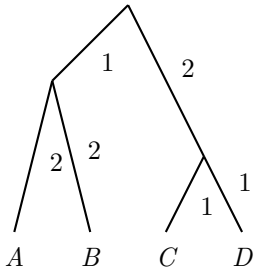

Fig. 10: Newick rooted: $((A:2, B:2):1, (C:1, D:1):2);$

This tree contains three interior nodes, which are represented by three pairs of matched parentheses. Inside the parentheses, separated by commas, are the nodes that are direct descendants of the interior nodes, respectively. E.g. A and B are direct descendants of the interior node left to the root and C and D of the one right to the root. These two interior nodes themselves are direct descendants of the root, which is represented by the outer pair of parentheses.

Leaves are represented by their names, where a name can be a 'any string of printable characters except blanks, colons, semicolons, parentheses and square brackets' (new [2]) or it may be empty. Note that is also possible to assign names to interior nodes.

Branch lengths can (but do not have to) be included in the representation by putting a colon followed by a real number after a node.

Every tree ends with a semicolon.

When the Newick Format is used to represent an *unrooted* tree, an arbitrary node is chosen as its root. Given that the tree is *binary*, it is still possible to distinguish between a rooted and an unrooted tree. Whereas in a rooted binary tree all nodes, in particular the root, have exactly two direct descendants, the *arbitrary root* in an unrooted binary tree has three. All other interior nodes, however, have two direct descendants.

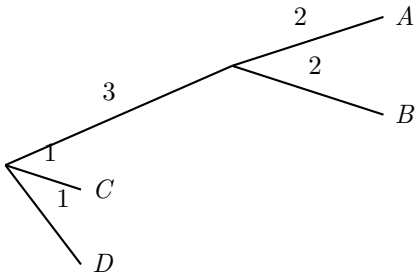

Fig. 11: Newick unrooted: $((A:2, B:2):3, C:1, D:1)$;

7.2 Calculation of the Fair Proportion Index

The *Fair Proportion Index* can be computed easily using its definition (see (4.1)). In a first step, we loop over the edges in T and calculate their contribution $\theta(e)$ to the *Fair Proportion Index* by dividing the edge length λ_e by the number of descendants D_e, respectively. In a second step, we loop over the leaves of T and calculate the *Fair Proportion Index* by summing up those contributions for all edges on the path from the leaf to the root.

Algorithm 1 Fair Proportion Index

1: $n \coloneqq$ number of leaves of T;
2: **for** $e = 1, \ldots, 2n-2$ **do** ▷ $2n-2$ is the number of edges
3: $\quad \theta(e) \coloneqq \frac{\lambda_e}{D_e}$;
4: **end for**
5: **for** $l = 1, \ldots, n$ **do**
6: $\quad FP(l) \coloneqq \sum_e \theta(e)$,
7: $\quad$ where the sum runs over all edges e on the path from l to the root;
8: **end for**

7.3 Calculation of the modified Shapley Value for rooted trees

The *modified Shapley Value* can either be derived from the *Fair Proportion Index* or be calculated directly from its definition. The latter is quite complex, thus for large trees the former is preferable. Still we present both ways.

Using (5.7) the *modified Shapley Value* for a taxon a can be derived from the *Fair Proportion Index* by subtracting $PD(a)/n$.

Algorithm 2 Modified Shapley Value - Version 1

1: Use Algorithm 1 to calculate $FP(l)$ for all $l \in N$;
2: **for** $l = 1, \ldots, n$ **do**
3: $\quad$ Calculate $PD(l)$;
4: $\quad \widetilde{SV}(l) = FP(l) - \frac{PD(l)}{n}$;
5: **end for**

Calculation of the *modified Shapley Value* by direct use of its definition (3.2) requires more considerations.

At first, one needs a way to represent the 2^n subsets of $N = \{1, \ldots, n\}$, reflecting all possible coalitions of taxa. One possible solution is to assign a binary number of length n to every coalition, where a 1 at position i means that taxa i is present in the coalition, whereas a 0 denotes that it is not. E.g. for $n = 3$ taxa this leads to the binary numbers $000, 100, 010, 001, 110, 101, 011$ and 111. Note that the digit sum of each binary number equals the size of the coalition.

In a second step, the $PD(S)$ has to be calculated for every coalition $S \subseteq N$, which is done by summing over all edge weights λ_e in the minimal spanning tree containing S and the root.

Based on this we can loop over all taxa, take into account all coalitions $|S| \geq 2$ (since we are looking for $\widetilde{SV}$) that contain the current taxon and calculate its *modified Shapley Value*.

Algorithm 3 Modified Shapley Value - Version 2

```
1: n := number of leaves of T;
2: Represent the 2^n subsets of N by binary numbers;
3: for all Coalitions S do
4:     Calculate PD(S) := sum_e λ_e,
5:     where the sum runs over all edges e in the minimal spanning tree
6:     containing the taxa in S and the root;
7: end for
8: for l = 1, ..., n do
9:     SV~(l) := 0;
10:    for all Coalitions S do
11:        if l is member of S then
12:            |S| := digit sum of binary number representing S;
13:            SV~(l) := SV~(l) + (|S| − 1)!(n − |S|)!(PD(S) − PD(S \ {l}));
14:        end if
15:    end for
16:    SV~(l) := SV~(l) / n!;
17: end for
```

7.4 Calculation of the Shapley Value for unrooted trees

Although the definition of the *Shapley Value* is the same for both rooted and unrooted trees, its computation differs. While the *Shapley Value* for rooted trees has to be calculated by direct use of its definition or by derivation from the *Fair Proportion Index*, the *Shapley Value* for unrooted trees can be calculated by considering all *splits* induced by the edges of the tree. This approach is less complex than the calculation by definition and therefore needs less computing time. This leads to the following definition (taken from Haake et al. [6]).

Definition 4. Let T be a phylogenetic tree with leaf set N and edge set E. For $a \in N$ and $e \in E$, the removal of edge e splits T into two subtrees. Let $C(a, e)$ denote the set of leaves in the subtree that contains a and let $F(a, e)$ denote the set of leaves in the other subtree, that is "far" from a. Let $c(a, e) := C(a, e)$ and $f(a, e) := F(a, e)$ be the number of leaves in the individual sets, respectively.

Example 5. Consider the following unrooted tree:

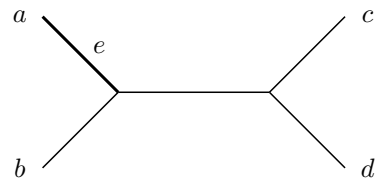

Fig. 12: Splits in unrooted trees

For taxon a, we for example have $C(a, e) = \{a\}$ and $F(a, e) = \{b, c, d\}$, thus $c(a, e) = 1$ and $f(a, e) = 3$ and for taxon d, we have $C(d, e) = \{b, c, d\}$ and $F(d, e) = \{a\}$, thus $c(d, e) = 3$ and $f(d, e) = 1$.
Note that we always have $c + f = n$.

Based on this knowledge, the *Shapley Value* for a taxon a from T can be calculated as follows:

$$\phi(a) = \sum_{e \in E} \lambda_e \frac{f(a, e)}{nc(a, e)}, \tag{7.1}$$

where the sum runs over all edges of T with edge weights λ_e and n is the total number of taxa (see Haake et al. [6], Martyn et al. [10]).

Example 6. Again consider the tree used by (Hartmann [7]), but think of it as unrooted (see Fig. 13). The *Shapley Value* for taxon b calculates as:

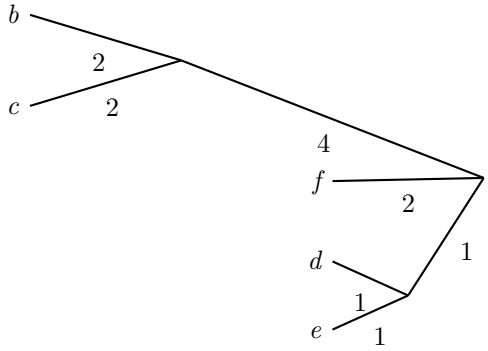

Fig. 13: Unrooted 5-leaf tree

$$\begin{aligned}
\phi(b) &= 2 \cdot \frac{4}{5 \cdot 1} + 2 \cdot \frac{1}{5 \cdot 4} + 4 \cdot \frac{3}{5 \cdot 2} + 2 \cdot \frac{1}{5 \cdot 4} + 1 \cdot \frac{2}{5 \cdot 3} + 1 \cdot \frac{1}{5 \cdot 4} + 1 \cdot \frac{1}{5 \cdot 4} \\
&= \frac{8}{5} + \frac{2}{20} + \frac{12}{10} + \frac{2}{20} + \frac{2}{15} + \frac{1}{20} + \frac{1}{20} \\
&= 3\tfrac{7}{30}
\end{aligned}$$

Analogously we get $\phi(c) = 3\frac{7}{30}$, $\phi(d) = 1\frac{59}{60}$, $\phi(e) = 1\frac{59}{60}$ and $\phi(f) = 2\frac{17}{30}$.

In order to compute the *Shapley Value* for unrooted trees, in a first step loop over all edges e and consider the bipartition of the leaf set N into two subsets $N_1(e)$ and $N_2(e)$, induced by the removal of e. Let $n_1(e)$ and $n_2(e)$ be the number of elements in $N_1(e)$ and $N_2(e)$, respectively. Then use (7.1) to calculate the *Shapley Value* for a taxon a. Note that

$$f(a,e) = \begin{cases} n_1(e) & \text{if } a \in N_2(e) \\ n_2(e) & \text{if } a \in N_1(e) \end{cases} \qquad \text{and} \qquad c(a,e) = \begin{cases} n_2(e) & \text{if } a \in N_2(e) \\ n_1(e) & \text{if } a \in N_1(e) \end{cases}.$$

Algorithm 4 Shapley Value for unrooted trees

```
1: n := number of leaves in T;
2: for e = 1, ..., 2n − 3 do
3:     Calculate the split of the leaf set N induced by the removal of edge e;
4:     N := N_1(e) ⋃ N_2(e);
5:     n_1(e) := |N_1(e)| and n_2(e) := |N_2(e)|;
6: end for
7: for l = 1, ..., n do
8:     SV(l) := 0;
9:     for e = 1, ..., 2n − 3 do
10:        if l ∈ N_1(e) then
11:            f(l,e) := n_2(e) and c(l,e) := n_1(e);
12:        else
13:            f(l,e) := n_1(e) and c(l,e) := n_2(e);
14:        end if
15:        SV(l) := SV(l) + λ_e f(l,e)/(n·c(l,e));
16:    end for
17: end for
```

7.5 Implementation

The computation of the *Shapley Value* and the *Fair Proportion Index* was done in the programming language Perl (Version 5-14), including modules from BioPerl (Version 1.6.901-4; Stajich [14]), and tested under the Linux Distribution *Linux Mint 16 Petra* on a 64-bit computer.

Modules needed to run the scripts are:

- `strict` (a perl pragma to restrict unsafe constructs)
- `warnings` (a perl pragma to control optional warnings)
- `Try::Tiny` (for error handling)
- `Bio::TreeIO`

- `Bio::Tree::Tree`
- `Bio::Tree::Node`

The last three are part of the BioPerl package and provide some helpful tools when dealing with phylogenetic trees. We will shortly describe their functionality.

- `Bio::TreeIO`
  is a Parser for Tree files. It allows to read in trees from data streams and creates `Bio::Tree::TreeI` objects. `Bio::TreeIO` supports different tree formats, of which the Newick Format was chosen in this project.

- `Bio::Tree::Tree`
  inherits `Bio::Root::Root`, `Bio::Tree::TreeFunctionsI` and `Bio::Tree::TreeI` and allows to access several characteristics of the input tree. Methods used in this project were:
  - `$tree->get_nodes()` to obtain a list of all nodes in the tree
  - `$tree->get_root_note()` to access the root of the tree
  - `$tree->get_leaf_nodes()` to receive a list of all leaves in the tree
  - `$tree->get_lineage_nodes($node)`, which returns a full list of the ancestors of a node

- `Bio::Tree::Node`
  inherits `Bio::Root::Root` and `Bio::Tree::NodeI` and provides several tools to get information about certain nodes in the tree. The methods used in this project were:
  - `$node->each_Descendent()`, which returns a list of the direct descendants of a node
  - `$node->branch_length()` to obtain the edge length between a node and its direct ancestor (note that the root node therefore does not have a value assigned to it)
  - `$node->id()`, which returns the human readable identifier of the node, e.g. the species name of a leaf
  - `$node->is_Leaf()` to check whether a node is a leaf or an internal node
  - `$node->get_all_Descendents()`, which returns a list of all descendants of a node (not just the direct descendants)

Based on that, we can now introduce the two Perl-scripts for computing the *Shapley Value* and the *Fair Proportion Index*. Both are command line programs.

7.5.1 shapley.pl

`shapley.pl` computes the *(modified) Shapley Value* for both rooted and unrooted trees. Since it uses Algorithm (3) for the former, it is not advisable to use it for large rooted

trees. Storing the PD for all 2^n subsets of N consumes a lot of memory and slows down the performance.

The command

```
~$ ./shapley.pl --help
```

yields an overview of the possible options to be chosen when running the script:

```
shapley.pl
Compute the (modified) Shapley Value
for phylogenetic trees in Newick format.

Please make sure that BioPerl is installed on your machine!

SYNOPSIS:
shapley.pl --in=filename (--out=filename)

OPTIONS:
--out=filename
Generates an output file containing the computed values.

DESCRIPTION:
Example: shapley.pl --in=myTree --out=myResults
```

As indicated above, this program is not suitable for larger rooted trees, i.e. trees with many taxa. The following table shows, how both computation time and memory usage rise with growing tree size:

Number of taxa	Computation Time (in sec)	Memory Usage (in kb)
5	<1	264
10	<1	428
15	5	7700
20	247	241004
21	529	484632
22	1185	923404

Table 9: Performance of `shapley.pl` for rooted trees

The analysis was done for randomly generated trees of different sizes by use of the Perl-module `Memory::Usage`. Its aim is not to provide absolute numbers for computation time and memory usage, but to indicate an overall tendency. The results show, how

both computation time and memory usage grow exponentially with the number of taxa, reflecting the exponential growth of coalitions to be considered in the computation of the *Shapley Value* for rooted trees.

Performing the same analysis for unrooted trees, shows that Algorithm (4) is less expensive and therefore also suitable for a great number of taxa:

Number of taxa	Computation Time (in sec)	Memory Usage (in kb)
20	<1	264
50	<1	396
100	<1	528
500	2	2120
1000	9	4124
2000	37	8036

Table 10: Performance of `shapley.pl` for unrooted trees

7.5.2 diversity_indices.pl

Depending on the options set by the user, `diversity_indices.pl` computes the *Fair Proportion Index* and/or the *Shapley Value* of one or several trees. Unlike `shapley.pl`, it is suited for large trees, since it uses the Algorithms (1, 2 and 4), which are not as expensive as Algorithm (3).

The command

```
~$ ./diversity_indices.pl --help
```

yields an overview of the possible options to be chosen when running the script:

```
diversity_indices.pl
Compute the Fair Proportion Index and (modified) Shapley Value
for phylogenetic trees in Newick format. Note,that the FP-Index
is only defined for rooted trees, while the SV-Index can be
calculated for both rooted and unrooted trees.

Please make sure that BioPerl is installed on your machine!

SYNOPSIS:
diversity_indices.pl --in=filename (--out=filename) (--values=value)

OPTIONS:
--out=filename
```

```
Generates an output file containing the computed values

--values=value
Choose the diversity index to calculate.
Options for value:
0 Calculation of FP
1 Calculation of SV
If the option is not chosen or used with an argument other than {0,1}
both the Fair Proportion Index and the Shapley Value will be
calculated.

DESCRIPTION:
Example: diversity_indices.pl --in=myTree --out=myResults --values=0
```

The following table gives an overview of the performance of `diversity_indices.pl`. The analysis was done for the computation of the *modified Shapley Value*. As it is derived from the *Fair Proportion Index*, the numbers already include the computation of the latter. As indicated above, `diversity_indices.pl` is much faster than `shapley.pl` and therefore should be preferred in practice.

Number of taxa	Computation Time (in sec)	Memory Usage (in kb)
20	<1	264
50	<1	396
100	<1	524
500	<1	1740
1000	1	3384
2000	3	6608
5000	14	16696

Table 11: Performance of `diversity_indices.pl` for rooted trees

8 Discussion

Biodiversity conservation often implies deciding on the species to conserve, since resources are limited. Two simple indices, which can be used as a guideline in the decision-making process, are the *Shapley Value* and the *Fair Proportion Index*. Both indices indicate the distinctiveness of a taxon and thus its importance to overall biodiversity. The *Shapley Value* is difficult to calculate, but offers an appealing biological interpretation. It can be seen as the average contribution a taxon makes to biodiversity and thus proves to be a sensible prioritization criterion. The *Fair Proportion Index*, in contrast, lacks a biological explanation, but is easier to calculate. Therefore it has been preferred to the *Shapley Value* in practical applications. This practice is justified as the *Fair Proportion Index* and the *Shapley Value* are actually the same or at least strongly correlated depending on whether the *original* or *modified* version of the latter is used. The *modified Shapley Value*, however, can be derived from the *Fair Proportion Index*, which allows for a less complex computation than by direct use of its definition.

Despite the strong correlation between the *Fair Proportion Index* and the *modified Shapley Value*, the two measures do not always result in the same ranking order of taxa. While they naturally lead to the same ranking order in case of ultrametric trees, it is possible to find non-ultrametric trees, in which they differ. Surprisingly, this seems to happen the more often the larger the tree. So far, we have not found a sufficient explanation for this phenomenon, thus it could be subject to further research. Still, it may be advisable to not only rely on the mere ranking order provided by either the *Fair Proportion Index* or the *Shapley Value*, but to compare the indices and examine the tree structure before deciding on the taxa to conserve.

A more profound approach to biological conservation might also have to take into account parameters not considered by neither the *Fair Proportion Index* nor the *Shapley Value*, such as the variable costs involved in conserving taxa or their individual risks of extinction. This makes the decision-making process more thorough, but also more complex to put into practice.

Therefore the *Fair Proportion Index* and the *Shapley Value*, which can be calculated and realized more easily, can be considered a useful, even though not exclusive, tool in biological conservation.

References

[1] www.edgeofexistence.org. [Online; accessed 06-August-2014].

[2] The newick tree format. http://evolution.genetics.washington.edu/phylip/newicktree.html, . [Online; accessed 22-August-2014].

[3] Newick format. http://en.wikipedia.org/wiki/Newick_format, . [Online; accessed 22-August-2014].

[4] http://biodiversity.ca.gov/Biodiversity/biodiv_def2.html, 2014. [Online; accessed 31-July-2014].

[5] M. Fuchs and E. Y. Jin. (almost) equality of shapley value and fair proportion index in phylogenetic trees.

[6] C.-J. Haake, A. Kashiwada, and F. E. Su. The shapley value of phylogenetic trees. *J. Math. Biol.*, 56(4):479–497, Sep 2007. ISSN 1432-1416. doi: 10.1007/s00285-007-0126-2. URL http://dx.doi.org/10.1007/s00285-007-0126-2.

[7] K. Hartmann. The equivalence of two phylogenetic biodiversity measures: the shapley value and fair proportion index. *J. Math. Biol.*, 67(5):1163–1170, Sep 2012. ISSN 1432-1416. doi: 10.1007/s00285-012-0585-y. URL http://dx.doi.org/10.1007/s00285-012-0585-y.

[8] K. Hartmann and M. Steel. Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the noah's ark problem. *Systematic Biology*, 55(4):644–651, Aug 2006. ISSN 1076-836X. doi: 10.1080/10635150600873876. URL http://dx.doi.org/10.1080/10635150600873876.

[9] N. J. Isaac, S. T. Turvey, B. Collen, C. Waterman, and J. E. Baillie. Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PLoS ONE*, 2(3):e296, Mar 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000296. URL http://dx.doi.org/10.1371/journal.pone.0000296.

[10] I. Martyn, T. S. Kuhn, A. O. Mooers, V. Moulton, and A. Spillner. Computing evolutionary distinctiveness indices in large scale analysis. *Algorithms for Molecular Biology*, 7(1):6, 2012. ISSN 1748-7188. doi: 10.1186/1748-7188-7-6. URL http://dx.doi.org/10.1186/1748-7188-7-6.

[11] R. F. Noss. Indicators for monitoring biodiversity: A hierarchical approach. *Conservation Biology*, 1990.

[12] C. on Biological Diversity. Article 2. use of terms. http://www.cbd.int/convention/articles/default.shtml?a=cbd-02. [Online; accessed 1-August-2014].

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org. ISBN 3-900051-07-0.

[14] J. E. Stajich. The bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, Oct 2002. ISSN 1088-9051. doi: 10.1101/gr.361602. URL http://dx.doi.org/10.1101/gr.361602.

[15] M. Vellend, W. K.Cornwell, K. Magnuson-Ford, and A. O.Mooers. Measuring phylogenetic biodiversity.

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.
This paper was not previously presented to another examination board and has not been published.

Greifswald, 9th October 2014