LONDON'S GLOBAL UNIVERSITY

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 7, 2019

Contents

List of Tables

List of Figures

ITIS: Integrated Taxonomic Information System PREDICTS SI

Introduction

1 Literature review

2 Collecting and imputing ecological trait data across terrestrial vertebrates

2.1 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 (). Moreover, it is now accepted that ecosystem functioning is positively correlated with species functional diversity (Tilman). 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. 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. 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 traits 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 Redlist, Birdzone). Trait data 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).

In this chapter, I collected ecological trait information for terrestrial vertebrates from diverse primary sources. Trait selection was motivated by two main reasons: (1) traits should be of ecological interest and be related to response of effect processes; (2) trait values should be available for many species, across the four terrestrial vertebrate classes, allowing for cross-classes comparative analyses. Selected targeted 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 the data for the other classes.

The present chapter details the methodology I employed to collate trait information. 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. To achieve full coverage, I imputed missing trait values using random-forest algorithms. Here, I briefly examine imputation error and robustness.

In October 2018, Cooke et al released a database of six mammalian and avian traits, using similar primary sources. They collated and imputed missing trait values for body mass, litter/clutch size, volancy, diel activity, primary diet and habitat breadth. I did not use their collected data for two reasons: first, similar primary sources were used in both our compilations; second, they used different missing data imputations methods. I used this freely accessible data as an opportunity to compare the results of both our data collection and imputation process. This chapter also presents the results of this comparison.

2.2 Methods

2.2.1 Ecological trait data collection

Primary data sources.

I collated ecological trait data for terrestrial vertebrates from the sources figuring in Table 2.1. 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 calculated from distribution data, extracted from the IUCN Red List.

Table 2.1: Data sources for trait compilation. I here show were I extracted trait data from for each class. These individual sources may 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. Target traits are bolded; 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	✓		✓	✓		✓	✓		✓		
Meiri								✓		✓		
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.

Diet and diet breadth. Even though diet was not available from any primary source for reptiles, I compiled diet information for all other classes. Species diet was described in primary sources as a binary variable recording whether food items were known to be consumed by a species or not. I calculated diet breadth as the number of food items a species was recorded to ingest. In addition, species were pooled into 5 categories in one of the source (Elton birds) according to their primary diet (food items that constituted more than 50% of the species diet). I adopted the same system and pooled species into the 5 following primary diet categories: (1) seed or plant consumers; (2) fruit or nectar consumers; (3) invertebrate consumers; (4) vertebrate consumers (including scavengers); (5) omnivores. More details on diet compilation are provided in the SI.

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 sensitive to different weight choices (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.

2.2.2 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).

2.2.3 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 RedList 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 RedList 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.An extra column indicates whether the species is known only at the genus level.

Outputs. I generated a dataset of vertebrate species names found across datasets, recording whether names were accepted or synonymic. For each name, the accepted name and the synonyms were stored, as well as additional taxonomic information (order, family, genus).

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 2.1). The species presenting the highest degree of pseudoreplication was the East African mole rat (*Tachyoryctes splendens*), which was figuring under 12 different names across primary sources (Figure 2.1B).

Despite the automation efforts, taxonomic redundancy persisted 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; Avibase; AmphibiaWeb). 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 ITIS were not comprehensive taxonomic sources, and for clades with high degrees of pseudoreplication in names, such as reptiles or amphibians, neither the Red List or ITIS contained enough information. 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.

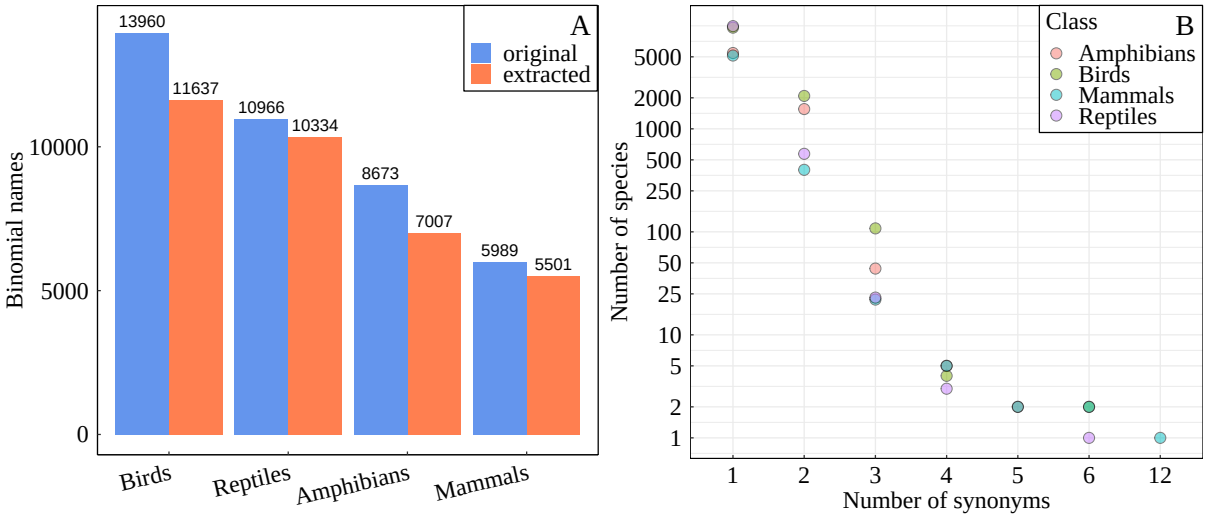

Figure 2.1: Difference in species number due to taxonomic correction (A) and distribution of number of synonyms across datasets (B). (A) shows the number of species across all primary sources (trait datasets, PREDICTS, phylogenies), before and after correcting for taxonomy. Replacing identified synonyms 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.

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 2.6% of mammalian, 1.5% of avian, 1% of amphibian and 1.5% 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 2.2). 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 example for each case of Figure 2.2).

Random species attachments. Some species in the trait datasets were not represented in the phylogenies. When applicable, and to increase representation, these species were attached to their genera in the trees at a random position. Only a small fraction of species that had no initial phylogenetic representation were randomly attached to their genera (Table 2.2).

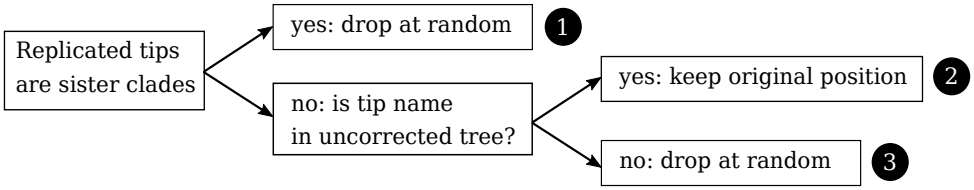

Figure 2.2: Procedure followed to drop replicated tips from phylogenies. Most of these were replicated twice. When replicated tips were sister clades, the tips to drop were chosen randomly, as it did not affect the 'true' phylogenetic position of the species (1). When replicated were not sister clades, 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).

Table 2.2: Species representation in phylogenetic trees (corrected taxonomy). The number of species randomly attached to their genera ranged from 94 (mammals) to 611 (reptiles). Finally, most avian and mammalian species were represented in the phylogenies, whereas more than half reptilian and amphibian species had no known phylogenetic position.

Class	Initially not in tree	Randomly attached	No final representation in tree
Amphibians	58% (4027 of 6888)	13% (510 of 4027)	**51%**
Birds	18% (2084 of 11637)	4.8% (100 of 2084)	**17%**
Mammals	7.4% (407 of 5502)	23% (94 of 407)	**5.7%**
Reptiles	62% (6391 of 10334)	9.6% (611 of 6391)	**56%**

Correcting for taxonomy in phylogenetic trees: conclusions. Overall, correcting for taxonomy in phylogenies improved species representation in the trees (Figure 2.3). Maximising the number of species represented in the phylogenies was important for further trait imputations and for estimating the amount of trait phylogenetic signal. For amphibian and reptilian species figuring in PREDICTS only, phylogenetic representation disproportionally increased (with a minimum representation of 76% for PREDICTS amphibians). Nevertheless, correcting phylogenetic tip labels generated replicates for a marginal number of tips, which then had to be dropped from the phylogeny.

Correcting for taxonomy increased trait coverage

Across all classes, correcting for taxonomy increased trait coverage, measured as the percentage of species for which trait information was available (Figure 2.4). Overall trait coverage was initially good for most traits for mammals and birds, with more than 50% coverage (Figure 2.4 A and B). Nevertheless, more than two-thirds of amphibian and reptilian traits had low coverage (below 50%, Figure 2.4 C and D). The trait coverage for the subset of species corresponding to PREDICTS vertebrates is provided in the SI. Coverage increased disproportionally for reptiles and amphibians for PREDICTS relevant species.

2.2.4 Biases in the completeness of trait information across classes

Trait coverage revealed taxonomic biases, with higher resolution of trait information across mammals and birds. For a species, the completeness of trait information, measured as the percentage of trait values that were not missing, showed the same taxonomic biases (Figure 2.5). The median completeness with taxonomic correction was 90% for mammals, 80% for birds, 25% for reptiles and

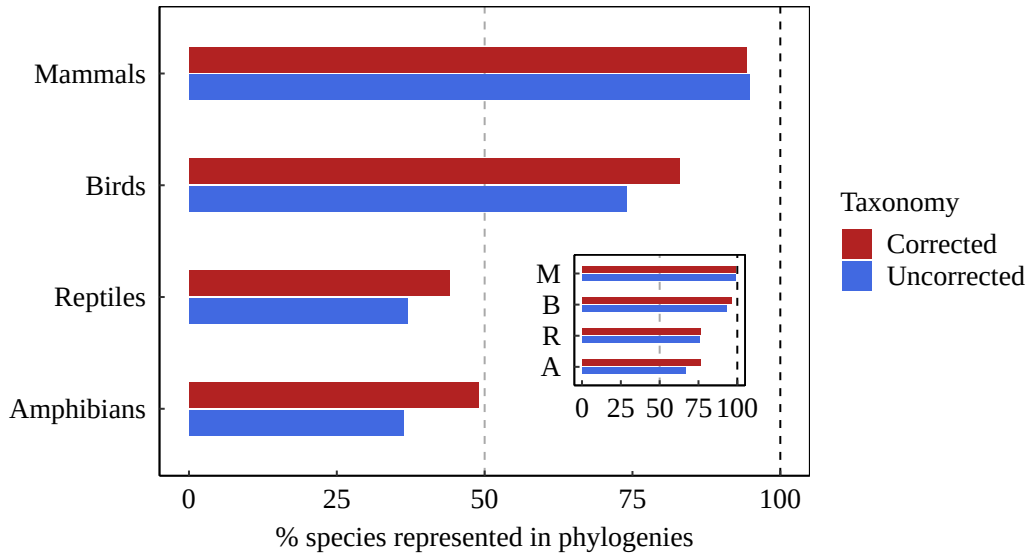

Figure 2.3: Percentage of species represented in the phylogenies for both corrected and uncorrected trait datasets. Overall, taxonomic correction increased species representation in phylogenetic trees. Representation for mammals and birds was high (after taxonomic correction: 83% of avian and 94% of mammalian species had a phylogenetic position). On the other hand, reptiles and amphibians were poorly represented (after taxonomic correction: only 44% of reptilian and 49% 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.

30% for amphibians.

These biases show that certain taxonomic groups are known much better than others. Within a class, trait information might be sampled non randomly, with certain families or order having a better resolution of trait information.

2.2.5 Imputing missing trait values

In order to achieve full trait 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 implemented with missForest (Stekhoven, 2011) and phylogenetic imputations implemented with phylopars (Goolsby, 2016)). Their study showed that the kNN approach resulted in significantly higher imputation errors 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 an interesting option for missing data imputation.

To further assess whether to use random forests rather than multivariate chained equations, I estimated the phylogenetic signal in traits. 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 in continuous traits (Munkemuller 2012). I used Pagel's

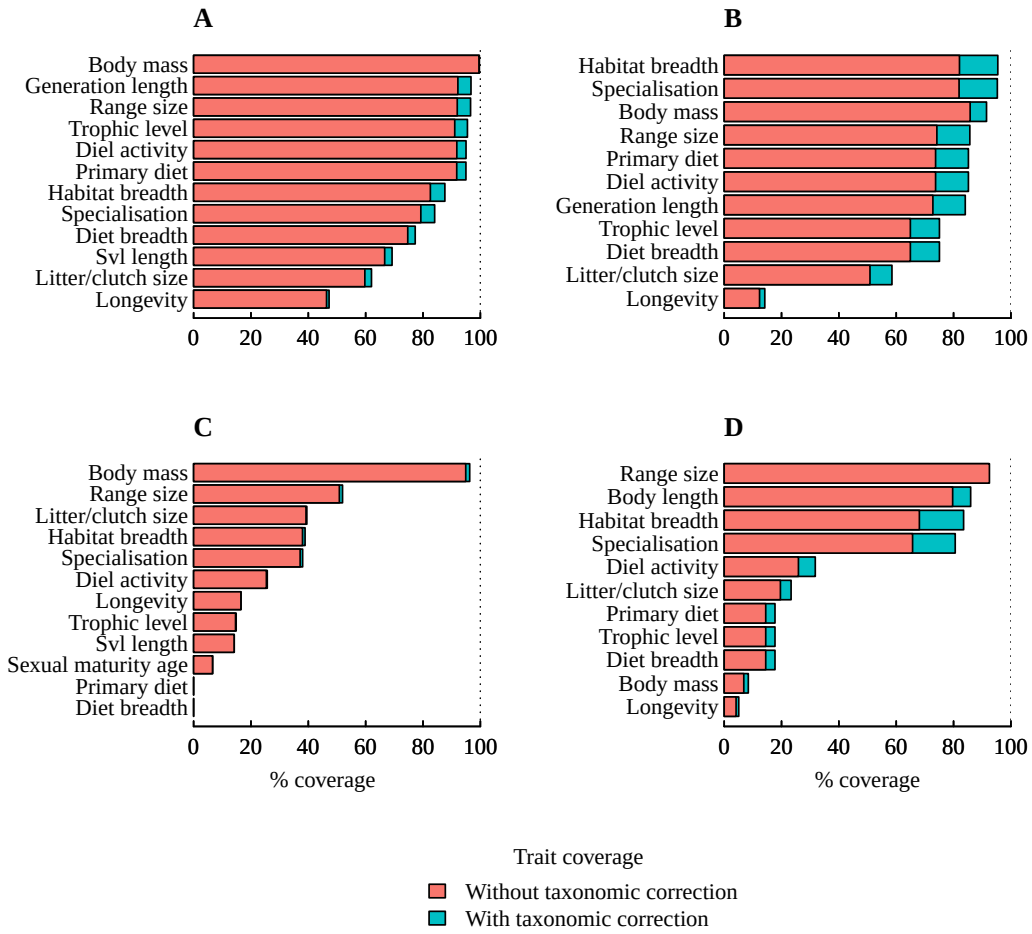

Figure 2.4: Trait coverage across all species before and after taxonomic correction. Here are shown all targeted traits as well as a few other traits used in imputations, as additional predictors (such as generation length for mammals and birds or body length for amphibians). **(A)** Trait coverage across mammals (5885 species before correction, 5502 and after correction); **(B)** coverage across birds (13554 species before correction, 11637 after correction); **(C)** coverage across reptiles (10722 species before correction, 10334 after correction) and **(D)**; coverage across amphibians (8643 species before correction, 7007 after correction). Trait coverage was calculated here 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 length were estimated for most species). On the other hand, trait coverage was poor (below 50%) for more than two thirds of collected reptilian and amphibian traits. A clear contrast in trait information appears between mammals and birds versus amphibians and reptiles, highlighting the existence of important taxonomic biases in data collection.

λ (function phylosig of the phytools package). Pagel's λ is a scaling component that measures the coefficient by which the trait covariance matrix should be weighted to fit a Brownian motion model of evolution. Indeed, under a Brownian motion model of evolution, the trait covariance matrix is expected to be influenced only by the phylogenetic history: changes in trait values happen at random and trait variance is proportional to evolutionary time. When other factors are at play, the observed covariance matrix is the expected covariance matrix transformed with the estimated λ. A value close to 1 indicates that trait values are more similar in closely related species than expected under a Brownian motion model of evolution.

Very few methods have been developed to measure and test phylogenetic signal in categorical

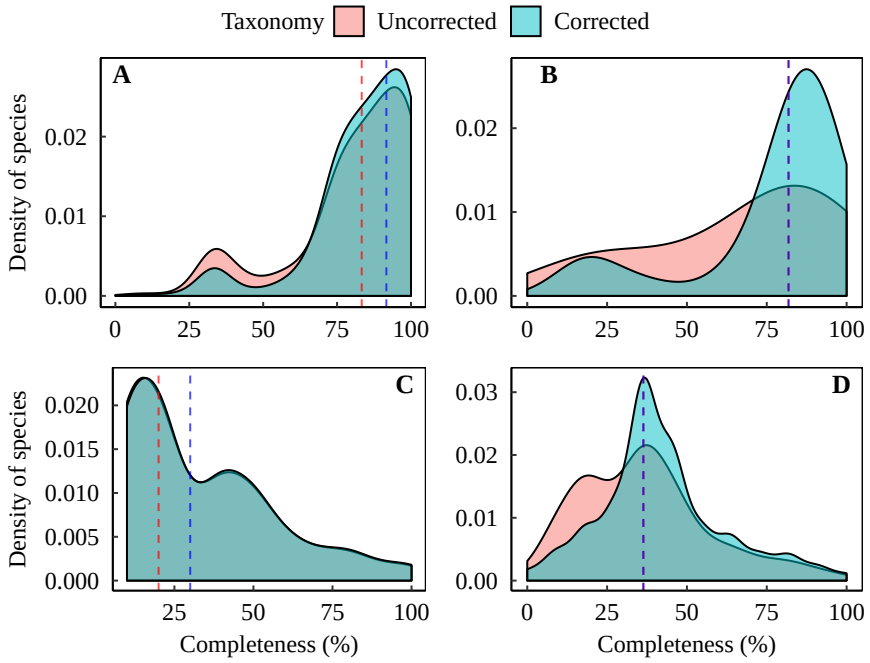

Figure 2.5: Completeness of trait information across species. (A) Mammals, **(B)** birds, **(C)** reptiles and **(D)** amphibians. Correcting for taxonomy did affected the median completeness (dashed lines) in mammals and reptiles, shifting the distributions to the right. Completeness is calculated here for the same set of traits shown in Figure 2.4.

traits. Fritz and Purvis (2010) introduced the D-statistic, which only applies to binary traits. Furthermore, D is based on a discretisation of the trait, which behaves as a continuous trait evolving under Brownian motion. Borges et al (2018) introduced a new statistic, δ, to measure phylogenetic signal in categorical traits. δ is based on Shannon entropy principles and uses Bayesian inferences for estimation. δ can take any positive number, with higher values indicating stronger signal. To test for the significance of the signal, the estimated value has to be compared to a null distribution of values. I generated null distributions of δ for each trait by simulating 100 random trait vectors and calculating δ for each. I then compared the simulated mean and 95% confidence intervals to the observed metric (see SI).

The results indicated strong phylogenetic signal in several categorical and continuous traits (Table 2.3). Despite differences in sample sizes, the signals were all significant (expect for amphibian body mass). Signals was nevertheless overall in mammals and birds, and this might be due to differences in sample sizes across classes.

As in most cases trait values were more similar among closer relatives than expected by chance, I 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, extracted from the phylogenies using the PVR package (Santos 2018). Penone et al (2014) also showed that including the first 10 eigenvectors minimised the imputation error. As not all species were represented in the phylogenies (Figure 2.3), phylogenetic eigenvectors presented some missing values. I added taxonomic orders as an additional predictor. All traits in Figure 2.4 were included in the imputations, except for primary diet and diet breadth in reptiles.

Table 2.3: Phylogenetic signal in 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. Phylogenetic signal in continuous traits was calculated using 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). All continuous traits had significant phylogenetic signal, except body mass in amphibians. This may have been the effect of the small sample size for this class, as body length showed a strong signal in that taxon. All categorical traits had significant phylogenetic signal. 'na' are traits that were not considered in a class but may have been used in another as a predictor in missing values imputations.

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	0.98*	0.99*	0.99*	0.90*	1.0*	1.0*	1.0*	na	0.99*	8.5*	NA	11*	1.2*
Birds	1.0*	0.99*	0.99*	0.92*	0.87*	1.0*	na	na	0.97*	11*	11*	460*	0.7*
Reptiles	0.39*	0.92*	0.99*	0.83*	na	na	0.99*	0.85*	0.65*	2.2*	na	2.9*	1.2*
Amphibians	<0.01	0.26*	0.86*	0.94*	0.49*	na	0.99*	na	0.78*	x	x	x	x

Imputation error and robustness

To assess imputation accuracy, I used the 'out-of-bag' error (OOB error) calculated by the missForest function. The missForest algorithm proceeds iteratively, training a random forest 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:

$$\Delta_{cont} = \frac{\sum_{j \in N} \left(X^{i,l} - X^{i,p}\right)^2}{\sum_{j \in N} \left(X^{i,l}\right)^2}, \tag{2.1}$$

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 on all continuous traits. For categorical variables, the difference Δ_{cat} is:

$$\Delta_{cat} = \frac{\sum_{k \in F} \sum_{j} J_{X^{i,l} \neq X^{i,p}}}{n(NA)}, \tag{2.2}$$

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, imputation error rates can be estimated. A mean square error (MSE) for each continuous trait and a proportion of falsely classified values (PFC) for each categorical trait are returned (the function can also return an overall normalised MSE for all continuous and overall PFC values for all categorical traits). The MSE for a trait is defined as:

$$\sqrt{\frac{mean\left(\left(X_t - X_i\right)^2\right)}{var\left(X_t\right)}}, \tag{2.3}$$

where X_t is a vector of the complete trait values and X_i a vector of the imputed trait values (Stekhoven 2011). For categorical traits, the is calculated as the PFC (Δ_{cat}, Equation 2.2). Imputation performance improves with decreasing error values.

I imputed 8 trait datasets for each class and plotted the MSE and PFC across all imputations. I then investigated whether imputations were robust examining whether values across imputations were congruent, or, on the other hand, showed a high variability.

2.3 Results

2.3.1 Outputs

I collected and imputed data for 10 traits across 11637 avian species, 5502 mammalian species, 10334 reptilian species and 7007 amphibian species. Produced synonym datasets are available alongside the trait data.

2.3.2 Imputation performance

Built-in errors

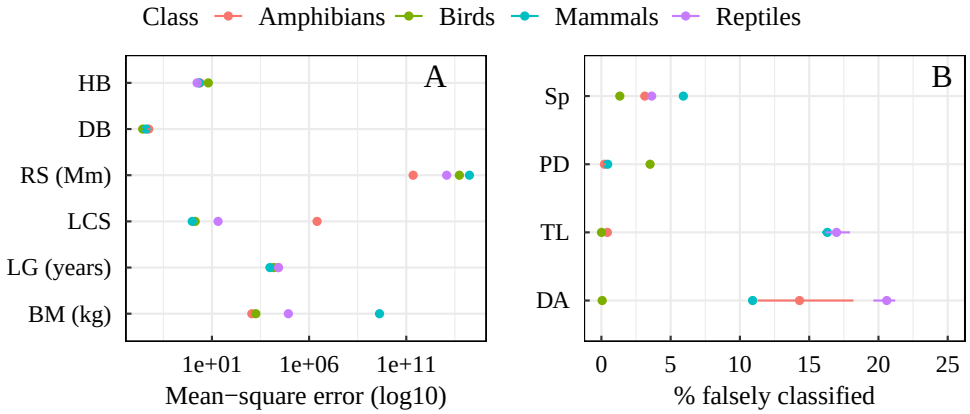

Figure 2.6

2.3.3 Imputation robustness

Congruence of several imputations

Comparison with another collected and imputed datasets for mammals and birds

2.4 Discussion

- Taxonomic challenges
- Biases in availability of trait information across classes
- Imputation robustness