**On the estimation of species age from phylogenetic trees**

**Introduction**

Species age expresses evolutionary history; thus, its association with ecological high-level processes sensu Vellend 2010 gives an insight into the interplay between evolution and ecology. For speciation, species age directly spurs from diversification rates and therefore it relates to the history and taxa’s relationships of the clade itself. In the case of selection, it has been postulated that species age, by its influence over traits associated with fitness, would affect survivorship both negatively and positively (Benton and other). Biotic and abiotic changes could make older species that originated in a different environment more prone to extinction than recent species. In contrast, strong competition from older species could inhibit recent species to establish, leading to a high early extinction. For dispersal, species age could be a measure of colonization time, especially in island systems or biotic invasions triggered by geological events, such as the formation of the Central American Isthmus for the Great American Interchange (Tanentzap and GAI). In the case of drift, species age could reflect population size. At its inception, species start with a small population being more prone to drift. On the other hand, if species age is positively related to specialization, old species would tend to have lower populations and experience greater drift. Thus, when associated with ecological processes, species age can offer insights into the history that molded past and extant biodiversity.

There are two ways of measuring species' ages depending on whether the species is extinct or extant. For the former, species age is measured in the fossil record through different statistical and probabilistic approaches based mostly on taxa’s stratigraphic duration. On the other hand, for extant species, some studies have calculated species ages from the branch lengths of time-calibrated phylogenies. Crown age is the interval between species origination, or the divergence from its nearest sister, to the present. In contrast, stem age is the lapse between the origination of the most recent common ancestor to the present. Nevertheless, estimating species age from phylogenetic trees (hereafter called “phylogenetic age”) can lead to misinterpretations due to three different features not captured by calibrated phylogenies.

First, deficient taxon sampling, especially in extreme habitats and the Global South, may reduce the accuracy of the tree topology, causing uncertainty in any phylogenetic estimation (heath2008, mynard 2023, cai2022). Second, based on the birth-death process, phylogenetic estimators cannot recover true extinction rates (Rabosky 2010). Omitting extinct or unsampled species from the phylogeny calibration would produce that the branch lengths of extant species would appear larger than they are. As a remarkable example, consider the Homo sapiens’ phylogenetic age when our extinct Homo relatives are not considered in the phylogeny; our species age goes to approximately 6 million years (to our common ancestor with the chimpanzee) being that the oldest fossil record for our species is of 300 thousand years .

Third, phylogenetic trees do not capture species’ evolution modes. For budding speciation or a diversification event where one new species originates while the ancestral species persists, the phylogenetic age can underestimate the true age because the branch lengths for the ancestral species are shorter than they truly are (Figure 1). Besides, regarding unsampled or extinct relatives, phylogenetic age can both under and overestimate true age. For bifurcating speciation or a diversification event that gives rise to two new species that replace the ancestral species there is no age misinterpretation. However, phylogenetic age can overestimate true age for this mode of evolution regarding unsampled or extinct species. For anagenetic speciation or a diversification event where one species arises, replacing the ancestral species, the phylogenetic age always overestimates the true age and the problem gets worst regarding missing species. Thus, we should not estimate extant species' ages merely from calibrated phylogenetic trees.

This paper aims to demonstrate the multiple sources of error associated with phylogenetic age. First, we used simulations to show how different error measures, both quantitative and qualitative, vary regarding different evolution modes and diversification rates. Next, we proposed three predictive models to partially correct species' age using features easily obtained from phylogenies. We evaluated the model’s performance with error measures and applying them in a comparison among phylogenetic age and fossil age for a set of extant species. We hope this paper will provoke discussions about the evolutionary information captured by phylogenetic trees and the pertinence of using it for inferences about the interplay between evolution and ecology.

**Methods**

The three methodological sections follow the same workflow:

* Phylogenetic trees simulation through the "sim.bd.taxa" function from the TreeSim package (stadler 2011), controlling the number of extant species, speciation and extinction rates.
* Taxonomy simulation through "sim.taxonomy" function from the FossilSim package (barido2019fossilsim), using the trees generated previously and controlling the speciation modes.
* The "sim.taxonomy" function allows us to save the True age of species. We pruned the extinct species from the phylogenies using the "prune.fossil.tips" from the FossilSim package (barido2019fossilsim). Then, we estimated the Phylogenetic ages from the branch lengths of the pruned trees.

*Quantifying the error*

First, we simulated 100 phylogenetic trees with 0.1, 0.5, and 1 speciation rates, a turnover vector varying from 0 to 0.99, and 100 extant species or tips (300 trees in total). Second, using the trees as inputs, we simulated four sets of taxonomies controlling the modes of speciation parameters as follows:

* Only Budding speciation.
* Only Bifurcating speciation.
* Budding speciation with anagenetic speciation proportional (0.5) to the speciation rate.
* Bifurcating speciation with anagenetic speciation proportional (0.5) to the speciation rate.

Thus, we obtained in total 120 000 extant species, with 30 000 corresponding to each speciation mode. From the True and Phylogenetic ages, we calculated the relationship between both variables for each speciation mode, the Mean Absolute Error (MAE) vs Turnover, and the Mean Absolute Percentage Error (MAPE) vs Turnover.

*Qualitative error*

First, we simulated 1000 phylogenetic trees with 0.9, 0.5, and 0 extinction rates, a fix speciation rate (1), and 100 extant species or tips (3000 trees in total). Second, using the trees as inputs, we simulated 3000 taxonomies contemplating only budding speciation due to the phylogenetic assumption of birth-death models (reference). Thus, we obtained in total 300 000 extant species. We used the corresponding root age of the phylogenetic tree to scale the True and Phylogenetic ages, and the three different extinction scenarios for the following graphical representations:

* The distribution of True and Phylogenetic ages.
* The difference between True and Phylogenetic age for each of the 30 000 species, indicating the percentage of correct estimation.
* One distribution accounting the difference between the oldest and youngest species for each tree according to the Phylogenetic ages and a second distribution using the same two species per tree but representing the True age difference.

*Predictive models*

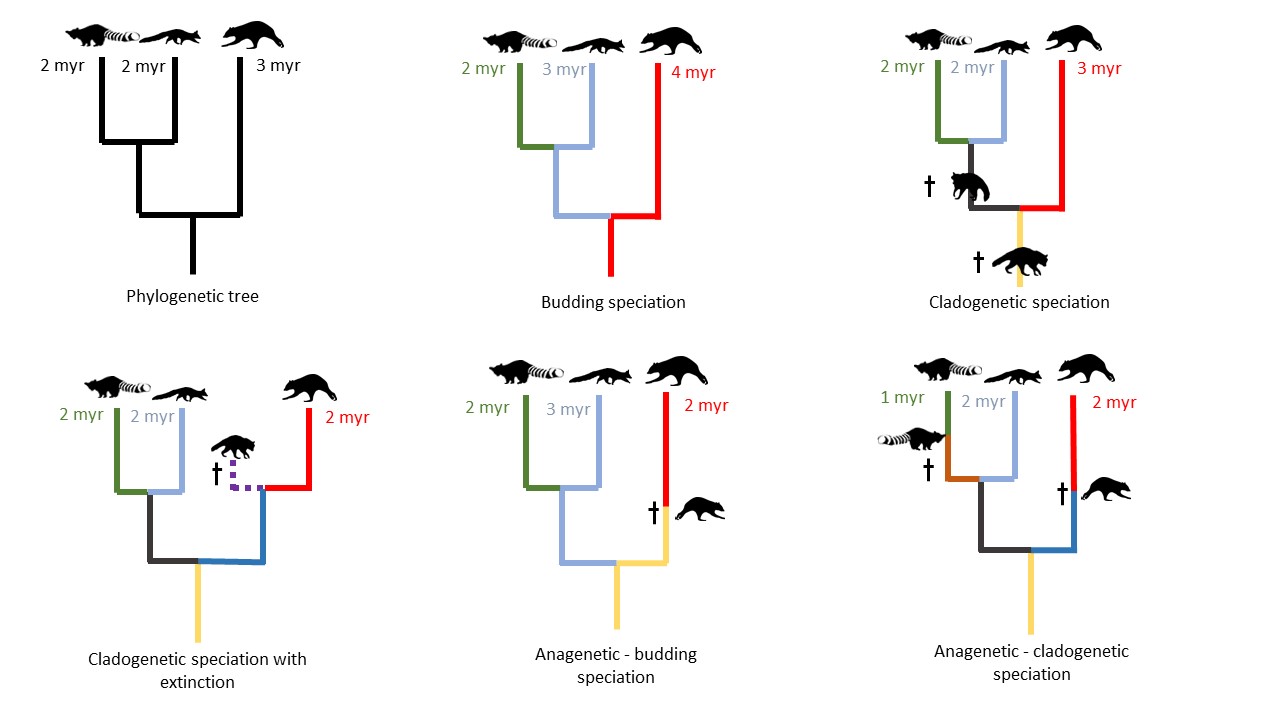
First, to obtain our training dataset, we simulated 100 phylogenetic trees with a random number of extant species (varying from 20 to 100) and assigning the speciation and extinction rates from diversification and turnover rates generated randomly (varying from 0 to 1). Second, using the phylogenies as input, we simulated four sets of 100 taxonomies controlling the modes of speciation parameters as follows:

* Only Budding speciation.
* Only Bifurcating speciation.
* Budding speciation with anagenetic speciation proportional (0.5) to the speciation rate.
* Bifurcating speciation with anagenetic speciation proportional (0.5) to the speciation rate.

Thus, we obtained 400 extant species or tips, with 100 corresponding to each speciation mode. We built the train dataset with the True age as response and Phylogenetic age, phylogenetic root age, number of species in the tree, number of sister species, mode of speciation and diversification and turnover rates. We log10 transformed all variables except mode of speciation, diversification and turnover rates. We fitted three predictive models: i) Bayesian Neural Network (BNN); ii) Bayesian Generalized Additive Models; iii) Linear Model (LM). Next, we simulated our test dataset using the same workflow as the training dataset but instead of 100 we simulated 25 trees. Then, we evaluated the performance of the models with the test data, the same variables log10 transformed, comparing the Mean Squared Error (MSE) between the corrected Phylogenetic ages (predicted by the models) and the True ages. Finally, we used our BNN model for predicting the corrected Phylogenetic ages from the following species: i) *Carcharhinus obscurus;* ii) *Trianodon obesus;* iii) *Balaena mysticetus;* iv) *Homo sapiens;* v) *Ursus arctos;* vi) *Vulpes velox;* and,vii) *Acinonyx jubatus*. We compared them to the fossil (the oldest record of the species) and the Phylogenetic ages. The sharks phylogenetic and fossil age were obtained from (bree2022); the *Homo sapiens* fossil age was obtained from (callaway2017oldest) and the Hominoidea phylogeny from (arnold201010ktrees). The rest of the mammals phylogenetic and fossil ages were obtained from (ask Silvestro).

**Results**

**Figure 1**



**Figure 2**