LONDON’S GLOBAL UNIVERSITY

UCL

University College London
Department of Genetics, Evolution and Environment

The influence of vertebrate species traits on their responses to land-use and climate change

Adrienne Etard

Primary supervision: Dr. Tim Newbold
Secondary supervision: Dr Alex Pigot

March 19, 2019

Abstract

Contents

List of Tables

List of Figures

List of abbreviations

BM	Body mass
BL	Body length
DA	Diel activity
Di	Diet
DB	Diet breadth
GL	Generation length
HB	Habitat breadth
L	Longevity
LCS	Litter/clutch size
TL	Trophic level
ITIS	Integrated Taxonomic Information System
LUCC	Land-use and climate change
MA	Maturity
PD	Primary diet
PREDICTS	Projecting Responses of Ecological Diversity In Changing Terrestrial Systems
RS	Range size
SI	Supporting Information

1 | Introduction

2 | Literature review

3 | Collecting and imputing ecological trait data across terrestrial vertebrates

3.1 Introduction

A growing body of research uses trait-based approaches to understand how biodiversity links to ecosystem functioning, and how environmental changes are likely to affect species non-randomly with respect to their traits (Hevia et al). Strictly, traits are defined as characteristics measurable the level of an individual, with an effect on organismal fitness or performance. They can be physiological (e.g., metabolic rates), morphological (e.g., body mass), behavioural (e.g., learning) or phenological (e.g., anthesis), or can relate to species life-history (e.g. longevity). This definition can be broadened to include characteristics measurable at the species level, such as the number of habitats known to be used by a species (habitat breadth). Here, I use this broader definition of traits and refer to these as ecological traits.

Many studies have shown that traits influence species responses to environmental pressures (ref). Moreover, it is now accepted that ecosystem functioning is positively correlated with species diversity (Tilman 2014). Species traits can provide a mechanistic understanding of both species roles in ecosystem functioning and of species responses to changes. Traits shape species fundamental and realised niches; for instance, physiological traits influence species thermal tolerances, participating in defining their geographical distributions. Traits such as trophic level or body mass structure food webs and affect inter- and intra-specific competition (ref). As such, traits determine and reflect species use of their environment. Specifically, effect traits define organismal contributions to ecosystem functions. Effect traits are underpinned by species resource use, and this applies at diverse scales, from single-celled nutrient cycling bacteria to large mammals. Response traits are those involved in determining species responses to environmental changes and can overlap with effect traits.

Although terrestrial vertebrates have been extensively studied in the past (Titley et al), the vast majority of research investigating the impact of environmental changes on ecosystem functions has focused on plants and invertebrates (Hevia et al). Vertebrates nevertheless play diverse ecosystem roles, and some are important keystone species. Vertebrate species particularly contribute in food web structures and population dynamics through predatory and herbivory activity. They are pollinators and seed dispersers, and overall participate in nutrient cycling at higher levels. Understanding how environmental changes may affect their ecological roles is important to predict future ecosystem functioning, and to put into place appropriate mitigation measures. The end-goals of my PhD thesis are to elucidate how species traits influence their responses to land-use and climate change, and how this links to changes in ecosystem functioning. Addressing these questions requires to use extensive trait data. Despite vertebrates having been the focus of much research, and despite the growing interest for trait-based approaches, there exist no comprehensive database of vertebrate ecological traits encompassing all classes. Consequently, collating trait data was a prerequisite for any further work, and this operation was constrained by the amount of information available in the literature. The present chapter focuses on data collection methods and missing trait values imputations. Thanks to past and recent efforts to release data in the public domain, at least four comprehensive ecological trait databases are now freely accessible (mammals: Pantheria, amphibians: Amphibio, amniotes: Myhrvold, mammals and birds: Cooke et al). Other trait datasets have been released on online platforms alongside published articles (e.g. Global Assessment of Reptile Distribution initiative, `http://www.gardinitiative.org/`), or can be downloaded from online databases (IUCN Red List (`https://www.iucnredlist.org/`), BirdLife data zone ((http://datazone.birdlife.org/home)). Trait data available from primary sources for mammals and birds is likely to be more abundant and more resolved than for reptiles and amphibians, due to systematic biases in sampling with regards to taxonomic groups (Newbold, *manuscript*).

The present chapter details the methodology I employed to collate trait information across terrestrial vertebrates. Primary sources offered a variety of traits, of which only a few were selected. Trait selection was motivated by two main reasons: (1) traits should be of ecological interest and be related to response or effect processes; (2) trait values should be available for many species, across the four terrestrial vertebrate classes, allowing for cross-classes comparative analyses. The selected target traits related to species life-history and morphology (body mass; longevity; litter/clutch size; diel activity; trophic level; diet) and to their habitat preferences (habitat breadth and specialisation). Reptilian diet was not readily available in primary data sources, and one exception was made as I extracted diet data for the other classes. Species mobility was hardly available across sources, and no similar variable could describe species mobility across classes; although species' abilities to move in their environment is likely to strongly impact their responses to threats, this trait was not considered for the above reasons.

The present chapter details the methodology I employed to collate selected traits. I elaborate on some of the challenges met when compiling data across many species, such as inconsistency of taxonomy across sources. Not unexpectedly, the amount of missing values was highly variable across classes and traits. I examined whether missing trait values presented patterns, both across and within classes. To achieve full trait coverage across species, I imputed missing trait values using random-forest algorithms.

In October 2018, Cooke et al released a comprehensive database of six mammalian and avian traits. They collated and imputed missing trait values for body mass, litter/clutch size, volancy, diel activity, primary diet and habitat breadth. As similar primary sources were used in both our data collection, I did not use their database to complement my sources. Moreover, the imputation methods they used to fill gaps in trait coverage differed from mine. I used this freely accessible compiled data as an opportunity to compare the results of both our data collection and imputation processes. Results of this comparison are available in the SI.

3.2 Methods

3.2.1 Ecological trait data collection

Primary data sources.

I collated ecological trait data for terrestrial vertebrates from the sources figuring in Table 3.1. Information was compiled for the following target traits: body mass, longevity, litter or clutch size, trophic level, diel activity, diet, and habitat preferences. I also compiled traits that were potentially correlated to either body mass or longevity, to be used as potential predictors in imputations of missing values. As such, body length information was compiled when available, as well as generation length or age at sexual maturity. Most notably, longevity was chosen over generation length or age at sexual maturity as it was the only common currency across classes reflecting generation turnover. In addition, species geographical range sizes were estimated from distribution data, extracted from the IUCN Red List.

Table 3.1: Primary sources used for each compiled trait. Primary sources may contain more traits than shown here. **BM**: body mass; **BL**: body length; **L**: longevity or maximum longevity; **GL**: generation length; **LCS**: litter or clutch size; **TL**: trophic level; **Di**: diet; **DA**: diel activity; **RS**: range size; **H**: habitat data. Bolded abbreviations highlight target traits; other traits were added for potential correlations in further imputations.

Sources	**Taxa**	**Traits**									**RS**	**H**
		BM	BL	**L**	MA	GL	**LCS**	**TL**	**Di**	**DA**		
Amphibio	Amphibians	✓	✓	✓	✓		✓	✓	✓	✓		
Cooper			✓				✓				✓	
Senior			✓									
Bickford			✓								✓	
Elton	Birds	✓							✓	✓		
Butchart		✓				✓						
Pantheria	Mammals	✓	✓	✓	✓		✓			✓		
Kissling1								✓				
Kissling2								✓				
Elton		✓							✓	✓		
Pacifici		✓		✓	✓	✓						
Scharf	Reptiles	✓		✓	✓		✓	✓		✓		
Vidan										✓		
Stark		✓		✓			✓			✓		
Schwarz							✓					
Novosolov1		✓						✓			✓	
Novosolov2							✓					
Slavenko		✓										
Myhrvold	Amniotes	✓	✓	✓	✓		✓					
IUCN	Vertebrates										✓	✓

Compilation methods.

Continuous traits. All continuous traits were averaged within species when different sources provided estimates. Longevity and maximum longevity were assumed to provide the same information and were averaged within species. No measure of intra-specific variability was compiled and estimates were provided as a single measure for each species.

Categorical traits.

Activity time. Species were described as being either nocturnal or non-nocturnal. Despite a higher resolution of activity time information in some of the primary sources (e.g. species being described as cathemereal, crepuscular or strictly diurnal), I adopted the classification of the primary source with the lowest resolution, in order to have consistent information across classes.

Diet and diet breadth. For mammals and birds, diet was compiled from the Elton Traits database (ref). Primary diet was available in the avian dataset and declined into five categories: (1) plant or seed consumers; (2) fruit or nectar consumers; (3) vertebrate consumers, including fish and carrion; (4) invertebrate consumers; and (5) omnivores. Primary diet was not available for mammals. Instead, mammal diet was only described as the percent use of different food items. I pooled these items together into the same five primary diet categories as for the avian dataset. Any food items for which percent use was equal to or above 50% were considered to be primary food items. Species for which no food item had percent use above 50% were considered to be omnivores. For amphibians, diet information was extracted from AmphiBIO. Diet information was available as binary variables for diverse food items. Percent use were not recorded, so these items were considered to form species primary diet. I pooled amphibian species into the five diet categories described above.

Trophic level. For amphibians and birds, trophic levels were partly inferred from the primary diet.

Habitat preferences. Species habitat preferences were compiled from IUCN habitat data files and were described as a binary variable recording whether a species was known to occur in a particular habitat. I calculated habitat breadth as the number of habitats a species was known to use. Weights were assigned to each habitat in this calculation depending on the recorded habitat suitability and importance; outcomes were not very sensitive to the presence of weights (compared to a non-weighted sum, see SI). Finally, a broad degree of habitat specialisation was produced. If any artificial habitat was recorded to be suitable, species were reported to be generalists; else, they were natural habitat specialists. More details on habitat preferences compilation are provided in the SI.

3.2.2 Phylogenetic information

I obtained phylogenetic trees for birds, amphibians, mammals and squamates from Hedges et al (2015) (available at `http://www.biodiversitycenter.org/ttol`, downloaded 06/07/2018). All trees were ultrametric and fully resolved, except for the amphibian tree which presented polytomies. All trees contained a few branches of length 0 (193 branches for mammals, 136 for amphibians, 189 for birds and 284 for reptiles).

3.2.3 Tackling taxonomic synonymy

Across the different primary sources, similar species could appear under different binomial names. This was a problem when matching datasets by species. It was also problem when matching species to the PREDICTS database. Moreover, it is possible than within a primary source, a given species was appearing under two or more different names. As such, taxonomic synonymy created 'pseudoreplicates' of the same species, overall falsely increasing the total number of species and artificially inflating the amount of missing trait values. Taxonomic synonymy was hence a major issue; due to the large number of species across datasets, extensive manual checks could not be applied. The presence of typos in species names had the same effect as synonymy, erroneously duplicating species. I attempted to correct for taxonomy first by correcting for typos, and second by identifying species which were entered under a synonymic name and replacing these with the accepted name. To this end, I developed an automated procedure, complemented with a few manual entries. Obvious cases where vernacular names had been entered in the place of binomial names were also treated manually; that was the case for 44 PREDICTS species (when possible, I best assigned binomial names to species common names; unidentifiable species were left empty and assigned to a genus (5 species)).

Automated procedure and outputs.

Extracting names from the IUCN Red List and the Integrated Taxonomic Information System (ITIS). The automated procedure consisted in extracting species accepted and synonymic binomial names from the IUCN Red List or from the ITIS, using the rredlist and taxize R packages. I started by generating a list of all names figuring across datasets (primary sources, phylogenies and PREDICTS). These 'original' names were corrected for typos; then, the IUCN Red List was queried and synonyms and accepted names were stored when possible. When species were not found in the IUCN Red List, information was extracted from ITIS. When species were not found in ITIS either, corrected names were assumed to be accepted. Family and order information was extracted using the same procedure and some entries were completed using the Global Biodiversity Information Facility taxonomic backbone (`https://www.gbif.org/tools/species-lookup`).
NB: for species entered with the forms *Genus cf.*, *Genus aff.* or *Genus spp.*, the accepted name was left empty.

Outputs. I generated a list of vertebrate species, recording whether species names were accepted or synonymic (for 14124, 8743, 6090, and 11183 names or identifiers found across datasets for birds, amphibians, mammals and reptiles respectively, including species names as they appeared in phylogenetic trees). For each name, the identified accepted name and the synonyms were stored when possible, as well as additional taxonomic information (order, family, genus). When queries did not succeed, species accepted names were assumed to be the original names found in the datasets.

Harmonising taxonomy in trait datasets. Taxonomy across datasets was finally homogenised by replacing recorded synonyms with their accepted scientific names. Overall, this procedure reduced the total number of species figuring in trait datasets (Figure 3.1). The species presenting the highest degree of pseudoreplication was the East African mole rat (*Tachyoryctes splendens*), which was figuring under 12 names identified as being synonymic across primary sources (Figure 3.1B), highlighting the need for normalising taxonomy across sources.

Despite the automation efforts, taxonomic redundancy persisted to a degree in the trait datasets. Indeed, at this stage, not all species in PREDICTS matched a species in the trait datasets. Additional manual inputs were required to resolve taxonomic synonymy for these species. Verifying the presence of PREDICTS species in trait datasets was important for further analyses. Taxonomic synonymy was resolved manually for 91 PREDICTS species that did not match any species in the trait datasets; in that case, information was extracted from other diverse sources (such as the Reptile Database (`http://www.reptile-database.org/`); Avibase (`https://avibase.bsc-eoc.org/avibase.jsp?lang=EN&pg=home`); AmphibiaWeb (`https://amphibiaweb.org/`)). After adding manual inputs to the synonym datasets, all PREDICTS species were represented in trait datasets.

The need to apply additional manual inputs underlines the fact that the automated procedure was not optimal. The Red List and the ITIS were not comprehensive taxonomic sources, and for

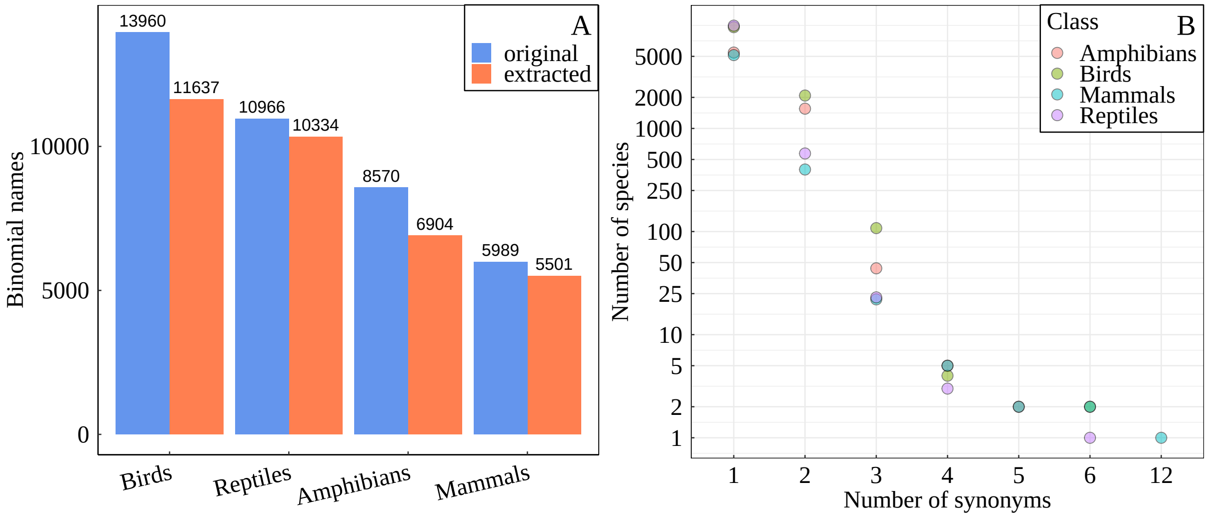

Figure 3.1: Difference in species number due to taxonomic correction (A) and distribution of number of synonyms across datasets (B). (A) shows the number of species across all primary sources (trait datasets and PREDICTS, excluding phylogenies), before and after correcting for taxonomy. Replacing identified synonyms by the extracted accepted name reduced the number of species in all classes, with the most drastic reduction for birds (decrease by 2,323 unique binomial names). The diminution was of 632 unique identified species for reptiles, of 1,666 for amphibians and of 488 for mammals. **(B)** shows the distribution of the number of synonymic names. In all four classes, more than 5,000 species (or binomial names) had no identified synonyms. Nevertheless, a large amount of species had two identified synonyms (range: 400 species for mammals - 2086 for birds). The most replicated species was the East African mole rat *Tachyoryctes splendens*, for which 11 synonyms were identified.

clades with high degrees of pseudoreplication in names, such as reptiles or amphibians, neither the Red List or the ITIS were fully resolved. As I only applied manual checks for PREDICTS relevant species, ‘pseudoreplication’ and taxonomic errors are likely to have persisted to a degree. Moreover, certain species were entered using the format *Genus subspecies* rather than *Genus species*; for these, automated queries may have failed to identify the species.

Harmonising taxonomy in phylogenetic trees and increasing species phylogenetic representation.

Taxonomic correction across tip labels. Efforts to correct datasets for taxonomy created problems for a marginal proportion of species when dealing with phylogenies. The idea of the procedure described above was to replace two or more identified synonyms by a single accepted name, and then collapsing dataset rows together by names. I applied the same method on phylogenies, replacing synonyms by their identified accepted names in trees’ tip labels. Not unexpectedly, in some cases, the procedure ended up assigning the same accepted name to different phylogenetic tips. This was the case for 2.8% of mammalian, 1.7% of avian, 1.6% of amphibian and 1.7% of reptilian species, which then had multiple phylogenetic positions (most having two different positions, see SI). Because keeping several putative phylogenetic positions for a species was problematic in further analyses, I selected one tip to conserve and dropped other tips from the phylogenies (Figure 3.2). To briefly describe the procedure, if replicated tips were sister clades, the tip to conserve was chosen randomly among the replicates. Else, I chose to conserve the tree tip whose position was closest to the position of the same tip in the uncorrected tree, when present. In all other few cases, tips to drop were chosen randomly. Further details on how replicated tips were dropped are available in the SI (with 3 examples for each case of Figure 3.2).

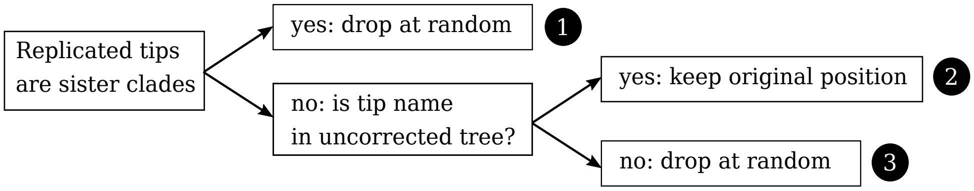

Figure 3.2: Procedure followed to drop replicated tips from phylogenies. Most of these were replicated twice. When replicated tips were sister clades, the tips to drop were chosen randomly, as it did not affect the 'true' phylogenetic position of the species (1). When replicated were not sister clades, I kept the tip whose position was closest to the position of the same tip in the uncorrected tree (2). In a few cases, the corrected name did not appear in the original tree. Those were problematic cases, and the tips to drop were chosen randomly (3). Nevertheless, occurences of that third case were rare (see SI).

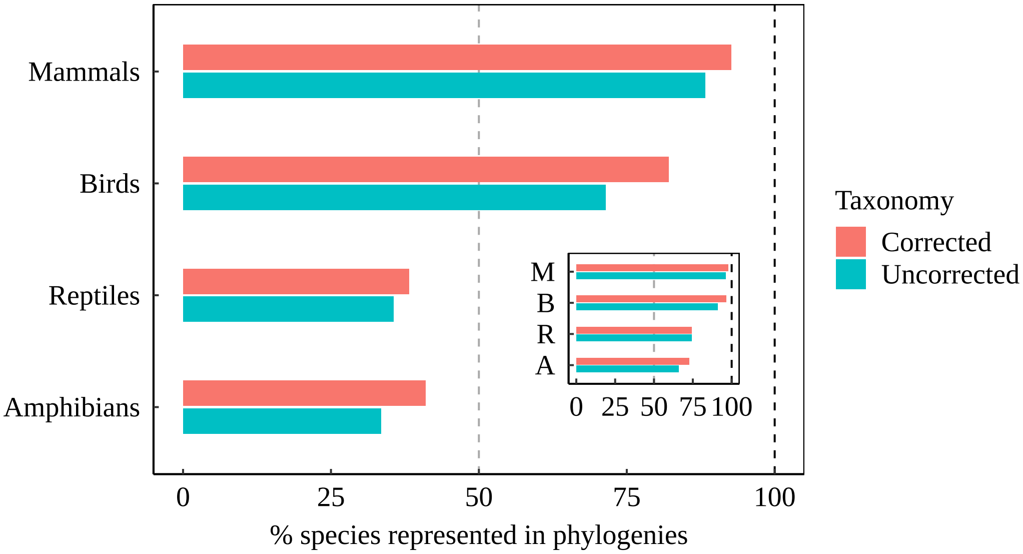

Figure 3.3: Percentage of species represented in the phylogenies for both corrected and uncorrected trait datasets. Overall, taxonomic correction increased species representation in phylogenetic trees. Representation for mammals and birds was high (after taxonomic correction: 82% of avian and 93% of mammalian species had a phylogenetic position). On the other hand, reptiles and amphibians were poorly represented (after taxonomic correction: only 38% of reptilian and 41% of amphibian species were placed in phylogenetic trees). The inset barplot shows representation for species figuring in PREDICTS. For these, species presence in phylogenetic trees after correction was high across all classes, with a minimum representation of 76% for amphibians.

Correcting for taxonomy in the phylogenies: conclusions. Overall, correcting for taxonomy in phylogenies improved species representation in the trees (Figure 3.3. For amphibian and reptilian species figuring in PREDICTS only, phylogenetic representation disproportionally increased (with a minimum representation of 76% for PREDICTS amphibians after correcting the trees for taxonomy, inset plot in Figure 3.3). Nevertheless, correcting phylogenetic tip labels generated replicates for a marginal number of tips, which then had to be dropped.

Species attachments to phylogenetic trees. Some species in the trait datasets were not represented in the phylogenies. Maximising the number of species represented in the phylogenies was important for further trait imputations. Indeed, if traits were evolutionary conserved, species phylogenetic position could be an important predictor of trait values. To maximise species representation, I added some species to the root of their genus, when possible (phytools package). Attaching species at the root of their genus created polytomies, which were resolved randomly (using multi2di and bifurcatr, ape and PDcalc packages). Resulting trees contained additional branches of length zero. Such modifications of the phylogenetic trees could have altered the significance and the strength of trait phylogenetic signal. I further verify whether these alterations of the trees had impacted phylogenetic signal, by qualitatively comparing the strength and the significance of phylogenetic signal for each trait, estimated using both original trees and augmented trees (see 'Assessing phylogenetic signal in traits').

Finally, a large number of species were attached to their genus in the trees (Table 3.2). For instance, only 38% of the species figuring in the reptilian trait dataset were initially found in the squamate phylogeny. After attaching non-represented species, 91% of the species were placed in the squamate phylogeny.

Table 3.2: Species representation in phylogenetic trees (corrected for taxonomy). The number of species attached to the root of their genus ranged from 175 (mammals) to 5438 (reptiles). Finally, most species were represented in the phylogenies, whereas more than half reptilian and amphibian species initially had no known phylogenetic position.

Class	Initially not in tree	Of which randomly attached	No final representation in tree
Amphibians	59% (4040 of 6904)	96% (3883 of 4040)	**2.3%**
Birds	18% (2085 of 11637)	75% (1574 of 2085)	**4.4%**
Mammals	7.4% (407 of 5502)	43% (175 of 407)	**4.2%**
Reptiles	62% (6391 of 10334)	85% (5438 of 6391)	**9.2%**

3.2.4 Exploring biases in the coverage and completeness of trait information across classes

Having normalised taxonomy and compiled trait data, I assessed trait coverage, defined as the percentage of species for which trait information was available for a given trait. I also estimated the amount of trait information available for a species by calculating trait completeness. For a species, trait completeness was defined as the proportion of traits for which information was available (number of non-missing trait values divided by total number of traits). In corrected datasets, species with 0% completeness in predictor traits were filtered out.

Further, I examined whether patterns in the distribution of missing values emerged within classes, as particular clades or parts of the phylogenies could be under-sampled compared to other clades. Whether values are missing at random is likely to impact imputation errors, notably if some taxa appear to be under-sampled. To assess whether missing values presented patterns, I represented within-family median completeness and coverage values in each branch of phylogenetic trees built at the family level. Tree branches were colour-coded to reflect median values in each family. Specifically, within-family trait completeness was calculated by aggregating species into their families and calculating the median trait completeness within each group.

Patterns of missing values in trait coverage were explored for each trait separately. Trait coverage was assessed within families as the number of species for which values were missing over the total number of species in each family. As families represented by very few species might present higher percentages of missing values, reflecting family size rather than randomness in sampling, I contrasted trait coverage plots against a plot showing how much each family contributed to the total number of species (number of species in each family over total number of species in the tree).

Trait coverage was highly variable across classes and traits (see Results). Trait coverage for species figuring in the PREDICTS database only overall improved compared to trait coverage for the whole set of species, particularly for reptiles and amphibians (see SI). Nevertheless, no trait reached 100% coverage in any class. Obtaining trait estimates for all of PREDICTS species was important, as otherwise, each species for which trait values were missing would have to be dropped in further analyses. Moreover, within-class biases in availability of trait information appeared (see Results). Consequently, dropping missing-value species could skew trait distributions and generate biases in further analyses. As such, rather than dropping missing-value species, I aimed to fill coverage gaps by imputing missing trait values.

3.2.5 Imputing missing trait values

In order to achieve full coverage across classes, I imputed missing trait values. Diverse imputation methods have been developed and used in published articles. Penone et al (2014) assessed the performance of four different imputation approaches (K-nearest neighbour (kNN, Troyanskaya 2001), multivariate imputation by chained equations (mice, van Buuren 2009, 2011), random forest algorithms as implemented in R by missForest (Stekhoven, 2011) and phylogenetic imputations implemented with phylopars (Goolsby, 2016)). Their study showed that the kNN approach resulted in significantly higher imputation error rates than the three other approaches. Both missForest and phylopars were the best methods when phylogenetic information was included. Nevertheless, phylopars was much slower than missForest, and could only handle continuous traits. missForest was faster and could deal with mixed type data. Without phylogenetic information, mice was found to be the best method, with fast imputations of mixed-type data. Of all these methods, missForest was the only one that did not make assumptions about data distribution (being a non-parametric approach), or that did not require a prior knowledge of some tuning parameters. As such, missForest appeared to be a robust option for missing data imputation. To further assess whether to use random forests rather than multivariate chained equations, I estimated the amount of phylogenetic signal in traits. Strong phylogenetic signal in traits would indicate than missForest could perform better than mice.

Assessing phylogenetic signal in traits

Measuring phylogenetic signal in continuous traits with Pagel's λ. Phylogenetic signal is a measure of the tendency of closely related species to resemble each other more than less related species. Diverse statistics have been developed to estimate phylogenetic signal, most of them applying to continuous traits (Munkemuller 2012). Here, I used Pagel's λ, estimated with the R function phylosig (phytools package), to assess the amount of phylogenetic signal in continuous traits. Pagel's λ is a scaling component that measures the transformation that should be applied to the phylogenetic tree for a trait to have evolved under a pure Brownian motion model of evolution. Under a Brownian motion model of evolution, changes in trait values happen at random along the branches and trait variance is proportional to evolutionary time. λ is then close to zero: the trait covariance matrix is scaled down and the tree loses its internal structure. When λ equals one, both the phylogeny and the trait covariance matrix remain unchanged and the structure of the tree explains trait evolution. As such, λ values close to one indicate that trait values are more similar in closer related species.

Using Pagel's λ, I assessed the strength of the phylogenetic signal. The phylosig function (phytools) also allows to test for signal significance (comparing the estimated λ to the null expectation of λ with a log-likelihood ratio test).

Measuring phylogenetic signal in categorical traits with δ (Borges et al, 2018). Very few methods have been developed to measure and test phylogenetic signal in categorical traits.

Fritz and Purvis (2010) introduced the D-statistic; nevertheless, D is based on a discretisation of categorical traits, which reduces them to binary variables. Borges et al (2018) introduced a new statistic, called δ, to measure phylogenetic signal in categorical traits of all types. Their approach uses Bayesian inferences to reconstruct trait evolution, that is, to infer trait values in ancestral nodes of the phylogeny. The underlying idea is that the better the phylogeny explains trait evolution, the lower the uncertainty is in ancestral state inferences. As such, δ relies on the quantification of the uncertainty associated with the reconstruction of ancestral states. δ can take any positive number, with higher values indicating stronger signal. To test for the significance of the signal, the authors propose to compare the estimated value of δ with the null expectation of δ.

I estimated phylogenetic signal in categorical traits with the δ statistic; implementation used the R code provided by Borges et al. To test for the significance of the signal, I generated null distributions of δ for each trait by randomising trait vectors 50 times (simulating Brownian motion model of trait evolution), and calculating δ for each randomised vector. I then calculated the median of simulated δ values as well as 95% confidence intervals. I tested whether the null-medians were significantly lower than the observed value of δ using one-sided Wilcoxon rank sum tests. Note that the function developed by Borges et al cannot be implemented if phylogenetic trees contain branches of length 0. As both original and corrected phylogenies contained 0-length branches, I added a very small number to these (10^{-10}) to remedy to this issue and to test for phylogenetic signal.

Significant phylogenetic signal in all traits All traits showed significant phylogenetic signal (Table 3.3 and SI for p-values of statistical tests), although the strength of the signal was variable across classes and traits. Overall, modifying the original phylogenies by correcting for taxonomy and by attaching species to the root of their genus did not, qualitatively, have a strong impact on the signal (Figures 3.4 and 3.5), although differences were bigger in reptiles and amphibians, where more than 80% of missing species were added to phylogenetic trees. In mammals and birds, phylogenetic signals remained similar. On the other hand,the stronger effects were observed for reptilian body mass, where adding species to the tree lowered the strength of the signal, and for amphibian trophic level, were the opposite was observed.

Phylogenetic signals in categorical traits were all highly significant (Figure 3.5; p-values for Wilcoxon signed rank test: see SI). The strength of the signal differed across classes and traits, with diel activity, trophic level and primary diet showing particularly strong signal in mammals and birds. Reptiles also showed strong signals for diel activity and trophic level. For amphibians, the results were more even across traits, and still highly significant. Overall, the signal for habitat specialisation was less strong.

Most mammalian continuous traits had very strong phylogenetic signal ($\lambda \geq 0.9$), except habitat breadth ($\lambda \approx 0.7$). In birds, both habitat and diet breadth showed weaker signal, but other continuous traits were highly conserved across closely related species . For amphibians and reptiles, signal strength was much more variable, which may be due to poorer initial trait coverage across phylogenetic tips (see Results). Nevertheless, body length showed high signal in both these classes ($\lambda \geq 0.9$).

Table 3.3: Phylogenetic signal in continuous and categorical traits and in range size. **BM**: body mass; **L**: longevity; **LCS**: litter/clutch size; **HB**: habitat breadth; **DB**: diet breadth; **GL**: generation length; **BL**: body length; **SM**: sexual maturity; **RS**: range size; **TL**: trophic level; **PD**: primary diet; **DA**: diel activity; **Sp**: specialisation. The phylogenetic signal in continuous traits was calculated with Pagel's λ. For categorical traits, the δ metric developed by Borges et al (2018) was used. A star indicates a significant signal (significant p-values scores for the log-likelihood ratio test in the case of λ; and significant difference from the simulated null distribution of δ for categorical traits, see SI). 'na' are introduced for traits that were not considered in a class but may have been used in another as a predictor in missing values imputations. All traits showed significant phylogenetic signal, with signals for BM, L, LCS, and GL being particularly strong in mammals and birds (above 0.9). Here all calculations were conducted with the corrected phylogenies, after species additions at the root of their genus. See SI for phylogenetic signals computed with the original phylogenies.

Class	**Continuous target traits, additional predictors and range size: λ**									**Categorical traits: δ**			
	BM	**L**	**LCS**	**HB**	**DB**	**GL**	**BL**	**SM**	**RS**	**TL**	**PD**	**DA**	**Sp**
Mammals	1.0*	0.97*	0.96*	0.70*	0.99*	0.99*	1.0*	na	0.99*	17*	50*	19*	1.4*
Birds	1.0*	0.87*	0.92*	0.50*	0.49*	0.99*	na	na	0.67*	10*	18*	$28{\cdot}10^3$*	1.6*
Reptiles	0.36*	0.81*	0.72*	0.39*	na	na	0.98*	0.87*	0.41*	4.3*	na	7.1*	1.5*
Amphibians	0.80*	0.33*	0.63*	0.77*	0.71*	na	0.96*	na	0.54*	18*	3.7*	2.9*	3.6*

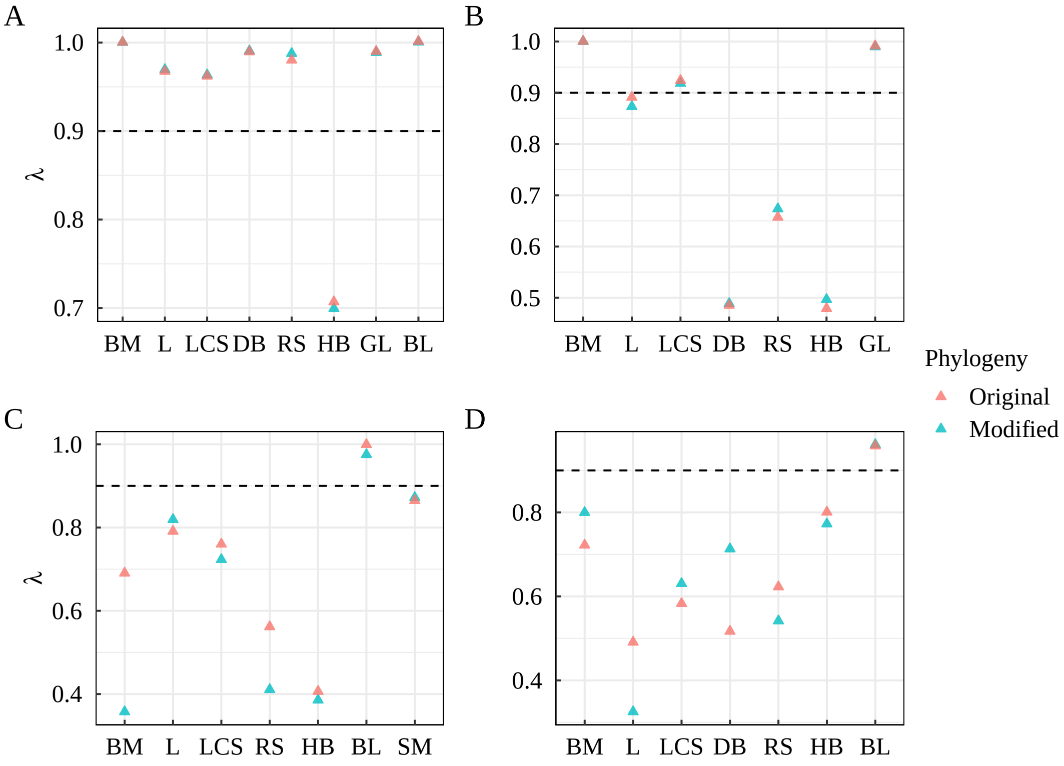

Figure 3.4: Phylogenetic signal in continuous traits (Pagel's λ) estimated with both original phylogenies and modified phylogenies. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Overall, altering the phylogenies by correcting for taxonomy and by increasing species representation did not have an important effect on λ.

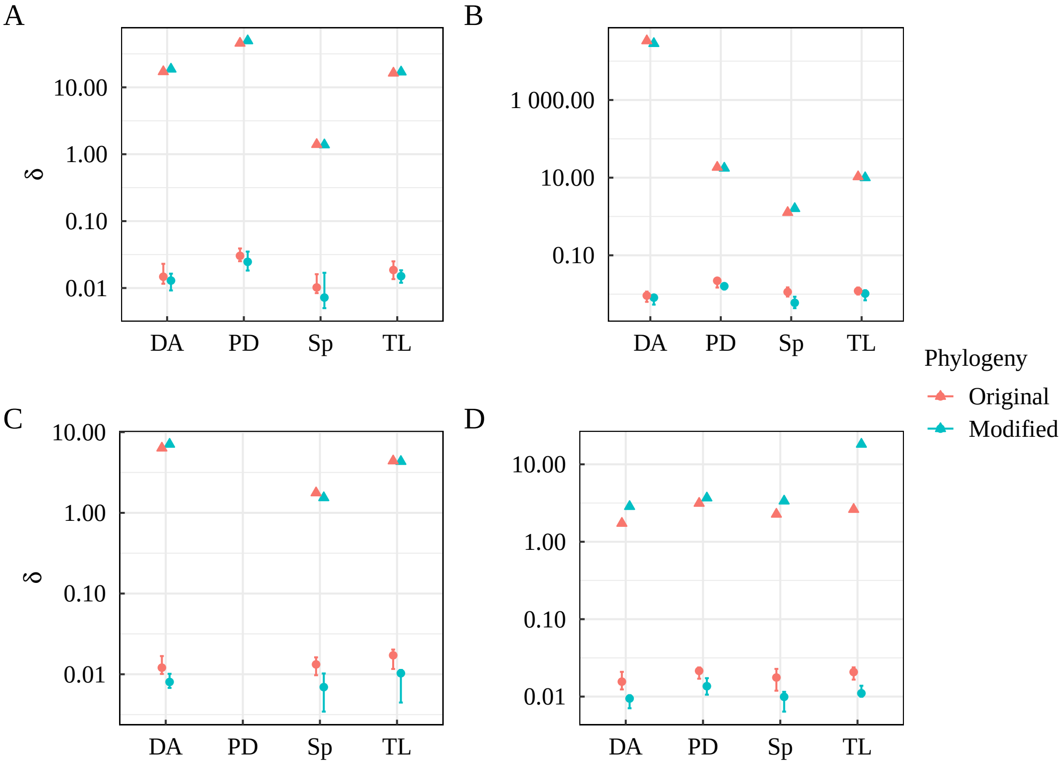

Figure 3.5: Phylogenetic signal in categorical traits (δ) estimated with both original phylogenies and modified phylogenies. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Triangle-shaped points represent the estimated phylogenetic signal in each trait; round-shaped points represent the median null expectation of the phylogenetic signal ($\pm 95\% CI$). Alterations of the phylogenies did not strongly impact δ.

Missing trait values: imputation implementation

Despite much variation in trait coverage across classes (see Results), results indicated strong phylogenetic signal in many categorical and continuous traits (Table 3.3). I hence imputed missing trait values using random forest algorithms, implemented by missForest. As stated above, missForest was shown by Penone et al (2014) to be the best method when including phylogenetic information for mixed-type variable imputations. Phylogenetic relationships were included as additional predictors in the form of phylogenetic eigenvectors (Diniz-Filho 2012), extracted from the phylogenies using the PVR package (Santos 2018). Phylogenetic eigenvectors are computed from a phylogenetic distance matrix, and calculated using principal coordinate analysis methods. They summarise the relationships among species, and the first set of eigenvectors reflect larger distances, capturing divergences closer to the root (Diniz-Filho 2012). Penone et al (2014) showed that including the first 10 eigenvectors minimised the imputation error when imputing missing trait values with missForest. As such, I included the first 10 eigenvectors as additional predictors of missing trait values.

As not all species were represented in the phylogenies (Figure 3.3), I also added taxonomic orders as an extra predictor variable in the random forest algorithm. All traits in Table 3.1 were included in the imputations (except for primary diet and diet breadth in reptiles). Tuning parameters of missForest were set to 10 maximum iterations (if the stopping criterion was not met beforehand, see below) and to 100 trees grown in each forests. To further examine imputation robustness and error, I imputed eight datasets in parallel (eight imputed trait datasets for each class).

Imputation error and robustness

Out-of-bag imputation error. To assess imputation accuracy, I used the ‘out-of-bag’ error (OOB error) returned by the missForest function. The missForest algorithm proceeds iteratively, training random forests on observed values first, then predicting missing values over several iterations. When the difference between the last imputed dataset and the previous imputed dataset increases, the stopping criterion is met. The penultimate imputed dataset is then returned. For continuous variables, this difference, Δ_{cont}, is defined as:

$$\Delta_{cont} = \frac{\sum_{j \in N} \left(X^{i,l} - X^{i,p}\right)^2}{\sum_{j \in N} \left(X^{i,l}\right)^2}, \tag{3.1}$$

where j is a continuous trait among N traits, $X^{i,l}$ is the last imputed dataset and $X^{i,p}$ is the penultimate imputed dataset. Δ_{cont} is a measure of the aggregated distance between two successive imputations across all continuous traits. For categorical variables, the difference Δ_{cat} is:

$$\Delta_{cat} = \frac{\sum_{k \in F} \sum_{j} J_{X^{i,l} \neq X^{i,p}}}{n(NA)}, \tag{3.2}$$

where k is a categorical trait among F categorical traits, $n(NA)$ is a the number of missing values for k and J is the j^{th} imputed values for which the consecutive imputations predicted contradicting results. In other words, Δ_{cat} measures the proportion of values that were found to be different between two successive imputations (see Stekhoven (2011) for more details).

When the stopping criterion has been met, out-of-bag imputation errors can be estimated. Out-of-bag errors refer to errors estimated from sub-samples of the data (bootstrap datasets, on which models are trained). Out-of-bag errors are estimated from these bootstrap datasets and as such differ from ‘true’ imputation errors, which require previous knowledge of the full dataset. The true root-mean square error (root-MSE) for continuous traits is defined as:

$$\sqrt{\frac{mean\left(\left(X_t - X_i\right)^2\right)}{var\left(X_t\right)}}, \tag{3.3}$$

where X_t is a vector of the complete trait values and X_i a vector of the imputed trait values (Stekhoven 2011). In case of an out-of-bag error, when the complete trait data is not provided, the MSE is calculated from the bootstrap datasets. For categorical traits, the out-of-bag PFC is calculated as the PFC (Δ_{cat}, Equation 3.2), using the bootstrap sub-samples. Breiman (1996) showed that OOB estimates provide accurate estimations of the true imputation error.

I retrieved OOB imputation errors (root-MSE and PFC) across eight imputed trait datasets in each class. I plotted the mean root-MSE and the mean PFC across the imputed datasets, as well as the range in errors (maximum error values and minimum errors values across all imputed datasets).

Imputation congruence. To further assess whether imputations were robust, I investigated whether similar values were imputed across the eight datasets in each class, or in other words, whether results were congruent across the imputed datasets. My expectation was that, for a trait, values imputed independently in different rounds should be nearly identical if imputations were robust. As such, for a continuous trait, pairwise correlations coefficients should be high across the eight datasets (correlation coefficients for the same trait imputed in pairwise independent rounds, see Table 3.4). For categorical traits, the random forest should predict the same values across the eight datasets.

Table 3.4: Conceptual design for examining imputation congruence for continuous traits. For one trait, pairwise correlation coefficients across eight independent imputation rounds are expected to be high if imputation are robust. To assess imputation congruence across eight imputed datasets, pairwise correlation coefficients were averaged (and the spread assessed using the range).

	Imputed 1	**Imputed 2**	**Imputed n**
Imputed 1	1	-	-
Imputed 2	corr(1,2)	1	-
Imputed n	corr(n,1)	corr(n,2)	1

For continuous traits, I assessed imputation congruence across the eight imputed datasets by averaging pairwise Pearson's correlation coefficients and plotting the mean (and range) for each trait. For categorical traits, I assessed congruence by assessing the percentage of species for which all eight imputed values were similar.

3.3 Results

3.3.1 Outputs

I collected and imputed data for 10 traits across 11637 avian species, 5502 mammalian species, 10334 reptilian species and 6904 amphibian species. Datasets recording species accepted and synonymic binomial names are available alongside the trait data.

3.3.2 Biases in the availability of trait information: non randomness in coverage and completeness and patterns in missing trait values

Increases in coverage and completeness due to taxonomic corrections.

Figure 3.6 shows the trait coverage within each class and for each trait, before and after correcting for taxonomy. Figure 3.7 shows the distribution of trait completeness before and after taxonomic corrections, as well as the median trait completeness for each class. Across all classes, correcting for taxonomy increased trait coverage (Figure 3.6). Nevertheless, the increase in coverage for reptiles was marginal, which may indicate that the procedure developed to extract and identify accepted names overall performed less well for reptilian species than for mammals, birds and amphibians. Similarly, correcting for taxonomy improved trait completeness in all classes (Figure 3.7). Wilcoxon rank sum tests, testing the null hypothesis that uncorrected and corrected completeness distributions came from the same population, rejected this hypothesis across all classes (alternative hypothesis: uncorrected medians were lower than corrected medians; mammals: p-value=$1.2 \cdot 10^{-9}$; birds: p-value$<2.2 \cdot 10^{-16}$; reptiles: p-value=0.025; amphibians: p-value$<2.2 \cdot 10^{-16}$). To conclude, correcting for taxonomy had a significant impact on trait completeness and increased coverage in most cases.

Among-class biases in the availability of trait information

Trait coverage. Trait coverage was highly variable across classes and traits. Trait coverage was initially good for most mammalian and avian traits, which had more than 50% coverage (Figure 3.6 A and B). Only longevity had a coverage lower than 50% for these classes, although generation length was above 80% in both cases. Conversely, trait coverage was overall much poorer for reptiles and amphibians (Figure 3.6 C and D). About two-thirds of amphibian and reptilian traits presented a coverage below 50%. Amphibians and reptiles appeared to be less sampled in all traits, except in body mass (reptiles) and in body length, range size and habitat variables (amphibians). As such, contrasting patterns of trait coverage appeared between, on the one hand, mammals and birds, and on the other hand, amphibians and reptiles. For species found in PREDICTS only, coverage increased disproportionally in reptiles and amphibians compared to the coverage for the full set of species (the figure for PREDICTS species only is available in the SI).

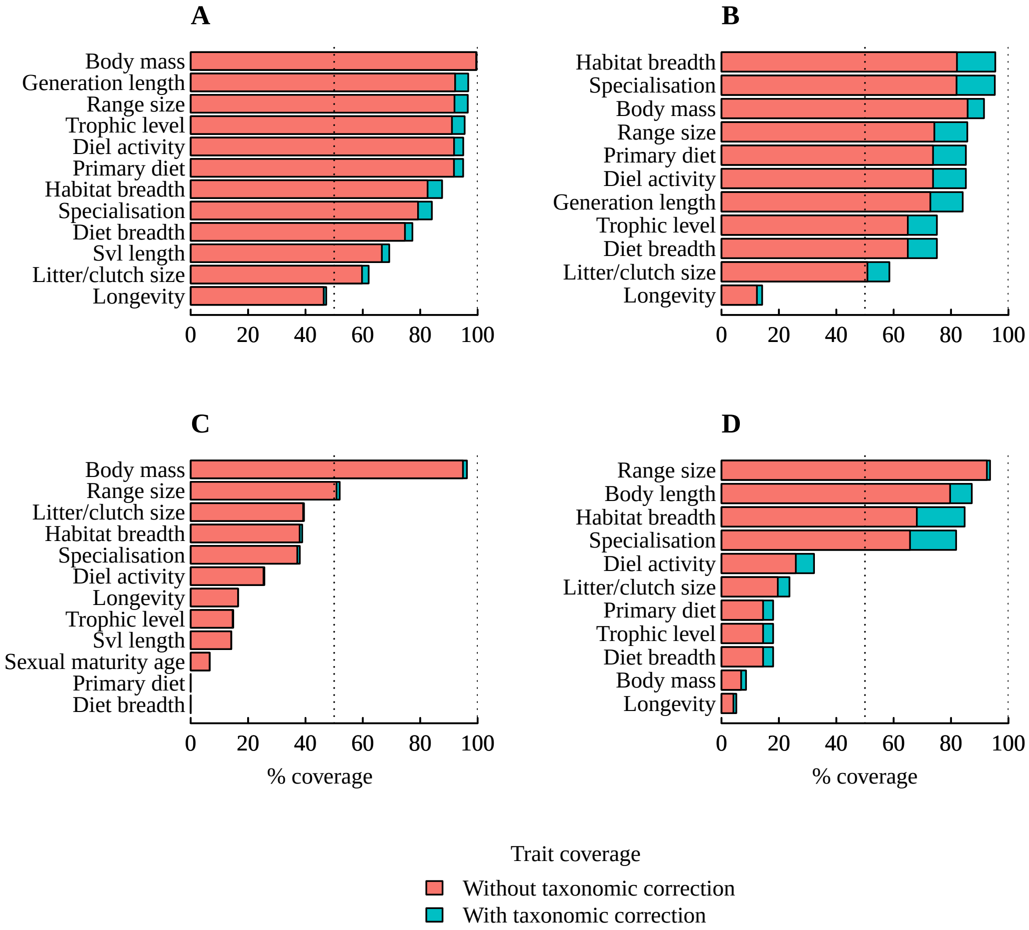

Figure 3.6: Trait coverage across all species before and after taxonomic correction. Here are shown all targeted traits as well as a few other traits used in imputations, as additional predictors (such as generation length for mammals and birds or body length for amphibians). **(A)** Mammals (5885 species before correction, 5502 and after correction); **(B)** birds (13554 species before correction, 11637 after correction); **(C)** reptiles (10722 species before correction, 10334 after correction) and **(D)**; coverage across amphibians (8643 species before correction, 6904 after correction). Trait coverage was calculated as the percentage of species for which trait information was available. Correcting for taxonomic synonymy improved coverage in most cases. For mammals and birds, all traits had an initial coverage of more than 50%, except longevity (but generation lengths were estimated for most species). On the other hand, trait coverage was poor (below 50%) for about two thirds of collected reptilian and amphibian traits. A clear contrast in trait information appeared between mammals and birds versus amphibians and reptiles, highlighting the existence of important taxonomic biases in data collection.

Trait completeness. Trait coverage revealed taxonomic biases, with higher resolution of trait information across mammals and birds. Trait completeness reflected similar biases. (Figure 3.7). The median completeness with taxonomic correction was high for mammals and birds (92% and 82%

respectively) but much lower for reptiles and amphibians (30% and 36% respectively). A pairwise Kruskall-Wallis rank sum test rejected the hypothesis that completeness distribution across classes originated from the same distribution (p-values$<2\cdot10^{-16}$ in all cases), showing that class had a significant effect on the availability of trait information.

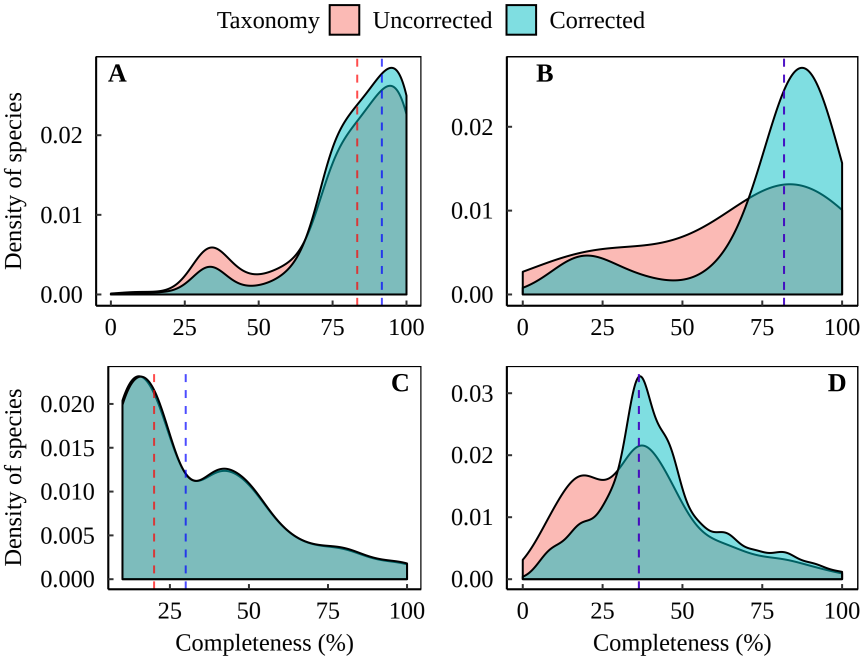

Figure 3.7: Distribution of completeness of trait information across species. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Completeness was calculated here for the same set of traits shown in Figure 3.6 (all predictor traits). Correcting for taxonomy affected completeness, significantly shifting the distributions to the right (alternative hypothesis, Wilcoxon rank sum tests: uncorrected medians were lower than corrected medians; mammals: p-value$=1.2\cdot10^{-9}$; birds: p-value$<2.2\cdot10^{-16}$; reptiles: p-value=0.025; amphibians: p-value$<2.2\cdot10^{-16}$). Class had a significant effect on median trait completeness (a pairwise Kruskall-Wallis rank sum test rejected the null hypothesis that completeness distributions across classes originated from the same distribution (p-values$<2\cdot10^{-16}$ in all cases)).

Non-randomness in trait information availability within classes: patterns of missing trait values with regards to phylogenies

Beyond cross-class biases in the availability of trait information, within-class patterns of missing values were revealed when plotting within-family median completeness and coverage against phylogenetic trees built at the family level.

Within-class patterns of trait completeness. Figure 3.8 shows within-family trait completeness for each class, colour-coded in the tree branches. For better visualisation, the trees are represented without tip labels. Figures providing tip labels are available in the SI (for each class, tip label information includes taxonomic order and family). As expected from the distribution of completeness values for mammals and birds, within-family completeness was high across most branches of the trees. In mammals, Chiropteras appeared to have lower median trait completeness than other orders (light blue cluster appearing in the middle of the tree, Figure 3.8A). In birds, no particular structure seemed to emerge in within-family median completeness (although the upper part of the phylogeny, corresponding to Procellariiformes, Charadriiformes, and Anseriformes appeared to be particularly well sampled, Figure 3.8B). In herptiles, nevertheless, clusters of similar completeness appeared at family levels. For reptiles, the lower part of the tree appeared to be particularly less well sampled than the above part of the tree (encompassing families such as Tropidophiidae, Lamprophiida or Typhlopidae: mostly, snakes; 3.8C). In amphibians, families in the Anura order showed both the best and worst median completeness (Figure 3.8D).

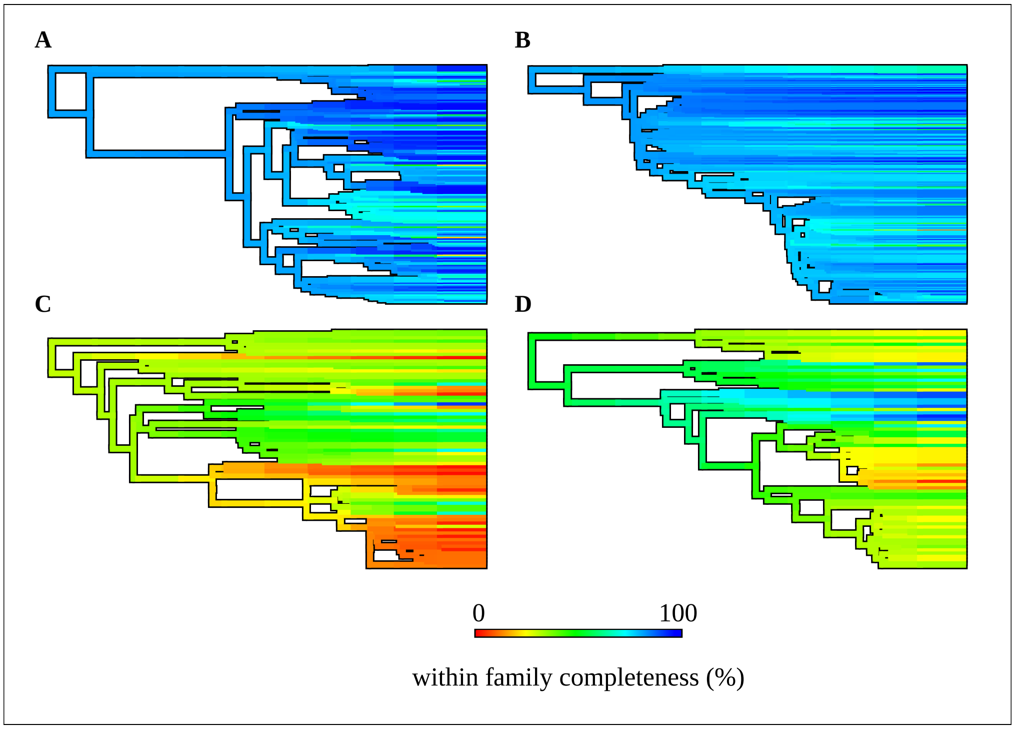

Figure 3.8: Median completeness across families. Tips labels are not shown here for better visualisation of the results; the same figures with tip labels are provided in the SI (zooming into the figure is necessary for mammals and birds); tip label information includes order and family. **(A)** Mammalian family tree; **(B)** avian family tree; **(C)** reptilian family tree and **(D)** amphibian family tree. Median trait completeness was calculated within families and colour-coded against tree branches. Family clusters of similar median trait completeness appear, particularly in reptile and amphibians.

Overall, these results showed that trait completeness was not random with regards to the phylogenetic relatedness of families. Closely related families seemed to share more similar median trait completeness than less closely related families. As such, the availability of trait information for a species may be dependent on its phylogenetic history; many other factors may interplay with species evolutionary history to explain these patterns.

Within-class patterns of trait coverage. Figures 3.9, 3.10, 3.11 and 3.12 show within-family median trait coverage. In each figure, the subplots are ordered from the trait showing highest overall coverage to the trait showing lowest overall coverage (as in Figure 3.6. In each figure, the last subplot represents the contribution of each family to the total number of species in the phylogeny.

Overall, these plots showed that for each trait, phylogenies seemed to be sampled non-randomly, with apparent clusters more visible as trait coverage decreased. Moreover, which families were better sampled was not

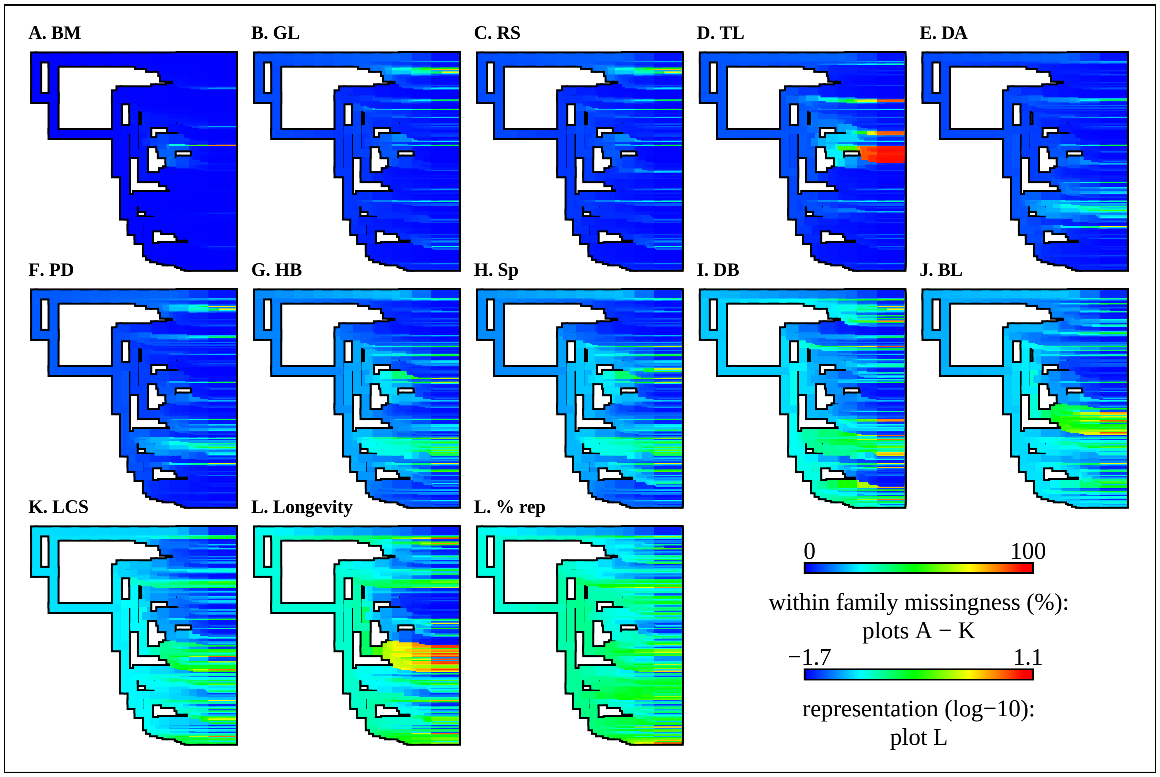

Figure 3.9: Within-family trait coverage in mammals.

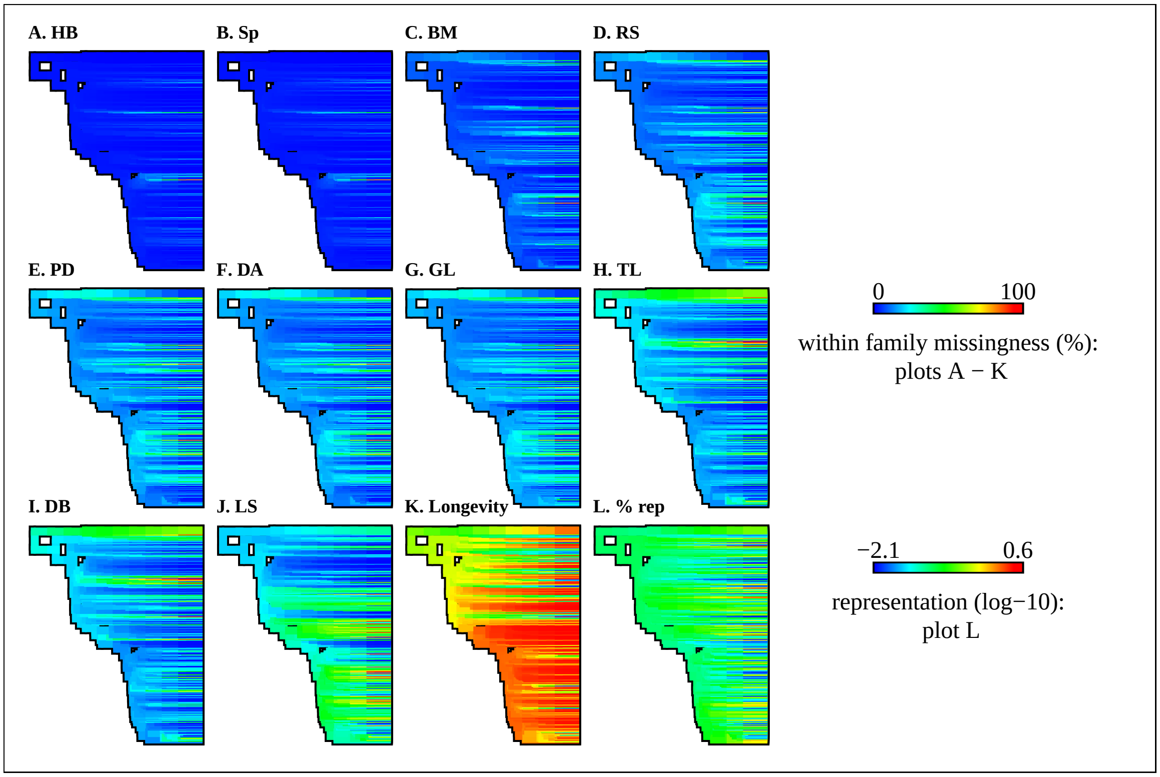

Figure 3.10: Within-family trait coverage in birds.

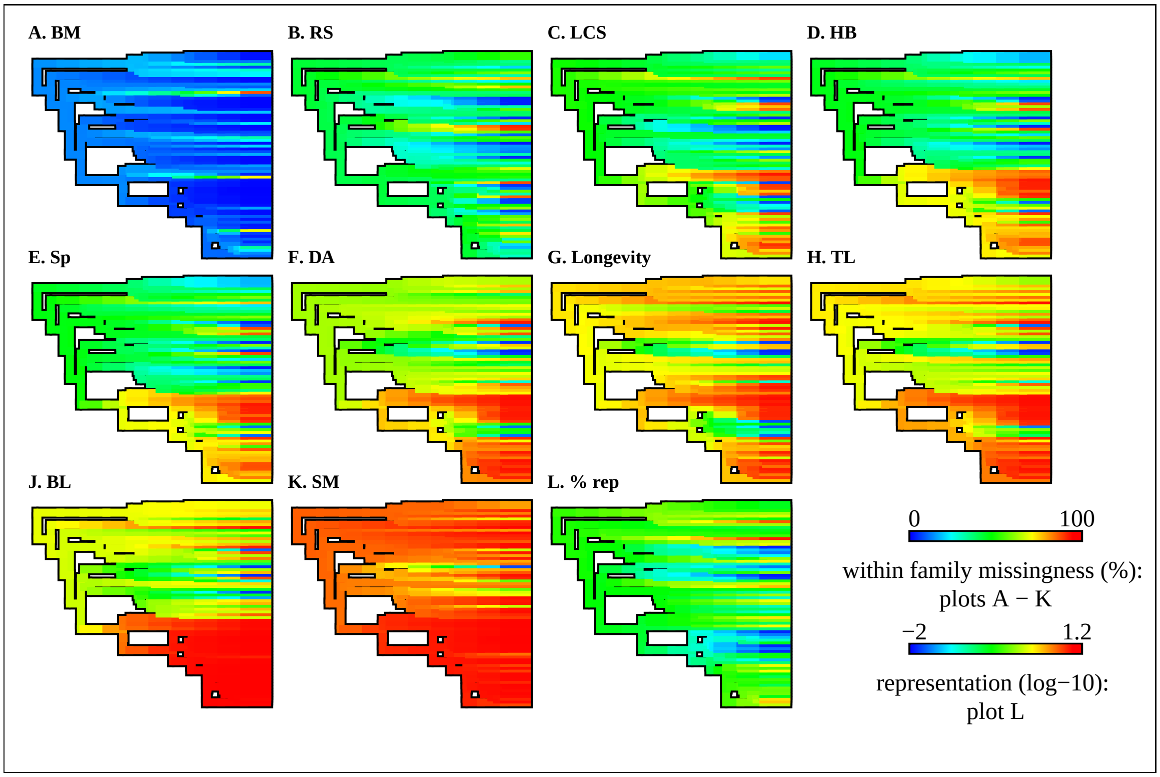

Figure 3.11: Within-family trait coverage in reptiles (squamates).

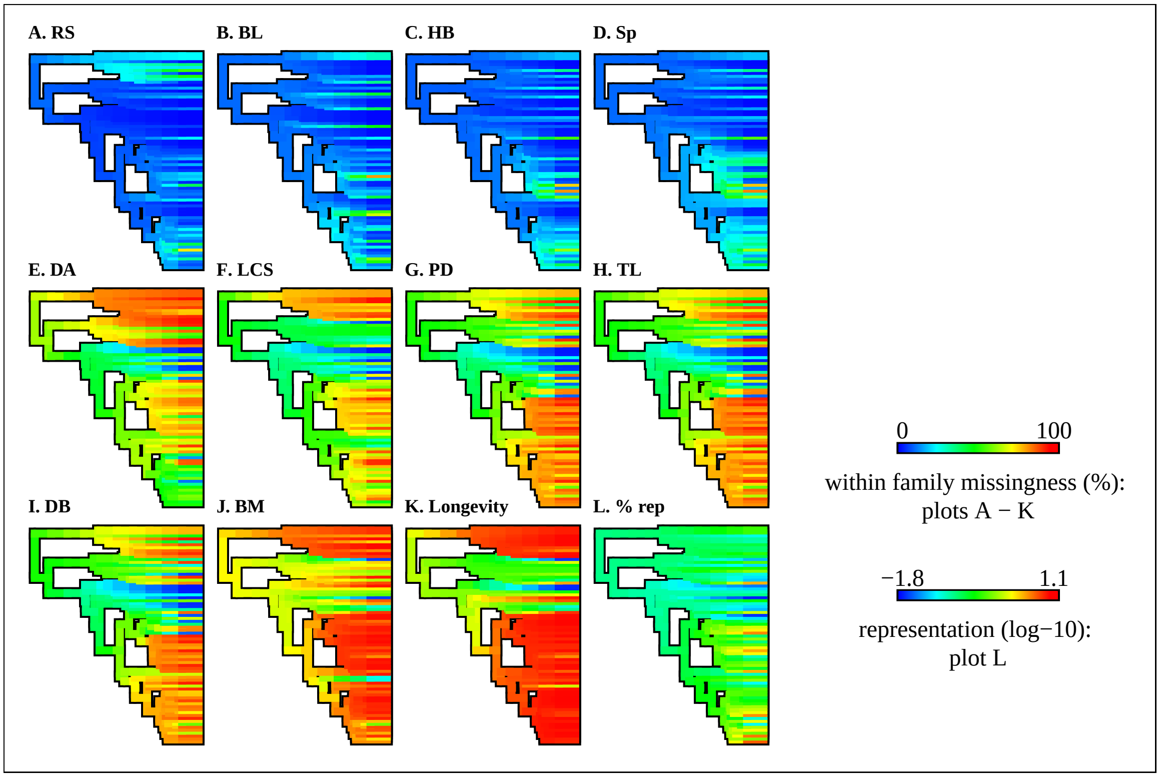

Figure 3.12: Within-family trait coverage in amphibians.

3.3.3 Imputation performance and robustness

Out-of-bag imputation errors

Figure 3.13A shows out-of-bag root-mean-squared errors for each continuous traits (shown here for one randomly selected imputed dataset). Figure 3.13B is the out-of-bag proportion of falsely classified values for categorical for the same imputed dataset. Estimated prediction errors for categorical traits were low to moderate (all below XX %). For continuous traits, estimated errors could be large (e.g., mammalian body mass, amphibian clutch size or range sizes). Nevertheless, such large errors were driven by high trait values in the dataset. Figure 3.14 shows the distribution of trait values after imputations; large prediction errors are estimated where traits can attain high values.

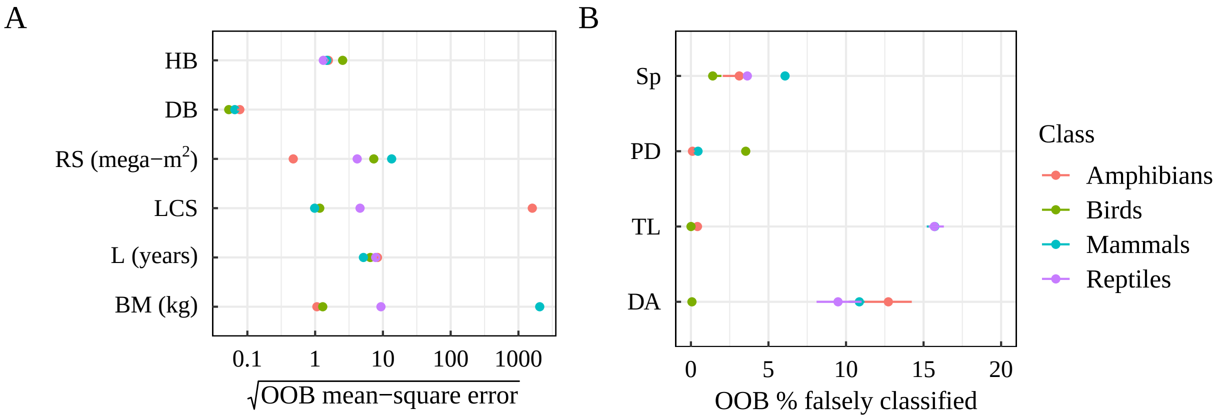

Figure 3.13: missForest out-of-bag root-mean-squared errors and proportion of falsely classified values. **(A)** Out-of-bag root-mean-square errors for continuous traits. **(B)** Out-of-bag proportion of falsely classified values.

Congruence of imputed values among eight imputed datasets

Figure 3.15A shows the range and mean of pairwise correlation coefficients obtained for each trait, across eight imputed datasets. Pairwise correlation coefficients were calculated for each trait, predicted in eight independent imputation rounds, so that high correlation values indicated more similar predictions for one trait across the eight datasets. Overall, imputation congruence was high for all continuous traits except habitat breadth. Imputation congruence was high across all classes for longevity (minimum mean correlation coefficient of 0.87 for reptiles), but more variable in other traits depending on the class. Figure 3.15B shows the proportion of species for which imputed values were similar across the eight imputed datasets. At least 50% of all species had similar predicted values across all imputed traits. Imputation congruence was high for trophic level (above 86% in all classes), and more variable in other traits depending on the class.

Mammals had the best imputation congruence scores in both continuous and categorical traits (minimum mean correlation coefficient of 0.85 for continuous traits and minimum percentage of agreement of 85% for categorical traits). Imputation congruence for birds was also very good, though scores were slightly lower for diet related variables (diet breadth and primary diet). For amphibians and reptiles, mean correlation coefficients were all above 0.60, except for habitat breadth. For amphibians in particular, imputation congruence on habitat breadth was poor. Overall, imputed results for amphibians were less congruent than for reptiles, birds and mammals.

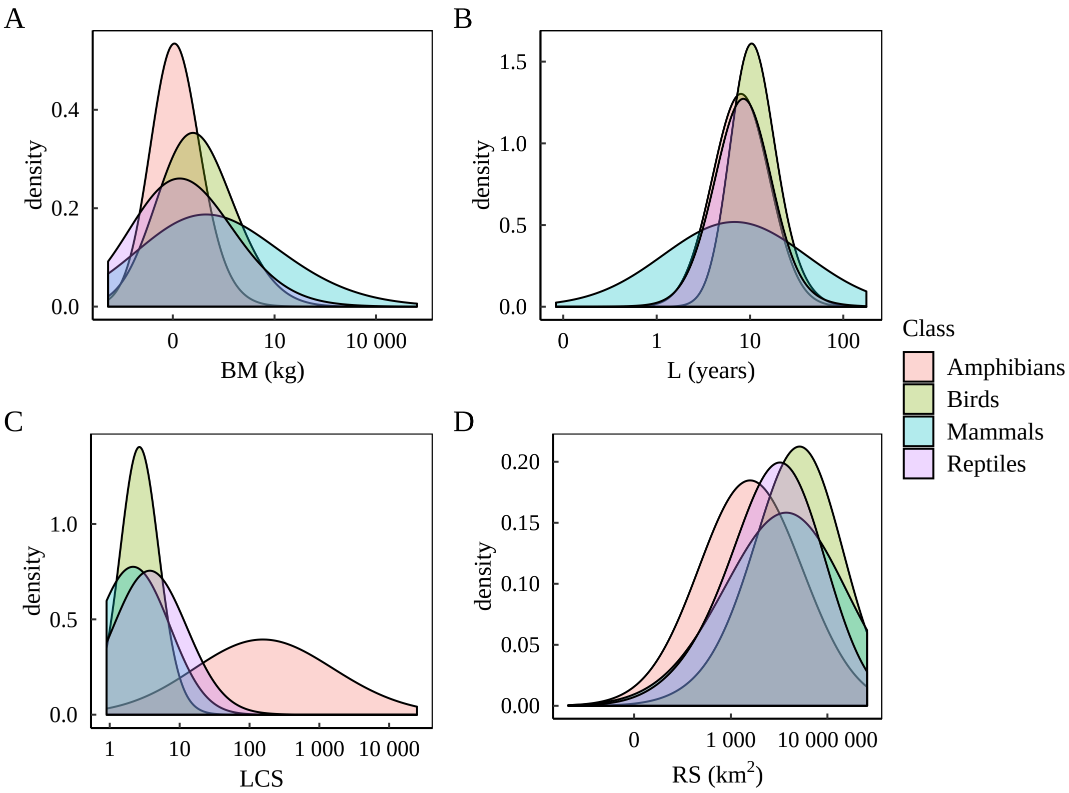

Figure 3.14: Distribution of trait values after imputation for body mass, longevity, litter/clutch size and distribution of range sizes.

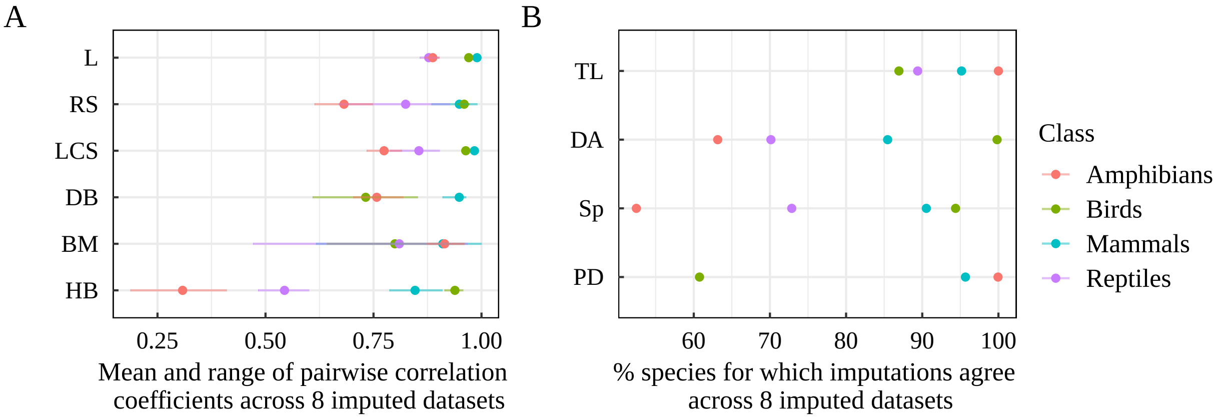

Figure 3.15: Imputation congruence across eight imputed datasets.

3.4 Discussion

In this work, I compiled and imputed data on 10 traits across x mammalian, x reptilian, x avian and x amphibian species. Traits related to species morphological characteristics (body mass), to their life-history (litter/clutch size, longevity, diel activity,), to their habitat preferences (habitat breadth, specialisation), and to their diet (trophic level; for mammals, birds and amphibians only, primary diet and diet breadth were also collated). To my knowledge, there is yet no published or freely available trait database encompassing all terrestrial vertebrates. As such, this work could constitute one of the first attempts to collate extensive trait information across all terrestrial vertebrates, enabled by all past and recent efforts to release trait information in the public domain.

Further developments could include enhancing the existing data to improve initial trait coverage. Alternatively, if novel primary sources were released, new variables could be added; even though the traits included in this work already encompass most of the ecological traits available in the literature across vertebrate classes, one notable omission was species mobility. Species abilites to both move within their habitats (home range) and to disperse and colonise new areas is likely to have a major impact on their aptitudes to cope with anthropogenic changes. Nevertheless, traits relating to mobility in amphibians or reptiles were unavailable. Relating to species movement, the only readily available variable that could have been added was volancy, assuming that most amphibians and reptiles were non-volant. Other information that could further enhance the dataset include reptilian diet, foraging strata and terrestriality (species habitat preferences along a vertical gradient: e.g. above versus below ground prefences). Note that the current compiled and imputed dataset contains fossil species, as some of the primary sources provided estimates for these. Such species could be filtered out in the future.

This works highlighted several frequent issues met when working with a large number of species or when working with dasasets from different origins. For categorical variables, the levels of the least resolved dataset had to be adopted across all classes, even though more detailed information was avaible in another class. Indeed, common denominators had to be found, at the expanse of highly resolved data. One example was diel activity time, that I had to constrain to two categories (nocturnal or non-nocturnal). The main reason for this was that most primary sources focused on one class only, while only a few encompassed more than two classes. Similarly, I did not compile any metric reflecting intra-specific variability in continuous traits. Intra-specific variability has been shown to have important effects on ecological systems, and a growing body of literature encourages trait-based research to include intraspecific variability in research studies. Here, metrics reflecting intraspecific variability were excluded due to both the scale of the data compilation and the lack of estimates across classes.

One major issue in this work was the taxonomic ‘pseudoreplication’ of species due to the presence of similar species under diverse names, and other taxonomic errors. To a lesser extent, where older taxonomic classification systems were used, inaccurate order or family information was provided in the primary sources. Taxonomic errors and replication of names are a major issue in ecology. This issue, which has been termed ‘taxonomic inflation’, is difficult to tackle at large scales (Isaac et al 2004). The lack of a comprehensive and universal database for species names complicates species identification, as well as unresolved taxonomy and taxonomic revisions (for example, one species being split into two subspecies, or the opposite). Taxonomic synonymy and taxonomic errors in general have been found to be a severe issue in diversity studies; for example, they can impediment accurate estimations of species number (Cardoso 2017). In this work, I showed that taxonomic synonymy artificially increased the amount of missing trait values by creating pseudoreplicates of the same species, also inflating the overall number of species and significantly lowering median trait completeness. The procedure that I developed to tackle taxonomic redundancy was itself highly dependent on the quality of the taxonomic information in the Red List and the ITIS. Overall, the procedure was not optimal, and a number of issues could be addressed to try and improved it. Nevertheless, it participated in reducing taxonomic mismatches and in increasing the number of matches across datasets. Some initiative, such as the Taxonomic Name Resolution Service for plants (http://tnrs.iplantcollaborative.org/) attempt to tackle taxonomic inflation by providing a free tool to standardise plant names and retrieve current accepted names. Up to five thousand species names can be submitted at once. The Global Information System provides a similar tool, which nevertheless does not systematically return species accepted names (but their status and a confidence level), and is as such less practical. Nonetheless, such databases are an invaluable source of information and should encourage researchers to try and standardise taxonomy.

In October 2018, Cooke et al released a dataset of six mammalian and avian trait. The methods they used to compile and impute trait data were very similar to the methods used in this work. The most notable divergences were the use of different imputation methods (multivariate chained equations) and the pre-selection of traits with more than 50% coverage in Cooke et al. Because very similar primary sources were used, I did not directly use their data in my work. Nevertheless, I compared the results of both data collection and imputation (see SI). Using non parametric random forest algorithms, as implemented in R by the missForest function, presented several advantages over other imputation methods. First, random forests could deal with mixed type variables, and estimate out-of-bag errors for each variable. Second, no underlying data distribution was assumed in the process, as it builds upon non-parametric processes. Third, missForest was computationally faster than other functions, which was an important criterion. Morever, missForest was found to be the best available method to use when imputing trait data with phylogenetic information. Nevertheless, even though missForest was found to be robust to missing values, no study has investigated imputations robustness with an amount of missing values as big as it was here. Moreover clustering of missing trait values in the phylogeny.

Despite these advantages, the robustness to imputations needs to be investigated further, for reasons that I detail below.

Conclusion I presented in this Chapter

The methods for compiling trait data may be revisited in the future, and the trait data is likely to be enhanced or to slightly change if imputed again.

Future work will build upon the trait data collated as described in this chapter. I illustrate the first use of this data with the next chapter, which investigates how land-use change impacts the functional diversity of vertebrate communities. In the last chapter, I detail some research questions that this data will allow to investigate in the future months on my PhD.

Completeness is likely to have an important effect on trait imputations, as it is a reflection of how many predictors have an estimate for a species.

4 Land-use change impacts on the functional diversity of vertebrate communities

4.1 Introduction

4.2 Methods

4.2.1 Trait selection for functional diversity metric calculations

In Chapter 1, I collected and imputed trait values for 10 traits across terrestrial vertebrates. I randomly selected one trait dataset among the eight imputed datasets (see Chapter 1) for all subsequent analyses. Continuous traits were transformed to improve normality. A log-10 transformation was applied to all continuous traits except habitat breadth, which was square-rooted. Traits were centred and scaled to zero-mean and unit-variance.

All traits were considered for inclusion into the calculation of functional diversity metrics, except variables relating to species diet, as these were unavailable for reptiles. Deciding which traits to include in functional diversity metrics was a critical step, as the metrics can be sensitive to the number of traits included (Mouillot et al 2014). On the one hand, not including enough variables may lead to missing important areas in the multidimensional trait space. On the other hand, collinearity among the traits could create a form of redundancy in the trait space, biasing estimated functional metrics. As such, a pre-analysis clean-up of the data was necessary to select traits to include in the calculations.

To first assess if collinearity could be a problem, I used a clustering approach for mixed-type data (factor analysis of mixed type data, or FAMD, ref) to represent the contribution of each variables to principal components. The graph shows that body mass - longevity might e correlated and that BM-LCS could be negatively correlated.

Next, I assessed Pearson's pairwise correlation coefficients among continuous traits, as high correlations can be an indicator of collinearity. Body mass and longevity were the highest correlated variables (Table XX, correlation: 0.51), nevertheless below or close to the threshold usually used for detecting potential collinearity (threshold of 0.7, Dormann 2012). Moreover, the determinant of the correlation matrix was 0.67. Values close to 0 indicate high degrees of collinearity in the dataset, while values close to one indicate low degrees of collinearity. A stepwise selection process using variance inflation factors (VIF) also failed to detect multicollinearity among continuous traits (with a threshold of 5; all VIF$\leq$1.4).

Consequently, all candidate traits were selected for inclusion in the functional metrics calculation. The selected traits were: body mass; longevity; litter/clutch size; trophic level; habitat breadth;

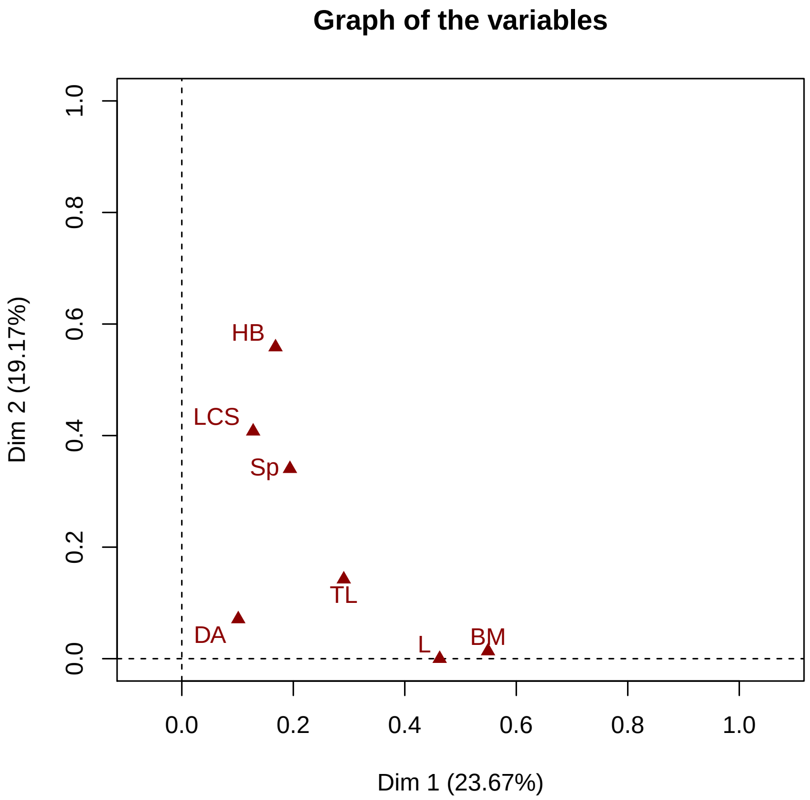

Figure 4.1:

degree of habitat specialisation; and diel activity.

4.2.2 Calculation of functional diversity metrics across PREDICTS vertebrate communities

Dendrogram-based functional richness

A Gower dissimilarity matrix was first computed from the trait dataset, using the gowdis R function (FD package). This distance matrix contained pairwise distances across all terrestrial vertebrates, based on their trait values. Gower distances allowed to include mixed type variables in the computation. In a second step, this dissimilarity matrix was clustered, to obtain a functional dendrogram, where species presenting similar functional characteristics were closer than more dissimilar species. I used the hclust function, which offers a range of clustering methods. As different clustering methods can have a strong influence on the output dendrogram, I selected the clustering method that best reflected the initial distances in the Gower matrix (correlation coefficient between cophenetic distances obtained from the cluster dendrograms and between the initial dissimilarities in the Gower matrix). The ‘average’ method (unweighted pair group method with arithmetic mean, UPGMA) presented the best correlation coefficient and was as such selected. The resulting cluster dendrogram was a functional dendrogram with 34377 tips, where each tip represented a species. Species position in the tree depended on their functional attributes: species that were functionally more similar were more closely related in the tree.

Finally, functional richness was calculated across all PREDICTS sites. At each site, vertebrate

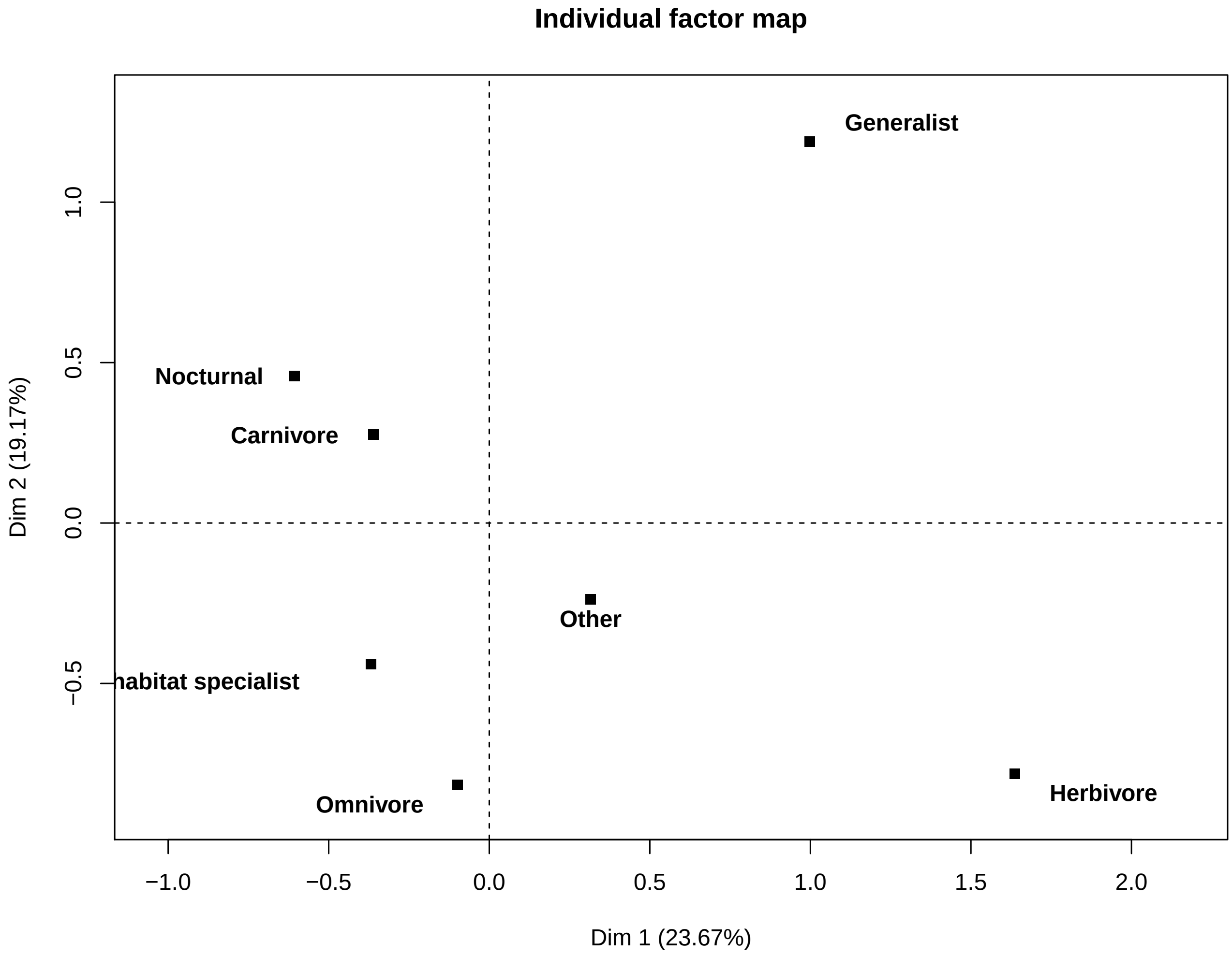

Figure 4.2:

community composition was assessed (species presence/absence), and the functional dendrogram was subsetted according to local community composition. For a site, functional richness was calculated as the sum of the branch length, from root to tip, for the local subset of the functional dendrogram (treedive).

194 studies to start with; presence absence matrices: 10 studies were excluded for abundance matrices: 30 studies were excluded.

4.2.3 Impact of land-use change on the functional diversity of vertebrate communities

4.3 Results

4.4 Discussion

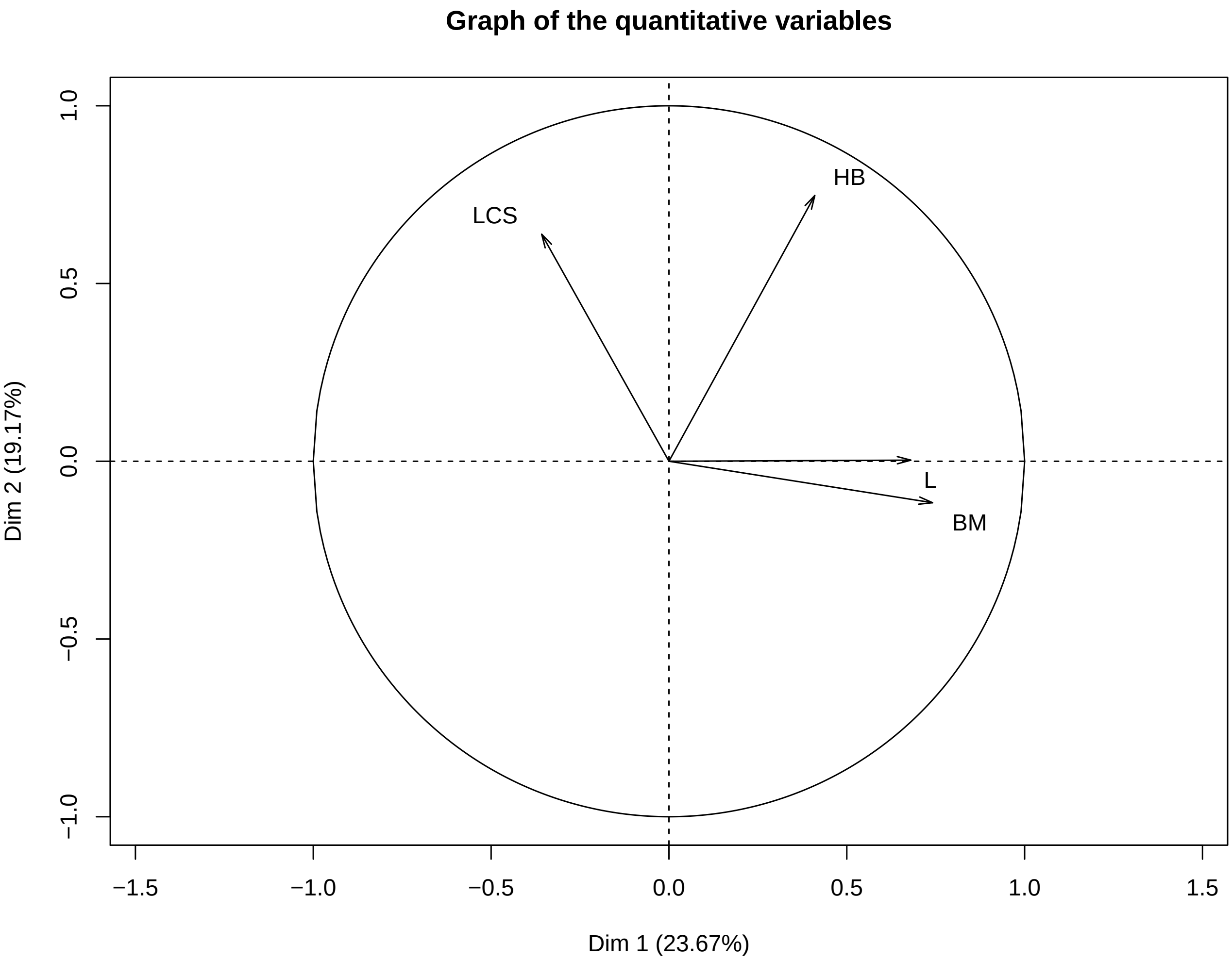
Graph of the quantitative variables
HB
LCS
L
BM
1.0
0.5
0.0
−0.5
−1.0
Dim 2 (19.17%)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Dim 1 (23.67%)

Figure 4.3:

5 | Outline and research questions for the next years

6 | Conclusion