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 28, 2019

Abstract

Contents

List of Tables **3**

List of Figures **4**

1 Introduction **6**

2 Literature review **7**

3 Collecting and imputing ecological trait data across terrestrial vertebrates **8**

3.1 Introduction 8

3.2 Methods 11

3.2.1 Ecological trait data collection 11

3.2.2 Phylogenetic information 13

3.2.3 Tackling taxonomic synonymy 14

3.2.4 Investigating biases in the coverage and completeness of trait information across classes 19

3.2.5 Imputing missing trait values 20

3.3 Results 27

3.3.1 Outputs 27

3.3.2 Biases in the availability of trait information: non randomness in coverage and completeness and patterns in missing trait values 28

3.3.3 Spatial biases of trait completeness 35

3.3.4 Imputation performance and robustness 36

3.4 Discussion 38

List of Tables

3.1 Primary sources used for each compiled trait. 12
3.2 Species representation in phylogenetic trees (datasets corrected for taxonomy) 18
3.3 Phylogenetic signal in continuous and categorical traits and in range size 23
3.4 Conceptual design for examining imputation congruence for continuous traits 27
3.5 Model coefficients 36

List of Figures

3.1 Difference in species number due to taxonomic correction (A) and distribution of number of synonyms across datasets (B) 16
3.2 Procedure followed to drop replicated tips from phylogenies 17
3.3 Percentage of species represented in the phylogenies, with and without taxonomic corrections 17
3.4 Phylogenetic signal in continuous traits (Pagel's λ) estimated with both original phylogenies and modified phylogenies 24
3.5 Phylogenetic signal in categorical traits (δ) estimated with both original phylogenies and modified phylogenies 24
3.6 Trait coverage across all species before and after taxonomic correction 29
3.7 Distribution of completeness of trait information across species 30
3.8 Median completeness across families 31
3.9 Within-family median trait coverage in mammals 33
3.10 Within-family median trait coverage in birds 33
3.11 Within-family median trait coverage in reptiles (squamates) 34
3.12 Within-family median trait coverage in amphibians 34
3.13 Relationship between trait completeness and species geographical range size 35
3.14 missForest out-of-bag root-mean-square errors and proportion of falsely classified values 36
3.15 Distribution of trait values after imputation for body mass, longevity, litter/clutch size and distribution of range sizes 37
3.16 Imputation congruence across eight imputed datasets 38

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

Planetary anthropogenic threats are reshaping patterns of species diversity (Böhm et al., 2013; Schipper et al., 2008; Spooner et al., 2018; Stuart et al., 2004). Land-use change globally impact local species richness (Newbold et al., 2015). In turn, species losses can negatively affect ecosystem functioning (Hooper et al., 2005, Hooper et al., 2012). Understanding how increasing pressures will affect ecosystem functioning and the services they provide is vital to put into place efficient mitigation measures.

The earliest experiments investigating the effect of species richness on the stability of ecosystem processes were conducted in the late twentieth century (Naeem et al., 1994; Tilman and Downing, 1994). Hundreds of grassplot experiments have since then confirmed that higher diversity promotes higher primary productivity and increased ecosystem stability (Balvanera et al., 2006; Tilman et al., 2014). Intuitive mechanistic explanations of this phenomenon include increased niche complementary and resource partitioning, favouring more efficient resource use. Specifically, species properties (or traits) shape how species interact with their biological and physical environment. As such, traits are central to elucidate how species diversity links to ecosystem functions.

Strictly, traits are defined as phenotypic characteristics measurable the level of an individual, with an effect on organismal fitness or performance (McGill et al., 2006; Violle et al., 2007). They can be physiological (e.g., metabolic rates), morphological (e.g., body mass), behavioural (e.g., activity time), phenological (e.g., anthesis); they can also relate to species life-history (e.g., longevity) or diet (e.g, trophic level). Traits shape species fundamental and realised niches; for instance, physiological traits influence species thermal tolerances, participating in defining their geographical distributions

(Calosi et al., 2010; Khaliq et al., 2017). Morphological attributes (e.g. body mass, eye position, etc.) participate in shaping the structure of food webs, and in determining the strength of inter- and intra-specific competition (Gravel et al., 2016; Laigle et al., 2018). As such, traits determine how species use and impact on their environment. Specifically, effect traits underpin species resource use and define organismal contributions to ecosystem functions. Response traits are those involved in determining species responses to environmental changes and can overlap with effect traits. This distinction between response and effect trait led to the development of a conceptual framework, the 'response–effect' paradigm (Lavorel and Garnier, 2002; Luck et al., 2012; see Chapter 1), which aims to understand how traits shape species responses to environmental changes, and how these changes in turn affect ecosystem functioning.

A looser, more flexible definition of trait is sometimes adopted. Some characteristics only measurable at the species level can be referred to as 'ecological' traits. Examples of ecological traits, only measurable in relation to species occurrence, include habitat preferences (breadth or thermal/moisture preferences). In this chapter, I use this more flexible definition of trait, and consider species 'ecological' traits.

Trait-based approaches are increasingly used to understand processes underpinning species coexistence and biodiversity–ecosystem functioning relationships. Notably, ~~their~~ are widely employed in the context of the response–effect paradigm, and publications in this field have increased exponentially since the 2000s (Hevia et al., 2017). Nevertheless, studies investigating how environmental changes are likely to affect species and ecosystem functions non-randomly with respect to species traits are both extremely taxonomically biased in favour of plants and invertebrates (more than 75% of analysed studies in the metanalysis conducted by Hevia et al. (2017)), and spatially biased towards local scales (about 60% of the studies in the metanalysis conducted by Hevia et al. (2017); more than 90% of the studies were conducted at local or national scales). Consequently, our understanding of biodiversity–ecosystem functioning relationships at various spatial scales need to be refined (Isbell et al., 2018; Thompson et al., 2018). Moreover, although terrestrial vertebrates have been extensively studied in the past (Titley et al., 2017), how environmental changes may affect their global contributions to ecosystem functions needs to be investigated further.

Indeed, vertebrates play diverse and important ecosystem roles. Through frugivory, they participate in seed dispersion (McConkey et al., 2012; Mokany et al., 2014; Wandrag et al., 2015). They are significant pollinators in numerous ecosystems (Ratto et al., 2018). Vertebrate herbivores impact global plant diversity patterns through top-down regulation (Lin et al., 2018; Zhang et al., 2018).

They also contribute in regulating animal populations through their predatory activity (Barber et al., 2010; Letnic et al., 2012; Luck et al., 2012; Paine et al., 2016; Salo et al., 2010). As scavengers, they participate in nutrient cycling and energy transfers (Cunningham et al., 2018; Inger et al., 2016; Wilson and Wolkovich, 2011). Moreover, they are culturally important (Albert et al., 2018; Hirons et al., 2016), and a source of protein for many people (Alves et al., 2018).

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 include elucidating how species traits influence their responses to land-use and climate change at global scales, and how changes in community composition may affect ecosystem functions.

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. This constituted the aim of this first chapter: here, I collected and imputed trait data across the four terrestrial vertebrate classes (mammals, birds, reptiles and amphibians).

In this chapter, I present the methods I used to collect and impute trait values across terrestrial vertebrates. 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: Jones et al. (2009); amphibians, AmphiBIO: Oliveira et al. (2017); amniotes: Myhrvold et al. (2015); both mammals and birds: Cooke et al. (2019)). Other trait datasets have been released on on-line 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).

Data collection was constrained by the amount of information available in the literature. All 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. Targeted traits related to species life-history, morphology, behaviour and feeding habits (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 common variable could describe species mobility across classes. Although species' abilities to move in their environment is key in understanding how species will respond to anthropogenic pressures (Barbet-Massin et al., 2012; Pearson, 2006; Schloss et al., 2012), this trait was not considered for the above reasons. In this chapter, I detail the methodology I employed to collate targeted traits. I elaborate on some of the challenges met when compiling data across many species, such as inconsistency of taxonomy across sources, and problems posed by taxonomic inflation and synonymy.

Despite a wealth of information across primary sources, trait data was likely to be incomplete across terrestrial vertebrates. Many species were likely to present missing trait data for many traits; and taxonomic and geographical biases in the global trait knowledge were likely to exist (González-Suárez et al., 2012; Hortal et al., 2014). The gap in global trait knowledge was termed the 'Raunkiær shortfall' by Hortal et al. (2014). Here, I assessed the Raunkiær shortfall for terrestrial vertebrates. I investigated whether trait data presented taxonomic, phylogenetic and spatial biases.

After examining patterns in the gaps in trait data information, I imputed missing trait values. This chapter finally details imputation methodology and examines imputation performance and robustness.

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 trait 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 of Threatened Species.

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**		
Oliveira et al., 2017	Amphibians	✓	✓	✓	✓		✓	✓	✓	✓		
Cooper			✓				✓				✓	
ior			✓									
xford			✓									
Wilman et al., 2014	Birds	✓							✓	✓		
Butchart		✓				✓						
Jones et al., 2009	Mammals	✓	✓	✓	✓		✓			✓		
Kissling et al., 2014								✓				
Gainsbury et al., 2018								✓				
Wilman et al., 2014		✓							✓	✓		
Pacifici et al., 2013		✓		✓	✓	✓						
Scharf et al., 2015	Reptiles	✓		✓	✓		✓	✓		✓		
Vidan et al., 2017										✓		
Stark et al., 2018		✓		✓			✓			✓		
Schwarz and Meiri, 2017							✓					
Novosolov et al., 2017		✓						✓			✓	
Novosolov et al., 2013							✓					
Slavenko et al., 2016		✓										
Myhrvold et al., 2015	Amniotes	✓	✓	✓	✓		✓					
IUCN Red List	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 EltonTraits database (Wilman et al., 2014). 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 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 item for which percent use was equal to or above 50% was considered to be part of the primary diet. Species for which no food item had percent use above 50% were considered to be omnivores.

For amphibians, diet information was extracted from AmphiBIO (Oliveira et al., 2017). Diet information was available as binary variables for diverse food items (leaves, flowers, seeds, fruit, arthropods and vertebrates). Percent use ~~were~~ not recorded, so these items were considered to form species primary diet. I pooled amphibian diet 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 habitat data files provided by the IUCN Red List. They 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 suitability and importance of the habitat; 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). Hedges et al. (2015) built a time-calibrated phylogenetic tree representing more than 50,000

species, using meta-analytic methods: results from 2274 phylogenetics and molecular evolution studies were assembled to build a 'Super Time' tree of life. Tree subsets for diverse clades are available at http://www.biodiversitycenter.org/ttol (downloaded 06/07/2018). The trees for vertebrate species were all 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). Using trees that were built using molecular data, and not life-history traits, allows to avoid circularity in further imputations of missing trait values.

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 (Chamberlain, 2018, Chamberlain and Szöcs, 2013). 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 the ITIS. When species were not found in the 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 (GBIF, 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 clades with high degrees of pseudoreplication in names, such as reptiles or amphibians, neither the

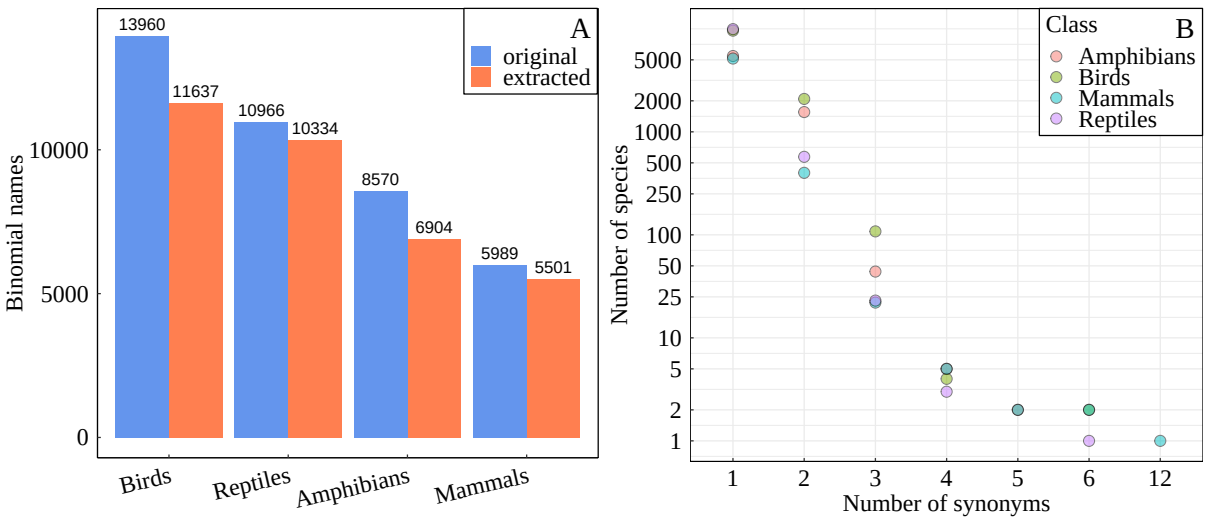

Figure 3.1: Difference in species number due to taxonomic correction (A) and distribution of number of synonyms across datasets (B). (A) shows the number of species across all primary sources (trait datasets and PREDICTS, excluding phylogenies), before and after correcting for taxonomy. Replacing identified 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.

Red List or the 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.

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).

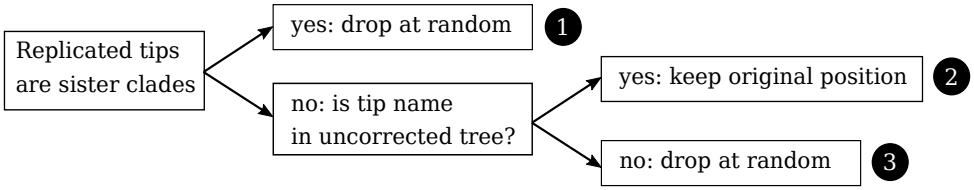

Figure 3.2: Procedure followed to drop replicated tips from phylogenies. Most of these were replicated twice. When replicated tips were sister clades, the tips to drop were chosen randomly, as it did not affect the 'true' phylogenetic position of the species (1). When replicated were not sister clades, 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).

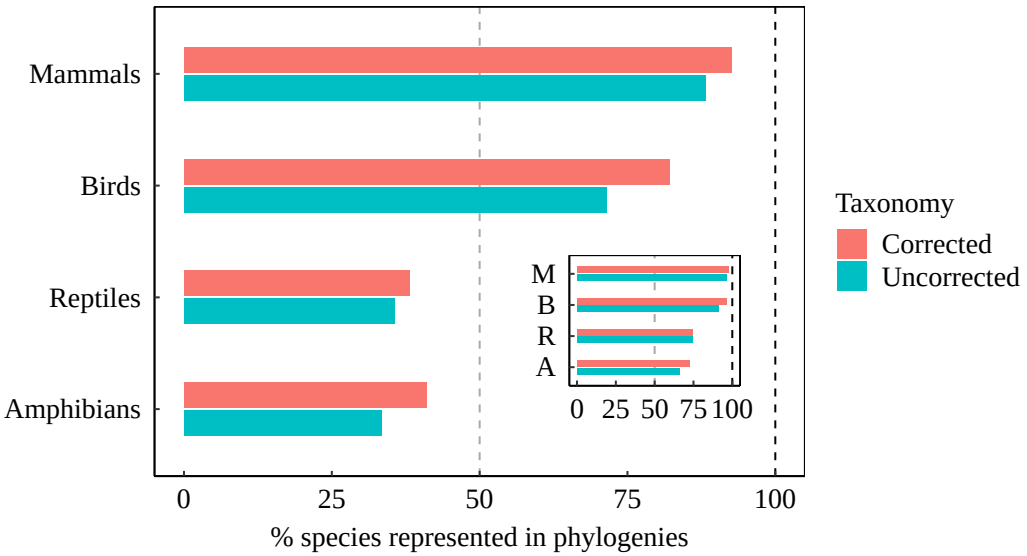

Figure 3.3: Percentage of species represented in the phylogenies, with and without taxonomic corrections. Overall, taxonomic correction increased species representation in phylogenetic trees. Representation for mammals and birds was high (after taxonomic correction: 82% of avian and 93% of mammalian 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. Figure 3.3 shows the phylogenetic representation of species figuring in the trait datasets. Overall, correcting for taxonomy

improved species representation in the trees. 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, even after taxonomic corrections (3.3). 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 attached non-represented species to the root of their genus, when possible (phytools package, Revell, 2016). Attaching species at the root of their genus created polytomies, which were resolved randomly (using multi2di in ape (Paradis and Schliep, 2018) and bifurcatr in PDcalc (Nipperess and Wilson, 2019)). 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 verified 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 modified 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 (datasets 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 Investigating biases in the coverage and completeness of trait information across classes

Taxonomic biases.

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. I tested whether taxonomic class impacted trait completeness using pairwise Kruskall-Wallis rank sum tests (the null hypothesis tested in each pair was that the distribution of completeness values were sampled from the same original distribution).

Phylogenetic biases.

Whether values are missing at random is likely to impact imputation errors, notably if some taxa appear to be under-sampled. 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. To assess whether missing values presented patterns, I represented within-family median completeness and within-family median 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 (using contMap, phytools package, Revell, 2016). 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).

Spatial biases.

I finally investigated whether trait completeness was spatially biased. Specifically, I tested the hypothesis that bigger geographical range sizes were correlated with better trait completeness. To that end, I fitted a generalised linear model with a Poisson error distribution: trait completeness was treated as count data (number of sampled traits). Class was added as an explanatory factor, interacting with range size. The model was written as: Completeness $\sim$ log(RS) + Class + log(RS):Class, and I specified a Poisson error distribution. I examined whether the fit of the model was good using a chi-squared test on the residual deviance.

NB. In all of the above, completeness was calculated over all predictor traits for each class (target traits and supplementary traits). As such, the total number of traits sampled in each class was not necessarily equal. The same analyses could be replicated on the set of target traits only. Analyses of spatial biases in trait sampling could also be developed in the future (see Chapter 5, Outline of future work).

Finally, I investigated whether correcting for taxonomy had an effect on trait completeness using Wilcoxon rank sum tests. I tested whether the median trait completeness was significantly higher for datasets corrected for taxanomy, than for uncorrected datasets, in each class.

Conclusion: imputing missing values to increase coverage. 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 (Cooke et al., 2019, Molina-Venegas

et al., 2018, Swenson, 2014). Penone et al., 2014 assessed the performance of four different imputation approaches (K-nearest neighbour (kNN, Troyanskaya et al., 2001), multivariate imputation by chained equations (mice, van Buuren and Groothuis-Oudshoorn, 2011), random forest algorithms as implemented in R by missForest (Stekhoven and Bühlmann, 2012, Stekhoven, 2016) and phylogenetic imputations implemented with PhyloPars (Bruggeman et al., 2009)). 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 (Münkemüller et al., 2012). Here, I used Pagel's λ (Pagel, 1999), estimated with the R function phylosig (phytools package, Revell, 2016), 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 (Münkemüller et al., 2012). 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 (phy-

tools) also ~~allowed~~ 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 et al., 2009 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 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., 2018 (https://github.com/mrborges23/delta_statistic). 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. I noted that the function developed by Borges et al., 2018 could not be implemented if phylogenetic trees contained 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 effect 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*

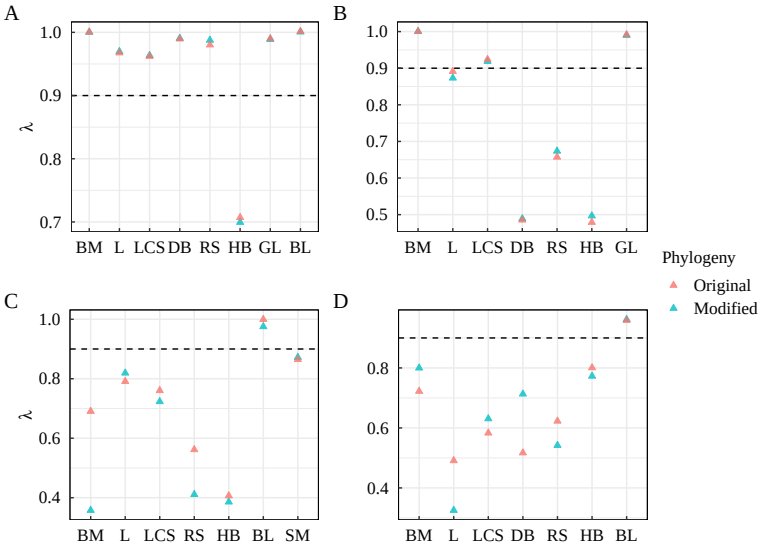

Figure 3.4: Phylogenetic signal in continuous traits (Pagel's λ) estimated with both original phylogenies and modified phylogenies. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Overall, altering the phylogenies by correcting for taxonomy and by increasing species representation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

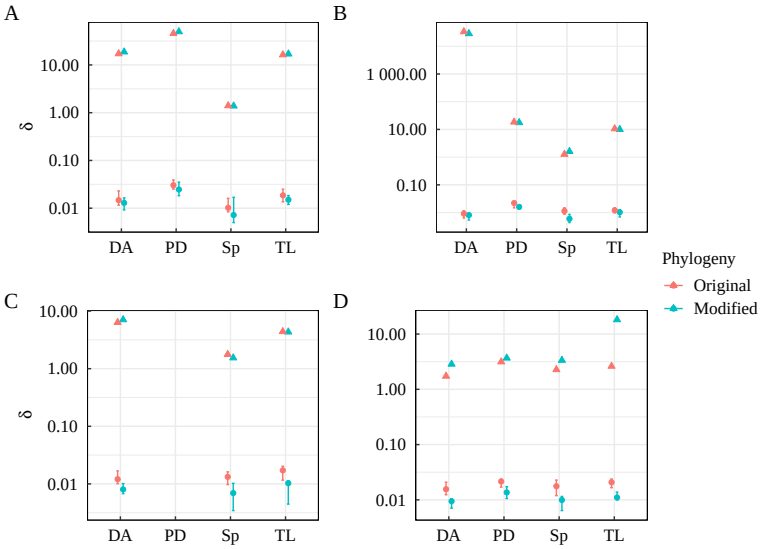

Figure 3.5: Phylogenetic signal in categorical traits (δ) estimated with both original phylogenies and modified phylogenies. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Triangle-shaped points represent the estimated phylogenetic signal in each trait; round-shaped points represent the median null expectation of the phylogenetic signal (±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. Moreover, Penone et al. (2014) also showed that adding phylogenetic information did not, in any case, decrease the accuracy of imputations.

Phylogenetic relationships were included as additional predictors in the form of phylogenetic eigenvectors (Diniz-Filho et al., 2012), extracted from the phylogenies using the PVR package (Santos, 2018). In this package, phylogenetic eigenvectors were computed from a phylogenetic distance matrix, and calculated using principal coordinate analysis methods. Phylogenetic eigenvectors summarised the relationships among species, and the first set of eigenvectors reflected larger distances, capturing divergences closer to the root (Diniz-Filho et al., 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: total of 32 imputed datasets).

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 and Bühlmann, 2012 for more details).

When the stopping criterion has been met, OOB imputation errors are estimated. OOB errors refer to errors estimated from sub-samples of the data (bootstrap datasets, on which models are trained). OOB 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). With the OOB error, when the complete trait data is not provided, the MSE is calculated from the bootstrap datasets. For categorical traits, the OOB PFC is calculated as the PFC (Δ_{cat}, Equation 3.2), using the bootstrap sub-samples. Breiman, 2001 showed that OOB estimates provide accurate proxies of the true imputation error.

To assess imputation accuracy, I retrieved OOB imputation errors (OOB root-MSE and PFC) across the 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 (Pearson 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 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.

Taxonomic biases in the availability of trait information: exacerbated Raunkiær shortfall in reptiles and amphibians.

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). To conclude, trait coverage revealed important taxonomic biases, with higher resolution of trait information across mammals and birds. A clear contrast in trait information appeared between mammals and birds versus herptiles,

highlighting the existence of taxonomic biases in data collection.

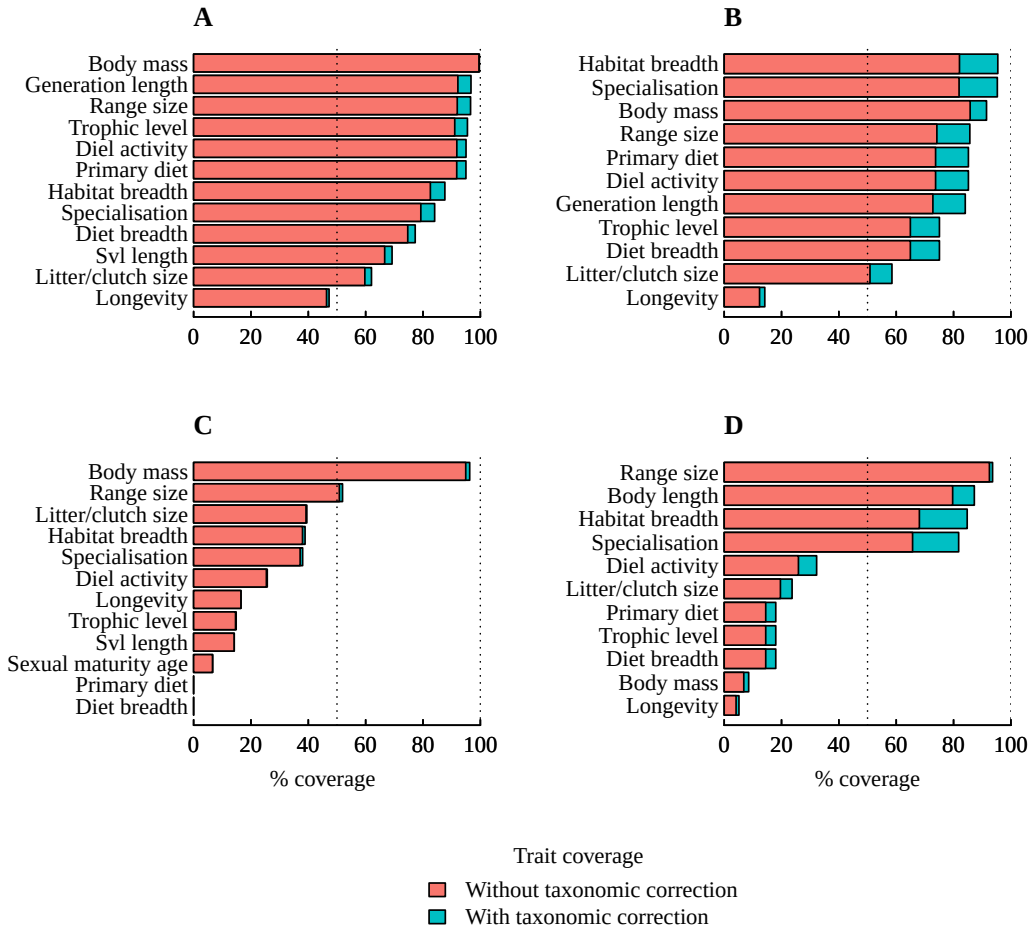

Figure 3.6: Trait coverage across all species before and after taxonomic correction. Here are shown target traits as well as a few other traits used in imputations as additional predictors (such as generation length for mammals and birds or body length for amphibians). **(A)** Mammals (5885 species before correction, 5502 after correction); **(B)** birds (13554 species before correction, 11637 after correction); **(C)** reptiles (10722 species before correction, 10334 after correction) and **(D)** 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.

Trait completeness. Trait completeness reflected similar biases as trait coverage (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$<2{\cdot}10^{-16}$ in all cases), showing that class had a significant effect on the availability of trait information.

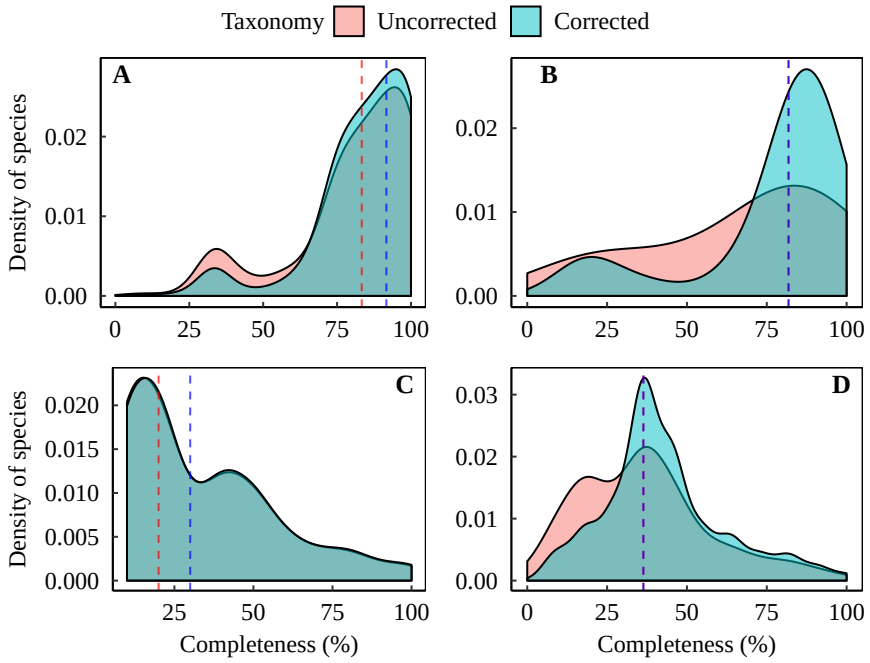

Figure 3.7: Distribution of completeness of trait information across species. (A) Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Completeness was calculated here for the same set of traits shown in Figure 3.6 (all predictor traits). Correcting for taxonomy affected completeness, significantly shifting the distributions to the right (alternative hypothesis, Wilcoxon rank sum tests: uncorrected medians were lower than corrected medians; mammals: p-value=$1.2{\cdot}10^{-9}$; birds: p-value<$2.2{\cdot}10^{-16}$; reptiles: p-value=0.025; amphibians: p-value<$2.2{\cdot}10^{-16}$). 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<$2 \cdot 10^{-16}$ in all cases)).

Non-randomness in trait information: within-class phylogenetic biases.

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 3.8 A). 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 3.8 B). 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; 3.8 C). In amphibians, groups of families in the Anura

order showed both the best and worst median completeness (Figure 3.8 D).

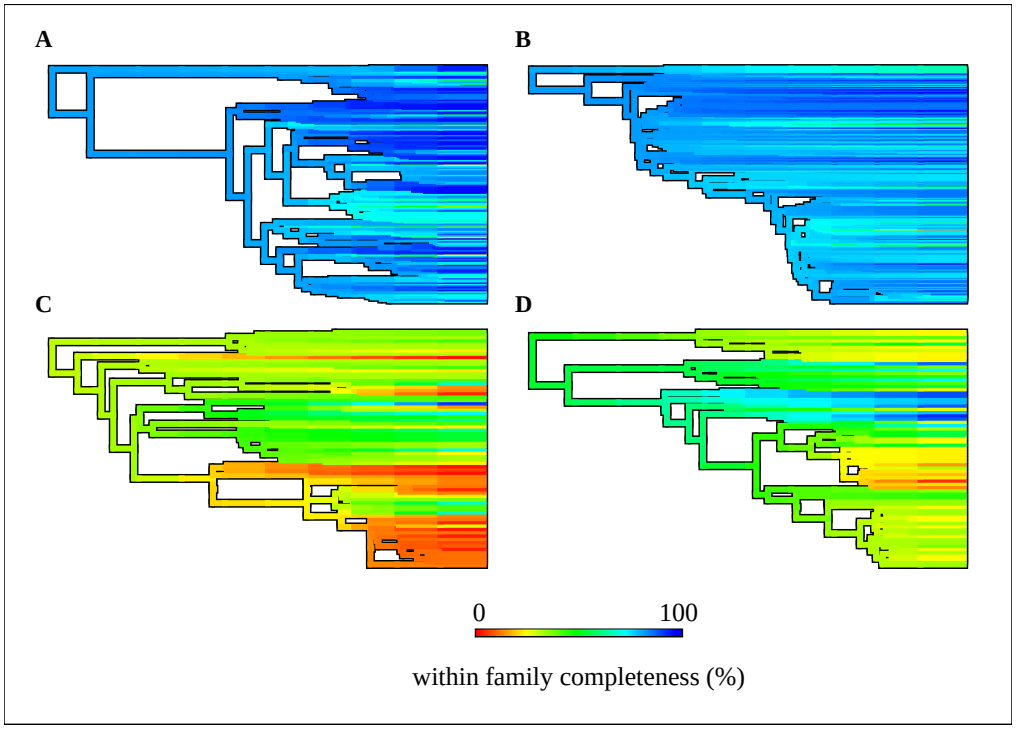

Figure 3.8: Median completeness across families. Tips labels are not shown here for better visualisation of the results; the same figures with tip labels are provided in the SI (zooming into the figure is necessary for mammals and birds); tip label information includes order and family. **(A)** Mammalian 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 reptiles and amphibians.

Overall, these results showed that trait completeness was not random with regard 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). The last subplot represents the contribution of each family to the total number of species in the phylogeny.

As trait coverage decreased, family clusters of similar median trait coverage became more visible. In mammals, a cluster of families showed low median coverage for trophic level; most of these families were in the Cetartiodactyla order, which contained marine and aquatic mammals. Families in the

Chiroptera order appeared to be less well sampled for three traits compared to other orders (body length, litter size, longevity, subplots J to L, Figure 3.9). Families in the Primates order also appeared to be less well sampled for certain traits (subplots E to I). Among the best sampled mammalian orders were the Diprotodontia and the Carnivora. For birds, the patterns were less clear (Figure 3.10). Diet information was less resolved for Struthioniformes (subplots H and L: top of the tree); overall, no systematic bias emerged. In reptiles, the lower part of the phylogeny was systematically less well sampled, with a few exception for families such as Boidae and Pythonidae (subplot E or F: green-blue areas within the red cluster). Overall, trait information for most snakes was systematically less well resolved than for other squamates. In amphibians (Figure 3.11), families in the Caudata order were overall better sampled (second branch from the root) , as well as a group of families in the Anura order (third branch from the root, above part). On the other hand, a large number of families in the Anura order, and most families in the Gymnophonia order (first branch from the root), were systematically less well sampled. These families nevertheless largely contributed to the total number of species.

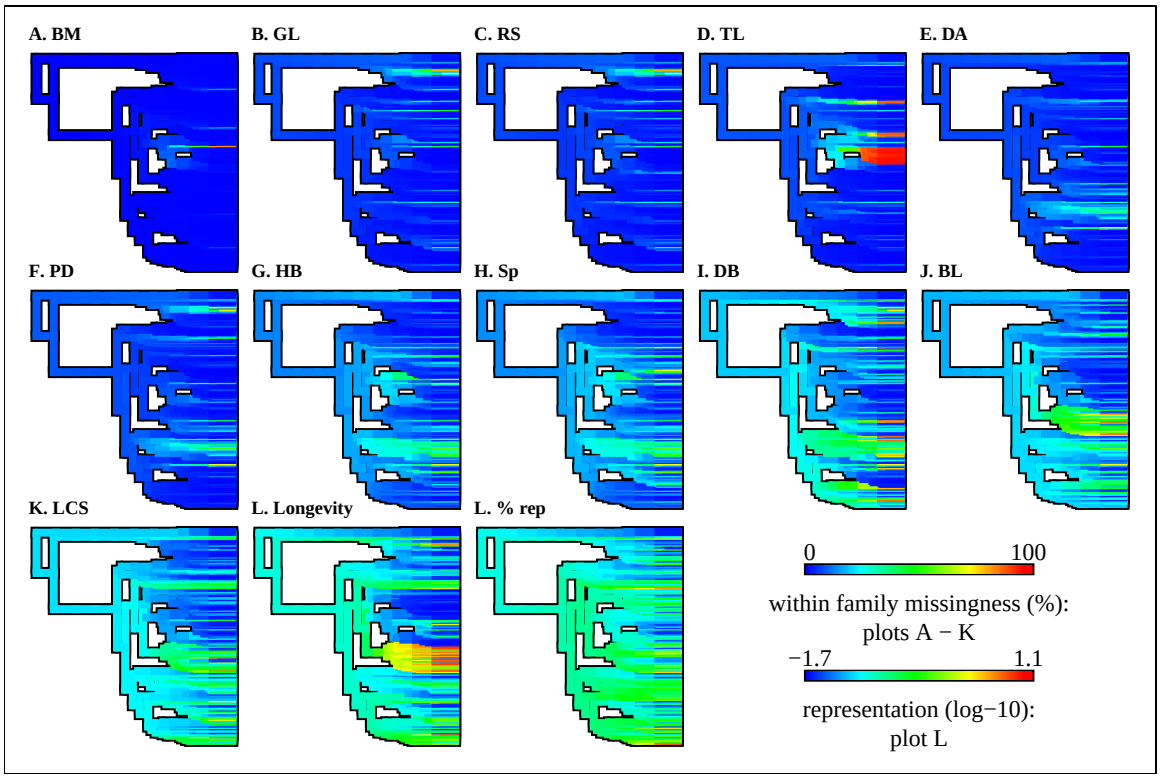

Figure 3.9: Within-family median trait coverage in mammals.

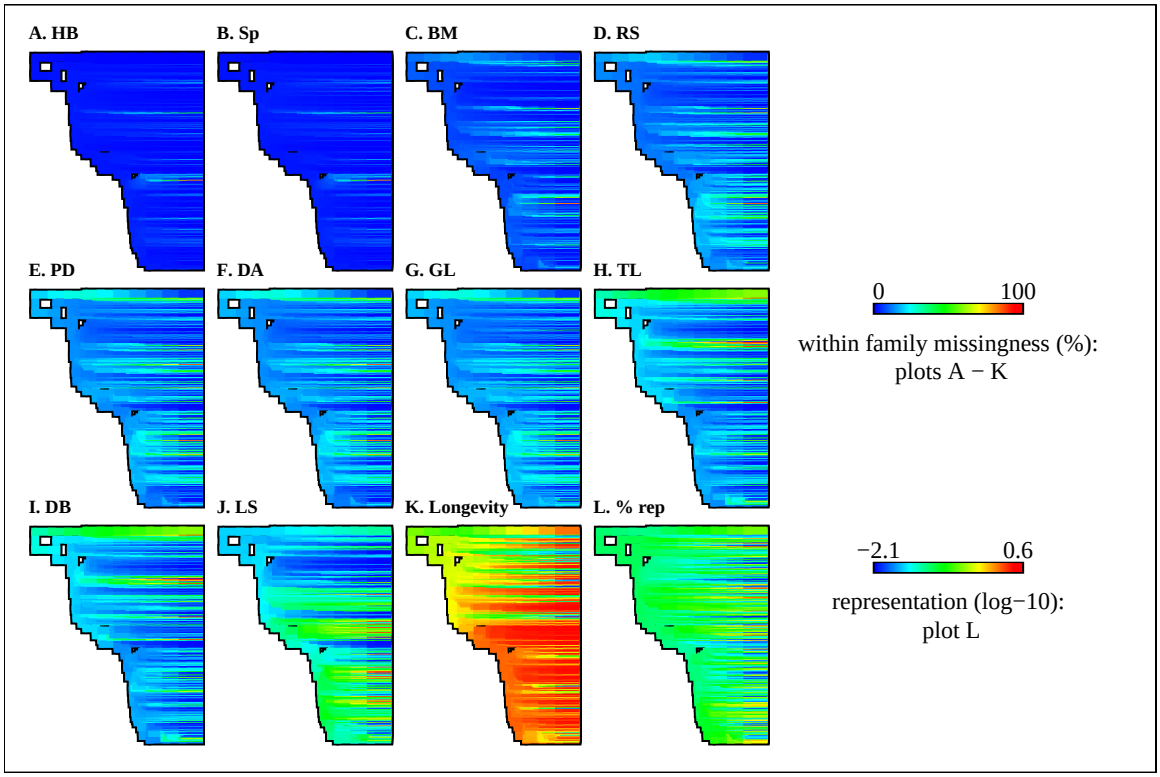

Figure 3.10: Within-family median trait coverage in birds.

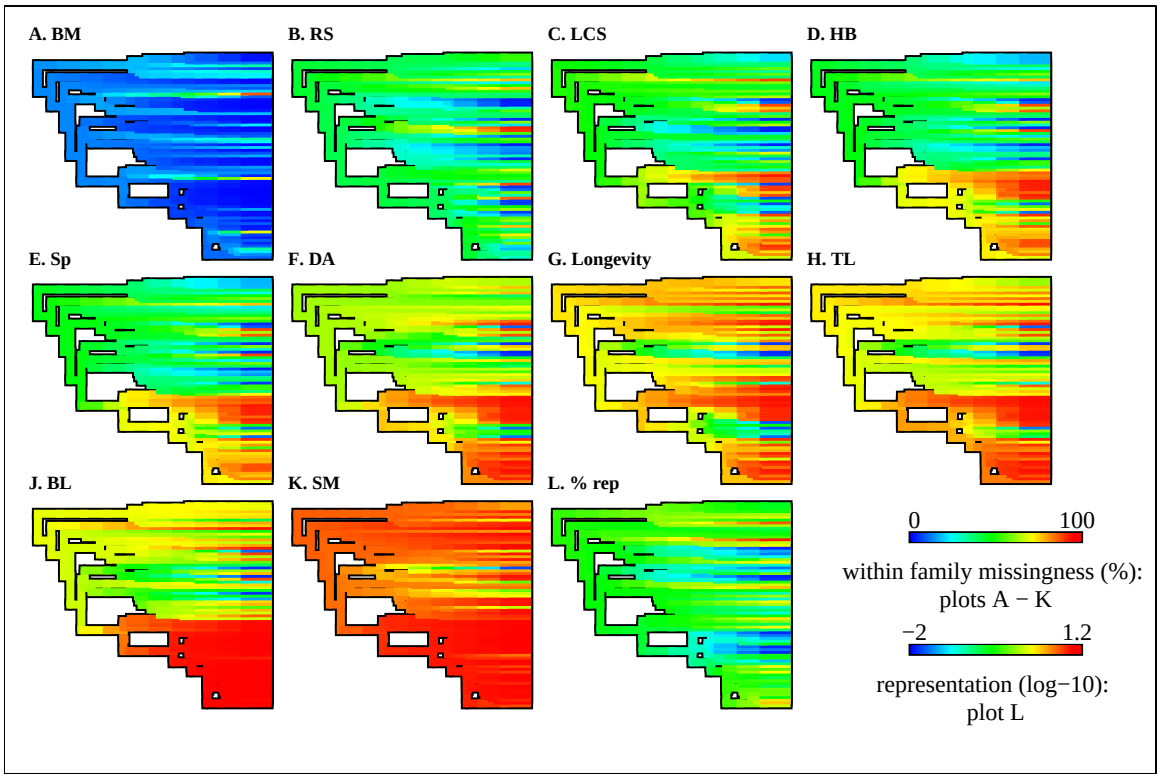

Figure 3.11: Within-family median trait coverage in reptiles (squamates).

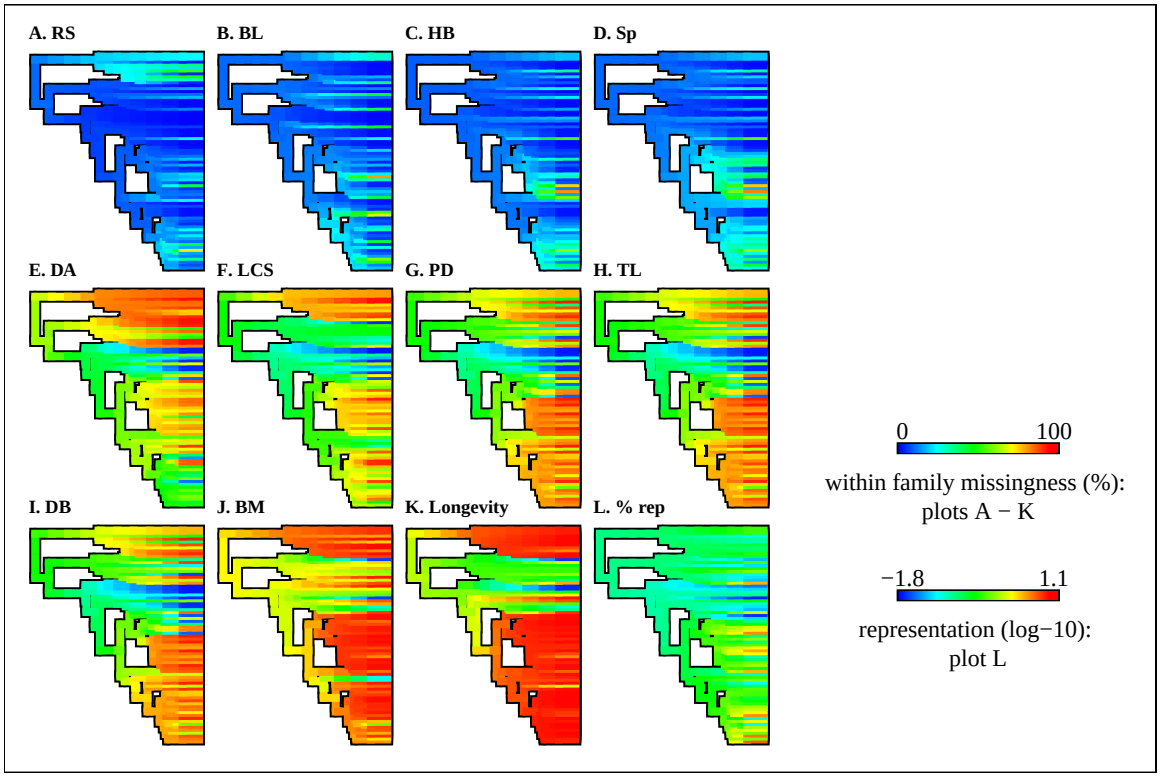

Figure 3.12: Within-family median trait coverage in amphibians.

3.3.3 Spatial biases of trait completeness

Geographical range size had a significant effect on the number of sampled traits (Table **??**, Figure 3.13). Class had a significant effect on the rate of increase in the number of sampled traits, except for reptiles (the rate of increase for reptiles was similar to the rate of increase for amphibians). Baseline rates were higher for mammals and birds than for herptiles, but rates of increase were higher for herptiles. A goodness-of-fit test on the residual deviance did not gather evidence that the model fitted badly (p=1, residual deviance: 14440 on 27105 degrees of freedom).

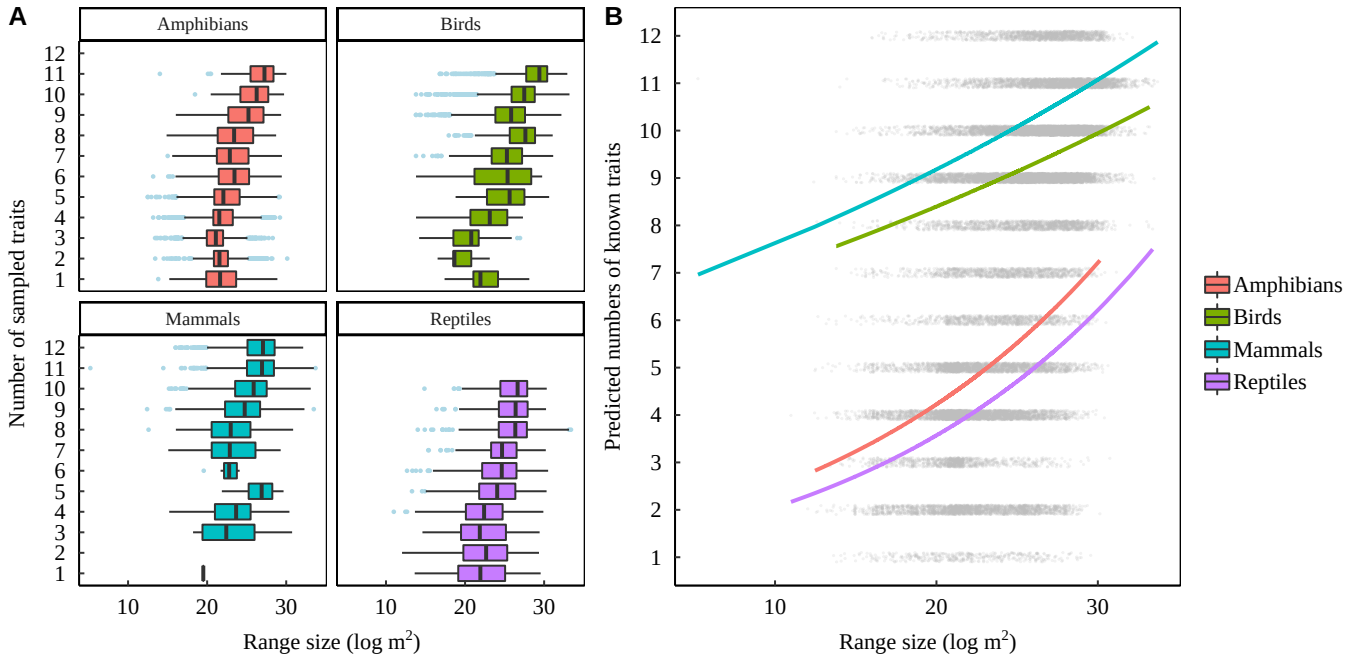

Figure 3.13: Relationship between trait completeness and species geographical range size. (A) Boxplots showing the number of traits sampled in each class against species geographical range sizes ($\log_{10}$). **(B)** Regression lines for the fitted generalised linear model. Grey points represent empirical values (not colour coded for better visual clarity). The model was fitted using a Poisson error distribution. Class was added as an explanatory variable with interaction.

Conclusion. Trait data across terrestrial vertebrates is taxonomically biased, with better resolution of trait information in mammals and birds. Phylogenetic biases are strong in reptiles and amphibians, where entire clades appeared to be systematically less well sampled. Finally, across all classes, species with bigger geographical range sizes are more likely to have more complete trait information. The effect of geographical range size is more pronounced for herptiles than for mammals and birds: the decrease in trait completeness corresponding to a decrease in range size is steeper for amphibians and reptiles.

Table 3.5: Model coefficients. A generalised linear model with a Poisson error distribution was fitted to the number of sampled traits, with range size and class as interacting explanatory variables. All effects were significant, except for the interaction between reptiles and range size.

Independent variable	Estimate	Std. Error	z-value	Pr(>|z|)
Intercept (Amphibians)	0.37	0.046	8.05	8.07E-16
log(RS)	0.05	0.002	26.72	2.94E-157
Birds	1.42	0.055	25.66	3.13E-145
Mammals	1.47	0.058	25.16	1.04E-139
Reptiles	-0.21	0.066	-3.15	1.64E-03
log(RS):Birds	-0.04	0.002	-15.89	7.47E-57
log(RS):Mammals	-0.03	0.002	-14.27	3.14E-46
log(RS):Reptiles	0.00	0.003	0.69	4.92E-01

3.3.4 Imputation performance and robustness

Out-of-bag imputation errors.

Figure 3.14 A shows OOB root-mean-squared errors for each continuous traits (shown here for one randomly selected imputed dataset). Figure 3.14 B is the OOB proportion of falsely classified values for categorical traits (for the same imputed dataset). Estimated prediction errors for categorical traits were low to moderate (all below 20%). 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 (see Figure 3.15, which shows the distribution of trait values after imputations; large prediction errors are estimated where traits can attain high values).

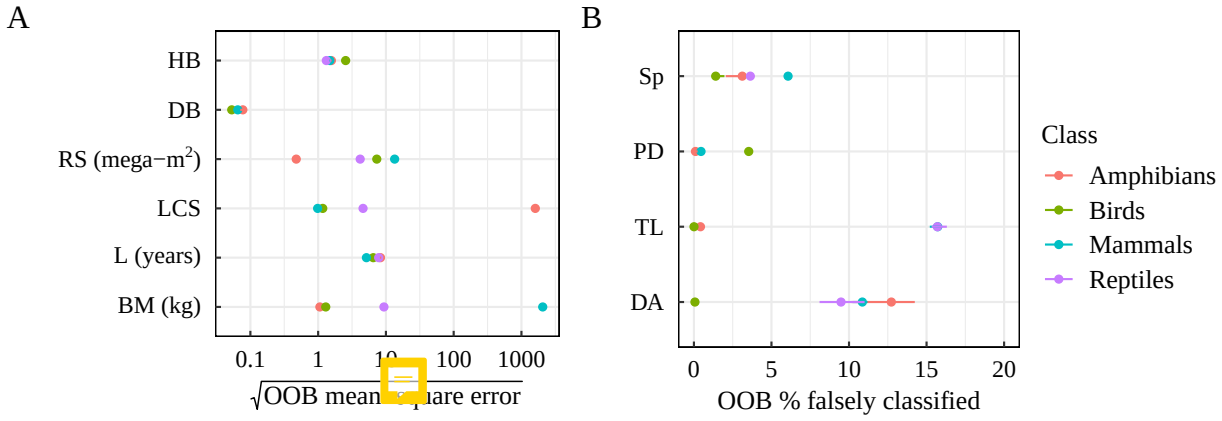

Figure 3.14: missForest out-of-bag root-mean-squared errors and proportion of falsely classified values. **(A)** Out-of-bag root-mean-square errors for continuous traits. **(B)** Out-of-bag proportion of falsely classified values.

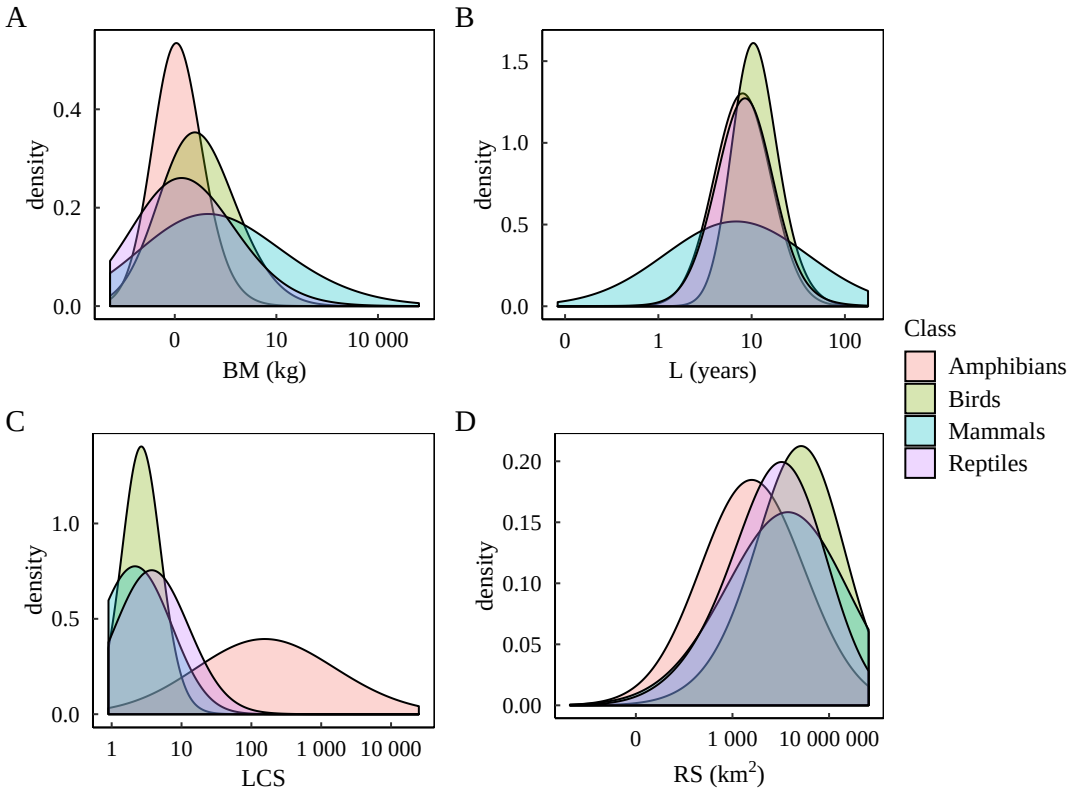

Figure 3.15: Distribution of trait values after imputation for body mass, longevity, litter/clutch size and distribution of range sizes.

Congruence of imputed values among eight imputed datasets.

Figure 3.16 A 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 3.16 B 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.

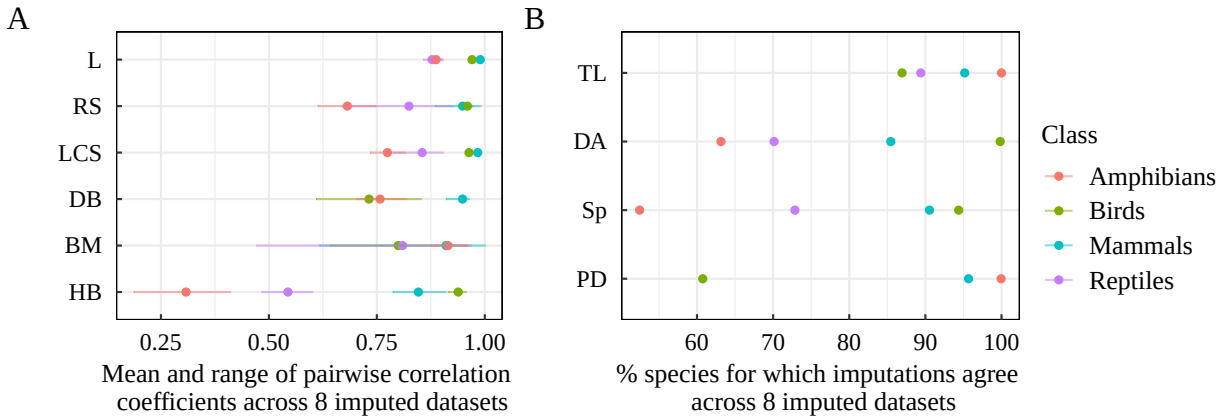

Figure 3.16: Imputation congruence across eight imputed datasets.

Conclusion. OOB imputation errors and imputation congruence showed that predictions were overall robust. Habitat breadth was the only variable for which imputation congruence was highly variable across classes, and below 50% for amphibians. Imputation accuracy may be impacted by the phylogenetic biases in trait completeness. Further work could investigate the impact of non-randomness in the sampling of trait values on imputation accuracy.

3.4 Discussion

In this work, I compiled and imputed data on 10 traits across 5502 mammalian, 10334 reptilian, 11637 avian and 6904 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, which was enabled by all past and recent efforts to release trait information in the public domain. Note that the current imputed dataset contains fossil species, as some of the primary sources provided estimates for these. Some marine and aquatic mammals are also represented. Both

fossil and non-terrestrial species could be filtered out in the future.

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 to the dataset. 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 abilities to both move within their habitats (home range) and to disperse and colonise new areas has a major impact on their aptitudes to cope with anthropogenic changes and on ecosystem functions (Schloss et al., 2012; Tucker et al., 2018). Nevertheless, traits relating to mobility in amphibians or reptiles were unavailable. The only readily available variable that could have been added was volancy. 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 preferences).

The data collection revealed important biases in the availability of trait information across terrestrial vertebrates. Mammals and birds were better sampled than reptiles and amphibians, even for species with similar range sizes. In herptiles, trait information was strongly phylogenetically biased. These results illustrate the biases in global biodiversity knowledge identified by Hortal et al. (2014). Identified gaps are consistent with biases found in González-Suárez et al. (2012) (study conducted on mammals only). Such biases have important consequences on macro-ecological studies. For instance, some analyses make inferences from certain taxa, for which data is available, to other missing-value taxa. Nevertheless, if the sample of studied species is not drawn at random, extrapolations may not be valid. As such, eliminating missing data can, not only, reduce sample sizes, but also bias estimates Nakagawa and Freckleton, 2008. Imputing missing values therefore appeared to be an interesting option; nevertheless, non-randomness in missing values could bias imputation accuracy.

Penone et al. (2014) conducted a simulation study where missing values were introduced in a trait dataset (10 to 80% missing values). Values were removed in three different ways: completely at random; at random with respect to only one trait; and finally, at random with respect to phylogenies. Their results showed that differences in imputation error in these three cases were marginal and not significant. For some traits, there was a trend for bigger imputation error where missing data were clustered in closely related species. Nevertheless, Penone et al. (2014) showed that missForest imputations were robust even when trait data was phylogenetically biased.

Using non parametric random forest algorithms to impute missing trait values, 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 OOB errors for each variable. Second, no underlying data distribution was assumed in the process. Third, missForest was computationally faster than other methods, which was an important criterion. Finally, missForest has been shown to outperform or perform as well as other approaches (Penone et al., 2014; Stekhoven and Bühlmann, 2012). Moreover, as stated above, Penone et al. (2014) showed that missForest imputations were robust even when missing values were not missing at random. Congruence results and imputation errors obtained here showed that missForest performed overall rather well. Habitat related variables showed less congruence and may as such be more difficult to impute.

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 available 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).

I also did not compile any metric reflecting intra-specific variability in continuous traits. Intraspecific variability has been shown to have important effects on ecological systems (Bolnick et al., 2011; Des Roches et al., 2018; González-Suárez and Revilla, 2013). A growing body of literature encourages trait-based research to include intraspecific variability (Carmona et al., 2016; Violle et al., 2012). 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. Here, taxonomic synonymy artificially increased the amount of missing trait values by creating pseudoreplicates of the same species, inflated the overall number of species and significantly lowered median trait completeness. Taxonomic uncertainty is a recurring problem in ecology and conservation (Isaac et al., 2004). For instance, Cardoso et al. (2017) showed than taxonomic inaccuracies and errors in species checklists lead to the overestimation of plant diversity in the Amazon. The lack of a universal, standardised database for species names complicates species identification. Nevertheless, the production of such a database is difficult to achieve, partly because different conceptual definitions of what a species is can lead to different taxonomic systems (Isaac et al., 2004). Moreover, frequent updates would be necessary for the database to be in line with the most recent taxonomic revisions. Here, the procedure that I developed to tackle taxonomic redundancy built upon taxonomic information contained

in the Red List and the ITIS. Overall, the procedure was not optimal, as these databases did not contain standardised taxonomic information. Nevertheless, it participated in reducing taxonomic mismatches and in increasing trait coverage. Initiatives such as the Taxonomic Name Resolution Service for plants (http://tnrs.iplantcollaborative.org/) or the Global Biodiversity Information Facility should encourage researchers to inspect taxonomic uncertainty when working with a large number of species.

Finally, Cooke et al. (2019) released a comprehensive dataset of six mammalian and avian traits. 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. (2019). 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. The results figure in the SI for comparative purposes.

Conclusion. Future work will build upon the trait data compiled and imputed as presented in this chapter. Even though collation methods may be revisited in the future, the framework is likely to remain similar. More work could be dedicated to assess the impact of phylogenetic biases in trait completeness on imputation accuracy.

I illustrate the first use of this data in the next chapter, which investigates how land-use change impacts the functional diversity of vertebrate communities at global scales. In the last chapter, I detail some research questions that this data will allow to investigate in the future months ~~on~~ my PhD.

Bibliography

Albert, C., Luque, G. M., and Courchamp, F. (2018). The twenty most charismatic species. *PLoS ONE*. DOI: 10.1371/journal.pone.0199149.

Alves, R. R. N., Souto, W. M. S., Fernandes-Ferreira, H., Bezerra, D. M. M., Barboza, R. R. D., and Vieira, W. L. S. (2018). Chapter 7 - The Importance of Hunting in Human Societies. *Ethnozoology*. Ed. by R. R. N. Alves and U. P. Albuquerque. Academic Press, 95 –118. DOI: https://doi.org/10.1016/B978-0-12-809913-1.00007-7.

Balvanera, P., Pfisterer, A. B., Buchmann, N., He, J. S., Nakashizuka, T., Raffaelli, D., and Schmid, B. (2006). Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology Letters*. DOI: 10.1111/j.1461-0248.2006.00963.x.

Barber, N. A., Mooney, K. A., Greenberg, R., Philpott, S. M., Van Bael, S. A., and Gruner, D. S. (2010). Interactions among predators and the cascading effects of vertebrate insectivores on arthropod communities and plants. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1001934107.

Barbet-Massin, M., Thuiller, W., and Jiguet, F. (2012). The fate of European breeding birds under climate, land-use and dispersal scenarios. *Global Change Biology*. DOI: 10.1111/j.1365-2486.2011.02552.x.

Böhm, M. et al. (2013). The conservation status of the world's reptiles. *Biological Conservation*. DOI: 10.1016/j.biocon.2012.07.015.

Bolnick, D. I., Amarasekare, P., Araújo, M. S., Bürger, R., Levine, J. M., Novak, M., Rudolf, V. H., Schreiber, S. J., Urban, M. C., and Vasseur, D. A. (2011). *Why intraspecific trait variation matters in community ecology*. DOI: 10.1016/j.tree.2011.01.009.

Borges, R., Machado, J. P., Gomes, C., Rocha, A. P., and Antunes, A. (2018). Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*. DOI: 10.1093/bioinformatics/bty800.

Breiman, L. (2001). Randomforest2001. *Machine Learning*. DOI: 10.1017/CBO9781107415324.004.

Bruggeman, J., Heringa, J., and Brandt, B. W. (2009). PhyloPars: Estimation of missing parameter values using phylogeny. *Nucleic Acids Research*. DOI: 10.1093/nar/gkp370.

Calosi, P., Bilton, D. T., Spicer, J. I., Votier, S. C., and Atfield, A. (2010). What determines a species' geographical range? Thermal biology and latitudinal range size relationships in European diving beetles (Coleoptera: Dytiscidae). *Journal of Animal Ecology*. DOI: 10.1111/j.1365-2656.2009.01611.x.

Cardoso, D. et al. (2017). Amazon plant diversity revealed by a taxonomically verified species list. *Proceedings of the National Academy of Sciences*, 114.40, 10695–10700. DOI: 10.1073/pnas.1706756114.

Carmona, C. P., Bello, F. de, Mason, N. W., and Lepš, J. (2016). *Traits Without Borders: Integrating Functional Diversity Across Scales*. DOI: 10.1016/j.tree.2016.02.003.

Chamberlain, S. (2018). *rredlist: 'IUCN' Red List Client*. R package version 0.5.0.

Chamberlain, S. A. and Szöcs, E. (2013). taxize : taxonomic search and retrieval in R [version 2; referees: 3 approved]. *F1000Research*. DOI: 10.12688/f1000research.2-191.v2.

Cooke, R. S., Bates, A. E., and Eigenbrod, F. (2019). Global trade-offs of functional redundancy and functional dispersion for birds and mammals. *Global Ecology and Biogeography*, October 2018, 1–12. DOI: 10.1111/geb.12869.

Cunningham, C. X., Johnson, C. N., Barmuta, L. A., Hollings, T., Woehler, E. J., and Jones, M. E. (2018). Top carnivore decline has cascading effects on scavengers and carrion persistence. *Proceedings of the Royal Society B: Biological Sciences*. DOI: 10.1098/rspb.2018.1582.

Des Roches, S., Post, D. M., Turley, N. E., Bailey, J. K., Hendry, A. P., Kinnison, M. T., Schweitzer, J. A., and Palkovacs, E. P. (2018). The ecological importance of intraspecific variation. *Nature Ecology and Evolution*. DOI: 10.1038/s41559-017-0402-5.

Diniz-Filho, J. A. F., Bini, L. M., Rangel, T. F., Morales-Castilla, I., Olalla-Tárraga, M. Á., Rodríguez, M. Á., and Hawkins, B. A. (2012). On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*. DOI: 10.1111/j.1600-0587.2011.06949.x.

Fritz, S. A., Bininda-Emonds, O. R., and Purvis, A. (2009). Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecology Letters*. DOI: 10.1111/j.1461-0248.2009.01307.x.

Gainsbury, A. M., Tallowin, O. J., and Meiri, S. (2018). *An updated global data set for diet preferences in terrestrial mammals: testing the validity of extrapolation*. DOI: 10.1111/mam.12119.

González-Suárez, M. and Revilla, E. (2013). Variability in life-history and ecological traits is a buffer against extinction in mammals. *Ecology Letters*. DOI: 10.1111/ele.12035.

González-Suárez, M., Lucas, P. M., and Revilla, E. (2012). Biases in comparative analyses of extinction risk: Mind the gap. *Journal of Animal Ecology*. DOI: 10.1111/j.1365-2656.2012.01999.x.

Gravel, D., Albouy, C., and Thuiller, W. (2016). *The meaning of functional trait composition of food webs for ecosystem functioning*. DOI: 10.1098/rstb.2015.0268.

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*. DOI: 10.1093/molbev/msv037.

Hevia, V., Martín-López, B., Palomo, S., García-Llorente, M., Bello, F. de, and González, J. A. (2017). *Trait-based approaches to analyze links between the drivers of change and ecosystem services: Synthesizing existing evidence and future challenges*. DOI: 10.1002/ece3.2692.

Hirons, M., Comberti, C., and Dunford, R. (2016). Valuing Cultural Ecosystem Services. *Annual Review of Environment and Resources*, 41.1, 545–574. DOI: 10.1146/annurev-environ-110615-085831.

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J. H., Lodge, D. M., Loreau, M., Naeem, S., Schmid, B., Setälä, H., Symstad, A. J., Vandermeer, J., and Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs*. DOI: 10.1890/04-0922.

Hooper, D. U., Adair, E. C., Cardinale, B. J., Byrnes, J. E., Hungate, B. A., Matulich, K. L., Gonzalez, A., Duffy, J. E., Gamfeldt, L., and Connor, M. I. (2012). A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature*. DOI: 10.1038/nature11118.

Hortal, J., Bello, F. de, Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., and Ladle, R. J. (2014). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*. DOI: 10.1146/annurev-ecolsys-112414-054400.

Inger, R., Cox, D. T., Per, E., Norton, B. A., and Gaston, K. J. (2016). Ecological role of vertebrate scavengers in urban ecosystems in the UK. *Ecology and Evolution*. DOI: 10.1002/ece3.2414.

Isaac, N. J., Mallet, J., and Mace, G. M. (2004). Taxonomic inflation: Its influence on macroecology and conservation. *Trends in Ecology and Evolution*. DOI: 10.1016/j.tree.2004.06.004.

Isbell, F., Cowles, J., Dee, L. E., Loreau, M., Reich, P. B., Gonzalez, A., Hector, A., and Schmid, B. (2018). Quantifying effects of biodiversity on ecosystem functioning across times and places. *Ecology Letters*. DOI: 10.1111/ele.12928.

Jones, K. E. et al. (2009). PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*. DOI: 10.1890/08-1494.1.

Khaliq, I., Böhning-Gaese, K., Prinzinger, R., Pfenninger, M., and Hof, C. (2017). The influence of thermal tolerances on geographical ranges of endotherms. *Global Ecology and Biogeography*. DOI: 10.1111/geb.12575.

Kissling, W. D., Dalby, L., Fløjgaard, C., Lenoir, J., Sandel, B., Sandom, C., Trøjelsgaard, K., and Svenning, J. C. (2014). Establishing macroecological trait datasets: Digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. *Ecology and Evolution*. DOI: 10.1002/ece3.1136.

Laigle, I., Aubin, I., Digel, C., Brose, U., Boulangeat, I., and Gravel, D. (2018). Species traits as drivers of food web structure. *Oikos*. DOI: 10.1111/oik.04712.

Lavorel, S. and Garnier, E. (2002). *Predicting changes in community composition and ecosystem functioning from plant traits: Revisiting the Holy Grail*. DOI: 10.1046/j.1365-2435.2002.00664.x.

Letnic, M., Ritchie, E. G., and Dickman, C. R. (2012). Top predators as biodiversity regulators: The dingo Canis lupus dingo as a case study. *Biological Reviews*. DOI: 10.1111/j.1469-185X.2011.00203.x.

Lin, F., Jia, S., Luskin, M. S., Ye, J., Hao, Z., Wang, X., and Yuan, Z. (2018). Global signal of top-down control of terrestrial plant communities by herbivores. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1707984115.

Luck, G. W., Lavorel, S., Mcintyre, S., and Lumb, K. (2012). Improving the application of vertebrate trait-based frameworks to the study of ecosystem services. *Journal of Animal Ecology*. DOI: 10.1111/j.1365-2656.2012.01974.x.

McConkey, K. R., Prasad, S., Corlett, R. T., Campos-Arceiz, A., Brodie, J. F., Rogers, H., and Santamaria, L. (2012). *Seed dispersal in changing landscapes*. DOI: 10.1016/j.biocon.2011.09.018.

McGill, B. J., Enquist, B. J., Weiher, E., and Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution*. DOI: 10.1016/j.tree.2006.02.002.

Mokany, K., Prasad, S., and Westcott, D. A. (2014). Loss of frugivore seed dispersal services under climate change. *Nature Communications*. DOI: 10.1038/ncomms4971.

Molina-Venegas, R., Moreno-Saiz, J. C., Castro Parga, I., Davies, T. J., Peres-Neto, P. R., and Rodríguez, M. A. (2018). *Assessing among-lineage variability in phylogenetic imputation of functional trait datasets*. DOI: 10.1111/ecog.03480.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*. DOI: 10.1111/j.2041-210X.2012.00196.x.

Myhrvold, N. P., Baldridge, E., Chan, B., Sivam, D., Freeman, D. L., and Ernest, S. K. M. (2015). An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. *Ecology*. DOI: 10.1890/15-0846R.1.

Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H., and Woodfin, R. M. (1994). Declining biodiversity can alter the performance of ecosystems. *Nature*. DOI: 10.1038/368734a0.

Nakagawa, S. and Freckleton, R. P. (2008). *Missing inaction: the dangers of ignoring missing data*. DOI: 10.1016/j.tree.2008.06.014.

Newbold, T. et al. (2015). Global effects of land use on local terrestrial biodiversity. *Nature*. DOI: 10.1038/nature14324.

Nipperess, D. and Wilson, P. (2019). *PDcalc: An implementation of the Phylogenetic Diversity (PD) calculus in R*. R package version 0.3.2.9000.

Novosolov, M., Raia, P., and Meiri, S. (2013). The island syndrome in lizards. *Global Ecology and Biogeography*. DOI: 10.1111/j.1466-8238.2012.00791.x.

Novosolov, M., Rodda, G. H., North, A. C., Butchart, S. H., Tallowin, O. J., Gainsbury, A. M., and Meiri, S. (2017). Population densityâĂŞrange size relationship revisited. *Global Ecology and Biogeography*. DOI: 10.1111/geb.12617.

Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C., and Costa, G. C. (2017). AmphiBIO, a global database for amphibian ecological traits. *Scientific Data*. DOI: 10.1038/sdata.2017.123.

Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottolo Marasini, G., Visconti, P., and Rondinini, C. (2013). Generation length for mammals. *Nature Conservation*. DOI: 10.3897/natureconservation.5.5734.

Pagel, M. (1999). *Inferring the historical patterns of biological evolution*. DOI: 10.1038/44766.

Paine, C. E., Beck, H., and Terborgh, J. (2016). How mammalian predation contributes to tropical tree community structure. *Ecology*. DOI: 10.1002/ecy.1586.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, xx, xxx–xxx.

Pearson, R. G. (2006). *Climate change and the migration capacity of species*. DOI: 10.1016/j.tree.2005.11.022.

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., and Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*. DOI: 10.1111/2041-210X.12232.

Ratto, F., Simmons, B. I., Spake, R., Zamora-Gutierrez, V., MacDonald, M. A., Merriman, J. C., Tremlett, C. J., Poppy, G. M., Peh, K. S., and Dicks, L. V. (2018). *Global importance of vertebrate pollinators for plant reproductive success: aÂămeta-analysis*. DOI: 10.1002/fee.1763.

Revell, L. J. (2016). Package âĂŸphytools'. *R topics documented*.

Salo, P., Banks, P. B., Dickman, C. R., and Korpimäki, E. (2010). Predator manipulation experiments: Impacts on populations of terrestrial vertebrate prey. *Ecological Monographs*. DOI: 10.1890/09-1260.1.

Santos, T. (2018). *Package âĂŸ PVR '. Phylogenetic Eigenvectors Regression and Phylogentic Signal-Representation Curve*.

Scharf, I., Feldman, A., Novosolov, M., Pincheira-Donoso, D., Das, I., Böhm, M., Uetz, P., Torres-Carvajal, O., Bauer, A., Roll, U., and Meiri, S. (2015). Late bloomers and baby boomers: Ecological drivers of longevity in squamates and the tuatara. *Global Ecology and Biogeography*. DOI: 10.1111/geb.12244.

Schipper, J. et al. (2008). The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science*. DOI: 10.1126/science.1165115.

Schloss, C. A., Nunez, T. A., and Lawler, J. J. (2012). Dispersal will limit ability of mammals to track climate change in the Western Hemisphere. *Proceedings of the National Academy of Sciences*. DOI: 10.1073/pnas.1116791109.

Schwarz, R. and Meiri, S. (2017). The fast-slow life-history continuum in insular lizards: a comparison between species with invariant and variable clutch sizes. *Journal of Biogeography*. DOI: 10.1111/jbi.13067.

Slavenko, A., Tallowin, O. J., Itescu, Y., Raia, P., and Meiri, S. (2016). Late Quaternary reptile extinctions: size matters, insularity dominates. *Global Ecology and Biogeography*. DOI: 10.1111/geb.12491.

Spooner, F. E., Pearson, R. G., and Freeman, R. (2018). Rapid warming is associated with population decline among terrestrial birds and mammals globally. *Global Change Biology*. DOI: 10.1111/gcb.14361.

Stark, G., Tamar, K., Itescu, Y., Feldman, A., and Meiri, S. (2018). Cold and isolated ectotherms: drivers of reptilian longevity. *Biological Journal of the Linnean Society*. DOI: 10.1093/biolinnean/bly153/5145102.

Stekhoven, D. J. (2016). Nonparametric Missing Value Imputation using Random Forest. *R Package version 1.4*. DOI: 10.1093/bioinformatics/btr597.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*. DOI: 10.1093/bioinformatics/btr597.

Stuart, S. N., Chanson, J. S., Cox, N. A., Young, B. E., Rodrigues, A. S. L., Fischman, D. L., and Waller, R. W. (2004). Status and trends of amphibian declines and extinctions worldwide. *Science*. DOI: 10.1126/science.1103538.

Swenson, N. G. (2014). Phylogenetic imputation of plant functional trait databases. *Ecography*. DOI: 10.1111/j.1600-0587.2013.00528.x.

Thompson, P. L., Isbell, F., Loreau, M., O'connor, M. I., and Gonzalez, A. (2018). The strength of the biodiversity-ecosystem function relationship depends on spatial scale. *Proceedings of the Royal Society B: Biological Sciences*. DOI: 10.1098/rspb.2018.0038.

Tilman, D. and Downing, J. A. (1994). Biodiversity and stability in grasslands. *Nature*. DOI: 10.1038/367363a0.

Tilman, D., Isbell, F., and Cowles, J. M. (2014). Biodiversity and Ecosystem Functioning. *Annual Review of Ecology, Evolution, and Systematics*, 45.1, 471–493. DOI: 10.1146/annurev-ecolsys-120213-091917.

Titley, M. A., Snaddon, J. L., and Turner, E. C. (2017). Scientific research on animal biodiversity is systematically biased towards vertebrates and temperate regions. *PLoS ONE*. DOI: 10.1371/journal.pone.0189577.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*. DOI: 10.1093/bioinformatics/17.6.520.

Tucker, M. A. et al. (2018). Moving in the Anthropocene: Global reductions in terrestrial mammalian movements. *Science*. DOI: 10.1126/science.aam9712.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45.3, 1–67.

Vidan, E., Roll, U., Bauer, A., Grismer, L., Guo, P., Maza, E., Novosolov, M., Sindaco, R., Wagner, P., Belmaker, J., and Meiri, S. (2017). The Eurasian hot nightlife: Environmental forces associated with nocturnality in lizards. *Global Ecology and Biogeography*. DOI: 10.1111/geb.12643.

Violle, C., Navas, M. L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., and Garnier, E. (2007). *Let the concept of trait be functional!* DOI: 10.1111/j.0030-1299.2007.15559.x.

Violle, C., Enquist, B. J., McGill, B. J., Jiang, L., Albert, C. H., Hulshof, C., Jung, V., and Messier, J. (2012). *The return of the variance: Intraspecific variability in community ecology.* DOI: 10.1016/j.tree.2011.11.014.

Wandrag, E. M., Dunham, A. E., Miller, R. H., and Rogers, H. S. (2015). Vertebrate seed dispersers maintain the composition of tropical forest seedbanks. *AoB PLANTS.* DOI: 10.1093/aobpla/plv130.

Wilman, H., Belmaker, J., Simpson, J., Rosa, C. de la, Rivadeneira, M. M., and Jetz, W. (2014). EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology.* DOI: 10.1890/13-1917.1.

Wilson, E. E. and Wolkovich, E. M. (2011). *Scavenging: How carnivores and carrion structure communities.* DOI: 10.1016/j.tree.2010.12.011.

Zhang, J., Qian, H., Girardello, M., Pellissier, V., Nielsen, S. E., and Svenning, J. C. (2018). Trophic interactions among vertebrate guilds and plants shape global patterns in species diversity. *Proceedings of the Royal Society B: Biological Sciences.* DOI: 10.1098/rspb.2018.0949.