LONDON'S GLOBAL UNIVERSITY

UCL

University College London
Department of Genetics, Evolution and Environment

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

Adrienne Etard

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

March 15, 2019

Abstract

Contents

List of Tables

List of Figures

List of abbreviations

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

1 Introduction

Anthropogenic activities are driving global biodiversity declines at unprecedented rates. Currently, habitat conversion and degradation – induced mainly by anthropogenic land-use change – are the primary causes of biodiversity loss (Pereira, Navarro and Martins, 2012; Newbold et al., 2015). Climate change is projected to be one of the biggest driver of biodiversity loss by 2070, matching or exceeding the deleterious impacts of land-use change on ecological communities (Newbold, 2018). Understanding how land-use and climate change (LUCC) act on biodiversity, separately and in combination, is key to project the future responses of species, and to consequently put into place efficient policies for biodiversity conservation. Furthermore, biodiversity losses affect ecosystem properties, and can adversely impact the delivery of ecosystem services (REF). Investigating if and how biodiversity decreases link to the loss of ecosystem functions is a key research area (Petchey and Gaston, 2006; Lefcheck et al., 2015) and can help mitigate the impacts of anthropogenic activities on ecosystem processes and services. It has now been established across diverse taxonomic groups that species traits mediate species responses to environmental changes, notably to LUCC (Newbold et al., 2013; Pearson et al., 2014; Pacifici et al., 2017; Estrada et al., 2018). McGill et al. (2006) defined traits as characteristics of organisms, measurable at the level of an individual across species. This definition can be broadened to include "ecological" traits, where species relation to their surrounding environment needs to be considered. Functional traits are those that particularly influence organismal fitness. Functional traits relate to species' abilities to exploit their biotic and abiotic environment and as such, shape ecosystem processes. Functional traits underpin both species' aptitudes to cope with environmental changes and their role in ecosystem functioning (Díaz et al., 2013). Specifically, 'response traits' affect species responses to disturbances, while 'effect traits' shape ecosystem processes. Certain traits can act as both effect and response traits. Conceptually, these are particularly interesting for investigating the impact of environmental changes on ecosystem processes and services, as they provide a mechanistic understanding of how stressors affect both species' responses and ecosystem processes (Luck et al., 2012; Hevia et al., 2017). Assessing the impacts of human activities on ecosystem functioning is increasingly important as pressures rise globally. Publications linking drivers of change and delivery of ecosystem services have increased exponentially since 2001 (Hevia et al., 2017); nevertheless, how species traits influence their responses to land-use and climate change, and how this relates to the loss of important ecosystem functions, remains to be largely explored. The aim of this PhD project is to investigate the effects of terrestrial vertebrate species' traits on their responses to LUCC, at global scales. Specifically, my main goals are (1) to elucidate which traits are likely to put species at greater risk from land-use and climate change; (2) to investigate whether future biodiversity declines triggered by these anthropogenic threats are likely to disrupt important ecosystem functions. Unlike previous published studies, this work will investigate these questions at a global scale, and simultaneously across the four terrestrial vertebrate classes – amphibians, birds, reptiles and mammals. This report synthetizes the work I have achieved throughout the first year. I start by briefly reviewing the literature to present the questions I have addressed in the context of the past and current ecological research (Chapter 1). Chapter 2 exposes the methods and results of data collection. Chapter 3 investigates how the functional diversity of vertebrate communities is affected by land-use change. Finally, I present an outline of the questions that I plan to investigate in the upcoming years.

2 | Literature review

2.1 Land-use and climate change, species traits and the functional composition of vertebrate communities

Currently, terrestrial land-use change is the most important driver of biodiversity declines (Newbold et al., 2015; Chaudhary and Kastner, 2016). With climate change projected to be catching up by 2070 (Newbold, 2018), it has become vital to understand how these threats will affect biodiversity, separately and in combination. By influencing species responses to environmental changes, response traits can provide a mechanistic understanding of how diverse threats shape ecological communities, an understanding particularly relevant for conservation policies. There is now empirical evidence across taxonomic groups that species traits influence their responses to LUCC. For instance, response traits to LUCC have been identified in terrestrial plant (Díaz, Noy-Meir and Cabido, 2001), fungal (Koide et al 2013), invertebrate (Williams et al., 2010; Hall et al., 2019), and vertebrate assemblages (Table 1). Overall, these studies, conducted on different taxa and at different scales, tend to show that larger, longer-lived specialist species with a lower reproductive output are more likely to be impacted negatively by LUCC (Table 1). Nevertheless, it is important to point out that in some cases, contrasting results are found (REF); this highlights the fact that studies may be context-dependent, with contingent limitations. As response traits determine whether a species is likely to be removed from a community due to the environmental filtering exerted by a pressure, changes in the distribution of trait values are expected in vertebrate communities that have faced LUCC. As such, some studies have documented changes in the trait composition of vertebrate communities that have faced climate change or alongside land-use gradients. Some studies summarise changes in the trait community composition using functional diversity indices (Flynn et al., 2009), which describe the diversity and variation of trait values across organisms. Other studies document changes in the distribution of values for particular traits (Rapacciuolo et al., 2017; La Sorte et al., 2018), which can be seen as a particular use of functional diversity indices when only one trait is considered. All these studies show that (1) anthropogenic LUCC is reshaping the functional composition of ecological assemblages, potentially disrupting important functions; (2) species sensitivity to LUCC depends on their traits. Nevertheless, despite this empirical evidence, there is still a need to refine our understanding of which traits significantly influence responses. As traits are commonly used to assess species vulnerability to threats or extinction risks (Pacifici et al., 2015; Willis et al., 2015; Bohm et al., 2016), it is particularly important to be confident about how they act on species responses. The interest for trait-based approaches highlights their potential to inform conservation policies. Trend-based approaches require important field work effort to monitor species populations. Getting extensive information on all species population trends is virtually impossible. The appeal of trait-based approaches is that, by providing mechanistic insights, they diminish the amount of population information needed. If species' responses to a threat consistently relate to certain traits, it is possible to generalise patterns across species for which data is less available (Verberk, van Noordwijk and

Hildrew, 2013). Nevertheless, for several reasons that I now expose, how species traits influence their responses to LUCC remains unclear. First, there is a lack of comprehensive understanding about which traits are important in shaping species responses to climate change. Wheatley et al. (2017) compared different published climate change vulnerability assessment frameworks, some of which trait-based, some trend-based, and some incorporating elements of both (hybrid). They found that the different frameworks, applied to the same set of species, did not yield consensual outputs and classified species inconsistently into different risk categories. Their work underlines that currently, trend-based vulnerability assessments perform better at identifying species at risks from climate change than trait-based approaches. This study highlights the current lack of unanimous understanding as to which traits to consider, and how, in vulnerability assessments. More broadly, their study stresses the need to clarify our understanding of how response traits to climate change act across different taxa. Wheatley et al. (2017)'s finding that there is no consensus across assessment frameworks might be explained by the fact that frameworks were initially designed and tested for a particular taxon – generally at the class level or lower ranks –, and do not hold when applied to other taxa. They nevertheless argue that frameworks should be universally applicable. Their findings put into question to our current ability to extrapolate the knowledge of response traits gathered for certain taxa to other taxonomic groups. To my knowledge, comparative studies looking at whether response traits to LUCC differ across taxonomic groups (at ranks higher than class), experiencing the same threat levels under similar conditions, are rare. The picture becomes even more complex when different studies find contradicting results within a taxon, such as was the case in some invertebrate species (larger body size having been found to influence species responses to land-use change both negatively (ref) and positively (ref)). The work by Bartomeus et al. (2018) further emphasises the idea that unless similar response traits to a threat are identified consistently across different systems and taxa, our ability to use traits as predictors of vulnerability or extinction risk remains limited. For these reasons, it is necessary to conduct comparative analyses across taxa, to identify response traits, verify whether they are conserved across species and whether they have the same importance in shaping responses across taxonomic groups and geographical areas. Second, another difficulty when identifying response traits is that different threats can be acting on the studied ecological community, so that observed modifications stem from the interactions of diverse response traits (Gonzalez-Suarez, Gomez and Revilla, 2013). Response traits must be identified for a single threat while controlling for others, before investigating potential interacting effects. Nevertheless, this is difficult to achieve when using global empirical data. Moreover, the importance of response traits may vary geographically. To conclude, potential taxon-, threat- and geographical dependence of response traits to land-use or climate change makes it difficult to generalise patterns observed at local scales. This stresses the need to conduct global, cross-taxon studies to verify whether empirical evidence supports the generalisation of any response trait.

2.2 Land-use and climate change, functional diversity and the disruption of ecosystem services

Response traits allow to understand and predict how environmental pressures are likely to modify ecological assemblages (changes in species richness and abundance). These alterations can lead to modifications in functional diversity (the diversity and variability of traits in a community). Functional diversity indices are interesting for at least two reasons. First, they can inform on how disturbances affect trait community composition. Second, as functional traits relate to ecosystem functions, measures of functional diversity can relate to ecosystem functioning. I will develop these two points in more detail further down.

2.2.1 Impact of land-use on the functional richness – species richness relationships

Several indices have been developed in the recent years to estimate diverse components of functional diversity (Schleuter et al., 2010). Functional richness, functional divergence (or dispersion) and functional evenness constitute cornerstone metrics that each quantify different aspects of functional diversity. Functional richness is conceptually similar to species richness but aims at reflecting the number of individual functional units across species in a community rather than quantifying the number of singular species. In other words, functional richness metrics estimate the amount of niche space that species occupy (Carmona et al., 2016). Several metrics have been developed to quantity functional richness alone (Legras, Loiseau and Gaertner, 2018). In this work, I use the dendrogram-based index developed by Petchey and Gaston (2002). Other indices. Functional richness indices have been shown to covary with species richness. In experimental studies and natural communities, a positive correlation between these metrics is often found (Petchey and Gaston, 2002). For this reason, examining whether functional richness indices inform on community dynamics differently from species richness is an important question to elucidate. Indeed, if species richness is as informative as functional richness, the latter is not worth measuring: species richness is then a proxy for functional richness. This question was at the heart of the study conducted by Cadotte, Carscadden and Mirotchnick (2011). By reviewing the literature, they found that the relationship between functional richness and species richness is context dependent, and that the shape of the relationship notably depends on the amount of functional redundancy in the community. In communities with a high degree of functional redundancy, functions can be maintained despite species loss. On the other hand, the loss or gain of functionally diverse species can lead to marked variations in functional richness (Figure). As anthropogenic land-uses globally negatively impact local species richness (Newbold et al., 2015), decreases in functional richness of local ecological communities are likely to take place, particularly in communities with low functional redundancy. Flynn et al. (2009) showed that the functional richness of bird, mammal and plant communities located in the Western hemisphere decreased because of agricultural intensification. In other words, land-use intensification impacted the functional richness-species richness relationship. Mayfield et al. (2010) also showed that the relationship between species richness and functional richness could be affected in different ways by human land-uses. They proposed diverse mechanisms building upon community assembly processes to explain how land-uses may influence species richness – functional richness trajectories. The recent development of functional indices, synthesising the diversity of functions in a community, reflects the importance of understanding how anthropogenic pressures will modify ecosystem processes. In the field of biodiversity-ecosystem functioning relationships, it is now well established that higher species diversity is associated with higher ecosystem productivity and stability, better use of limiting resource, as well as better resistance to biological invasions (Tilman at al 2014). I now explore the links between functional diversity indices and ecosystem functioning in more details.

2.2.2 Links between functional diversity and ecosystem functioning

Early experiments investigating the relationships between functional composition and ecosystem functioning classified species in broad functional groups; ecosystem functioning was measured in various ways, depending on the studied system. A higher number of functional groups was correlated to better ecosystem performance in several studies (References). The development of other functional diversity indices allowed for comparisons of diverse predictors of ecosystem functioning. Studies conducted on various systems showed that functional richness performed better at predicting ecosystem functioning than taxonomic richness (Díaz and Cabido, 2001; Flynn et al., 2011; Abonyi, Horváth and Ptacnik, 2018). All these results led to functional diversity emerging as an important predictor of ecosystem functioning. The use of functional diversity indices as predictors of ecosystem functioning raises two important points. First, ecosystem functioning should be clearly defined within the study system. Second, the functional traits, from which the processes of interest originate, should be identified and included in the calculation of functional indices. Cadotte, Carscadden and Mirotchnick (2011) underline that point; they highlight that traits should be linked with ecosystem functions for functional diversity indices to be useful. In other words, only relevant effect traits should be considered. Moreover, a larger number of effect traits can lead to higher functional differentiation among species, with potential impacts on the metrics. A careful selection of effect traits is thus vital for functional diversity indices to be informative on ecosystem processes. If effect traits inform on ecosystem processes, response traits mediate species responses to environmental change. As such, the link between environmental pressures and ecosystem functioning is conceptually realised with both response and effect traits. Specifically, the purpose of the "response-effect" framework is to is to understand how environmental changes alter ecosystem functioning by disentangling traits that render species sensitive to a threat (response traits) from traits that impact functioning (effect traits). A modification in the composition of a community could affect ecosystem functioning in two ways: functions can be lost directly through the removal of species, triggered by response traits (nestedness); and indirectly, functions can be affected by the resulting shift in composition (turnover). The response-effect framework relies on identified response traits to provide a mechanistic understanding of how disturbances modify the trait composition of communities, and how these changes link to alterations in functioning. Changes in functioning are driven by changes in the effect trait composition (effect traits being those that are involved in ecosystem functioning). When response and effect traits are similar, changes in ecosystem functioning are predicted by changes in species composition (direct effects of species loss or gain). In that case, overall functional diversity correlates with ecosystem functioning. However, when response and effect traits are decoupled, changes in functioning are defined by the shifts in effect trait composition only (indirect effects of species loss or gain). The application of the response-effect framework to real animal communities has been hindered by several issues (Luck et al 2012; Bartomeus). Luck et al. (2012) proposed a new trait-based framework to link environmental changes to ecosystem services in vertebrate species. They underline the need to develop robust and broadly applicable methods. Efforts to link drivers of change and ecosystem function responses have been disparate across taxonomic groups, with a major focus on plants and invertebrates in the past years. Indeed, Hevia et al. (2017) showed in a metanalysis that most studies investigating how species traits mediate the impacts of stressors on ecosystem processes focused on plants and invertebrates, such that there is an existing taxonomic bias in this area. Vegetation and invertebrates both represented an approximate 40% of the sampled papers, whereas only 17% were dedicated to vertebrates. Their metanalysis also shed light on other biases, such as the spatial scale of the papers, with most sampled studies being conducted at local or national scales. Therefore, although terrestrial vertebrates have a major cultural, economic and functional importance (Hocking, Babbitt and Hocking, 2014; Whelan, Şekercioğlu and Wenny, 2015; Ratto et al., 2018) and are over-represented in the overall biodiversity literature compared to other taxa (Titley, Snaddon and Turner, 2017), how disturbances affect the services they provide has not been extensively explored compared to other taxa. Efforts to understand ecosystem services provided by terrestrial vertebrates have mainly focused on pest control, seed dispersion, and protein provisioning. To understand how anthropogenic pressures may impact ecosystem processes sustained by vertebrate communities at global scales, there is a need to assess whether LUCC significantly affects the functional diversity of vertebrate communities, and, in particular, the effect trait composition; and to verify whether effect trait composition predicts ecosystem processes, and is, as such, a relevant measure for conservation and mitigation.

To conclude, examining how vertebrate species traits influence their responses to LUCC is the first step to (1) elucidate which traits are likely to put species at greater risk, and find out whether it is possible to generalise patterns across vertebrate species; (2) investigate whether future biodiversity declines triggered by these anthropogenic changes are likely to disrupt important ecosystem functions. The work I have achieved so far focuses on land-use change at global scales and aims at investigating the questions detailed below.

2.3 Research questions and hypotheses

As underlined in the introduction, functional traits can provide a mechanistic understanding of how environmental stressors affect both ecological assemblages and ecosystem processes. As such, they convey information most relevant to conservation policies. According to Hekkala and Roberge (2018), global assessments of how land-use change affects vertebrate functional diversity may have been limited so far by the amount of ecological information required to conduct such analyses, notably by the availability of species traits. There exist published databases of species traits (e.g. Pantheria: Jones et al., 2009; Myhrvold et al., 2015; AmphiBIO: Oliveira et al., 2017) but despite these collation efforts, some taxa remain under sampled (Newbold, unpublished manuscript). For this project, I will collate information on vertebrate traits prior to conducting any analysis, and I will impute missing values to increase the trait coverage across species. Because obtaining longitudinal information on compositional changes can be difficult, the effects of land-use change throughout this project will be studied using a 'space for time' substitution, whereby a spatial gradient is used as a proxy for temporal dynamics (De Palma et al., 2018). As such, the following analyses will build upon the PREDICTS database, a large collated dataset of species occurrence and abundance around the world across different land-uses (Hudson et al., 2014, 2017). To date, this database constitutes the most comprehensive global collection of biodiversity samples across different land-uses. It comprises 666 studies, each of which recording the occurrence and/or abundance of species at different sites. Each site is classified into a land-use category; the land-use gradient encompasses primary vegetation, secondary vegetation, plantation forest, cropland, pasture and urban. Primary vegetation refers to native vegetation undisturbed since its development under current climatic conditions. Where primary vegetation was destroyed (either by human actions or natural causes), recovering vegetation forms are referred to as secondary vegetation. Plantation forest, cropland and pasture refer to agricultural areas (crop trees grown for human purposes, biofuels and herbaceous crops, and areas grazed by livestock). Using this database, I aim to address the two following questions.

2.3.1 How does land-use affect the functional diversity of vertebrate communities?

In this part of the project, I aim to investigate how land-use change affects the trait composition of vertebrate communities. I hypothesise that by reducing local habitat heterogeneity, human-dominated land-uses promote functional homogenisation, whereby the similarity in trait composition across assemblages increases through the loss of certain functions (H.1). This hypothesis relies on the idea that strong environmental filtering will disproportionately remove certain functional types. To test this hypothesis, I will use various indices of functional diversity and functional β-diversity. First, I will calculate three cornerstone indices of functional diversity. For each PREDICTS site, I will measure functional richness, functional divergence and functional evenness. Specifically:
- I expect functional richness to decrease in more human-dominated land-uses, with habitat filtering reducing the amount of utilised trait space. Because functional richness is often correlated with species richness, I will investigate whether, for a given species richness SR, the functional richness

FR is predicted to be lower for sites under higher human land-use intensity (H.1.1). Assuming SR and FR are positively correlated for each land-use type, I expect the slope of the relationship to be lower for sites located within human-dominated land-uses (H.1.1).
- I also expect functional evenness and divergence to decrease along the land-use gradient, with more species within human-dominated land-use communities over-utilising central parts of the functional trait space (H.1.2). I expect to observe a convergence of trait values.

I will test whether observed effects are consistent across vertebrate classes and contingent on the geographical area. These indices will give insights into how land-use alters overall functional diversity across PREDICTS studies. Nevertheless, such indices are insensitive to changes in functional composition. Indeed, a decrease in functional richness could be explained by either functional turnover, whereby certain trait values are replaced by others, or by functional nestedness, whereby certain functions are lost (Figure 1.). Functional diversity indices will therefore not be informative as to what drives changes in functional composition.

To further investigate this point, I will use functional β-diversity measures. β-diversity indices allow to quantify the dissimilarity in functional composition across sites. Baselga (2010) developed an approach to partition total dissimilarity into one component accounting for functional turnover (potentially stemming from species turnover) and one component accounting for functional nestedness (potential loss or gain of functional trait space relating to the loss or gain of species). I expect nestedness to drive decreases in functional richness (H.1.3) with certain functions being disproportionally lost.

2.3.2 Which vertebrate traits confer sensitivity to land-use change?

The goal of this second analysis is to identify response traits to land-use change in vertebrate species. As opposed to the previous analysis, species traits will be used as explanatory variables. I aim to assess the individual effects of traits or trait combinations on species' sensitivity to land-use. Specifically, I hypothesize that, with increasing intensity of human land-use:

- Species with longer generation length are impacted more negatively than shorter-lived species. The rationale behind this hypothesis is that declining population trends in human-dominated land-uses could be compensated more rapidly in species with shorter generation lengths.

- Larger species (higher body masses) are impacted more negatively than smaller species. Indeed, a general ecological rule states that species with higher body masses have lower local population densities (Santini, Isaac and Ficetola, 2018); I hypothesize that higher body masses compromise species' persistence in disturbed habitats as those species will tend to be rarer than species with lower body mass (but see Vermeij and Grosberg (2018)).

- Species with larger litter or clutch sizes respond less negatively to land-use change than species with smaller litter or clutch sizes. Indeed, declining population trends in human-dominated land-uses could be compensated more in species with larger reproductive outputs.

- Specialist species, that have stricter requirements either in their habitat preferences or in their diet, are impacted more negatively than generalists. Indeed, such species could be more dependent on particular food sources or habitats that may be impacted negatively by land-use changes.

- Narrowly distributed species are impacted more negatively than species with larger range sizes. I hypothesize that narrowly distributed species have stricter habitat requirements and lower breadth in the dimensions of their fundamental niche, making them less able to cope with altered habitats than more broadly distributed species.

3 | Collecting and imputing ecological trait data across terrestrial vertebrates

3.1 Introduction

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

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

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

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

The present chapter details the methodology I employed to collate selected traits. I elaborate on some of the challenges met when compiling data across many species, such as inconsistency of taxonomy across sources. Not unexpectedly, the amount of missing values was highly variable across classes and traits. To achieve full trait coverage across species, I imputed missing trait values using random-forest algorithms. In this chapter, I briefly examine imputation performance, notably by assessing whether increasing species representation in phylogenetic trees affects imputation error.

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

Finally, the trait data collected and imputed in this chapter can be subject to future changes, and may not final at this stage.

3.2 Methods

3.2.1 Ecological trait data collection

Primary data sources.

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

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

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

Compilation methods.

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

Categorical traits.

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

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

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

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

3.2.2 Phylogenetic information

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

3.2.3 Tackling taxonomic synonymy

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

Automated procedure and outputs.

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

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

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

Despite the automation efforts, taxonomic redundancy persisted to a degree in the trait datasets. Indeed, at this stage, not all species in PREDICTS matched a species in the trait datasets. Additional

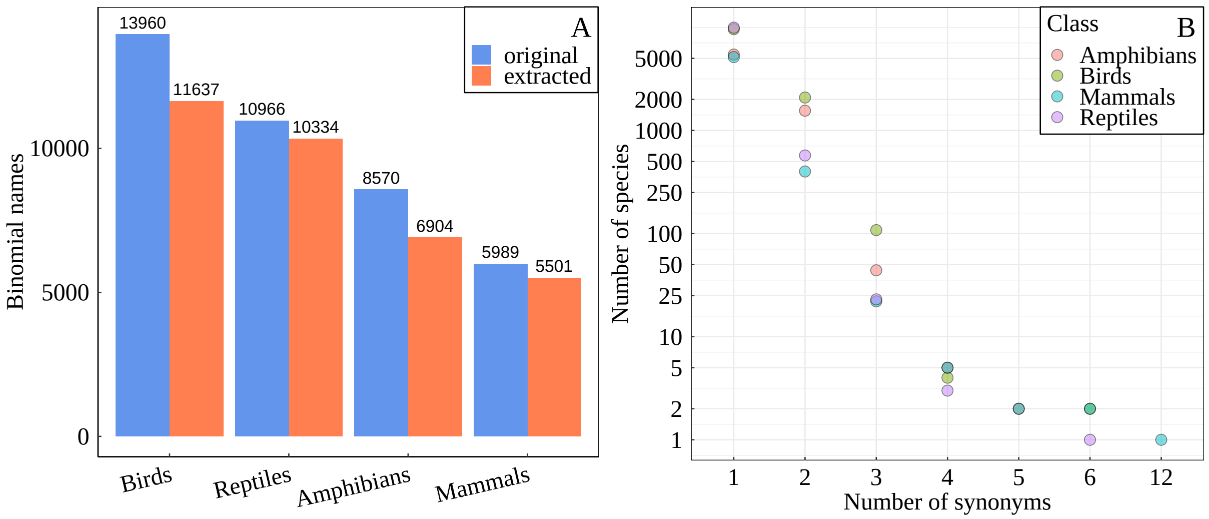

Figure 3.1: Difference in species number due to taxonomic correction (A) and distribution of number of synonyms across datasets (B). (A) shows the number of species across all primary sources (trait datasets and PREDICTS, excluding phylogenies), before and after correcting for taxonomy. Replacing identified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

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 Red List or the ITIS were fully resolved. As I only applied manual checks for PREDICTS relevant species, 'pseudoreplication' and taxonomic errors are likely to have persisted to a degree. Moreover, certain species were entered using the format *Genus subspecies* rather than *Genus species*; for these, automated queries may have failed to identify the species.

Harmonising taxonomy in phylogenetic trees and increasing species phylogenetic representation.

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

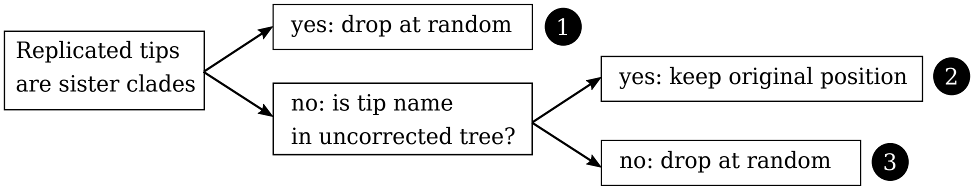
Replicated tips are sister clades
yes: drop at random
1
no: is tip name in uncorrected tree?
yes: keep original position
2
no: drop at random
3

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

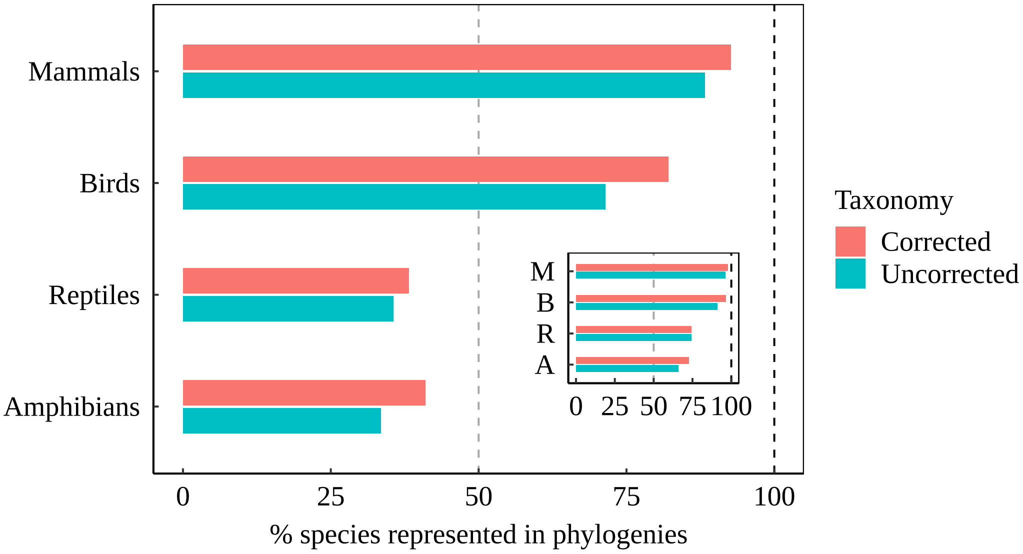

Figure 3.3: Percentage of species represented in the phylogenies for both corrected and uncorrected trait datasets. Overall, taxonomic correction increased species representation in phylogenetic trees. Representation for mammals and birds was high (after taxonomic correction: 82% of avian and 93% of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

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

Species attachments to phylogenetic trees. Some species in the trait datasets were not represented in the phylogenies. Maximising the number of species represented in the phylogenies was important for further trait imputations. Indeed, if traits were evolutionary conserved, species phylogenetic position could be an important predictor of trait values. To maximise species representation, I added some species to the root of their genus, when possible (phytools package). Attaching species at the root of their genus created polytomies, which were resolved randomly (using multi2di, ape package). Resulting trees contained branches of length zero. To facilitate further analyses, a small number (10^{-10}) was added to these branch lengths; consequently, the trees were not ultrametric. were Such a process could have altered the significance and the strength of trait phylogenetic signal. I further verify whether these alterations of the trees had impacted phylogenetic signal, by qualitatively comparing the strength and the significance of phylogenetic signal for each trait estimated using both original trees and augmented trees (see 'Assessing phylogenetic signal in traits').

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

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

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

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

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

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

3.2.5 Imputing missing trait values

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

Assessing phylogenetic signal in traits

Measuring phylogenetic signal in continuous traits with Pagel's λ. Phylogenetic signal is a measure of the tendency of closely related species to resemble each other more than less related species. Diverse statistics have been developed to estimate phylogenetic signal, most of them applying to continuous traits (Munkemuller 2012). I used Pagel's λ (function phylosig, phytools package). Pagel's λ is a scaling component that measures the coefficient by which the trait covariance matrix should be weighted to fit a Brownian motion model of evolution. Indeed, under a Brownian motion model of evolution, the trait covariance matrix is expected to be influenced only by the phylogenetic history: changes in trait values happen at random and trait variance is proportional to evolutionary time. When other factors are at play, the observed covariance matrix is the expected covariance matrix transformed with the estimated λ. A value close to 0 indicates that the covariance matrix need not be transformed by much to fit a Brownian motion. On the other hand, a value close to 1 indicates that trait values are more similar in closely related species than expected under a Brownian motion model of evolution. Using Pagel's λ, I assessed the strength of the phylogenetic signal. The phylosig function (phytools) also allows to test for signal significance (comparing the estimated λ to the null expectation of λ with a log-likelihood ratio test). Note that the function developed by Borges et al does not work if phylogenetic trees contain branches of length 0. As both original and corrected phylogenies contained 0-length branches, I added a very small number to these (10^{-10}) to remedy to this issue. As such, the trees with which δ was computed were not ultrametric.

Measuring phylogenetic signal in categorical traits with δ (Borges et al, 2018). Very few methods have been developed to measure and test phylogenetic signal in categorical traits. Fritz and Purvis (2010) introduced the D-statistic, which only applies to binary traits. Furthermore, D is based on a discretisation of the trait, which behaves as a continuous trait evolving under Brownian motion. Borges et al (2018) introduced a new statistic, δ, to measure phylogenetic signal in categorical traits. δ is based on Shannon entropy principles and uses Bayesian inferences for estimation. δ can take any positive number, with higher values indicating stronger signal. To test for the significance of the signal, the authors propose to compare the estimated value with a null distribution of values. I generated null distributions of δ for each trait by simulating 100 random trait vectors (simulating Brownian motion of trait evolution) and calculating δ for each. 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.

Significant phylogenetic signal in all traits All traits showed significant phylogenetic signal (Table 3.3), although the strength of the signal was

Despite much variation in sample sizes across classes, results indicated strong phylogenetic signal across both categorical and continuous traits (Table 3.3). The signals were all significant (expect for amphibian body mass, but the signal in body length was strong and significant in this class). Signal strength was overall higher for mammals and birds, which may be a consequence of missing value biases. p-values outputs of the likehood-ratio tests and the Wilcoxon rank sum tests are provided in the SI.

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*	$28\cdot10^3$*	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*

I hence imputed missing trait values using random forest algorithms, implemented by missForest. As stated above, missForest was shown by Penone et al (2014) to be the best method when including phylogenetic information for mixed-type variable imputations. Phylogenetic relationships were included as additional predictors in the form of phylogenetic eigenvectors, extracted from the phylogenies using the PVR package (Santos 2018). Penone et al (2014) also showed that includ-

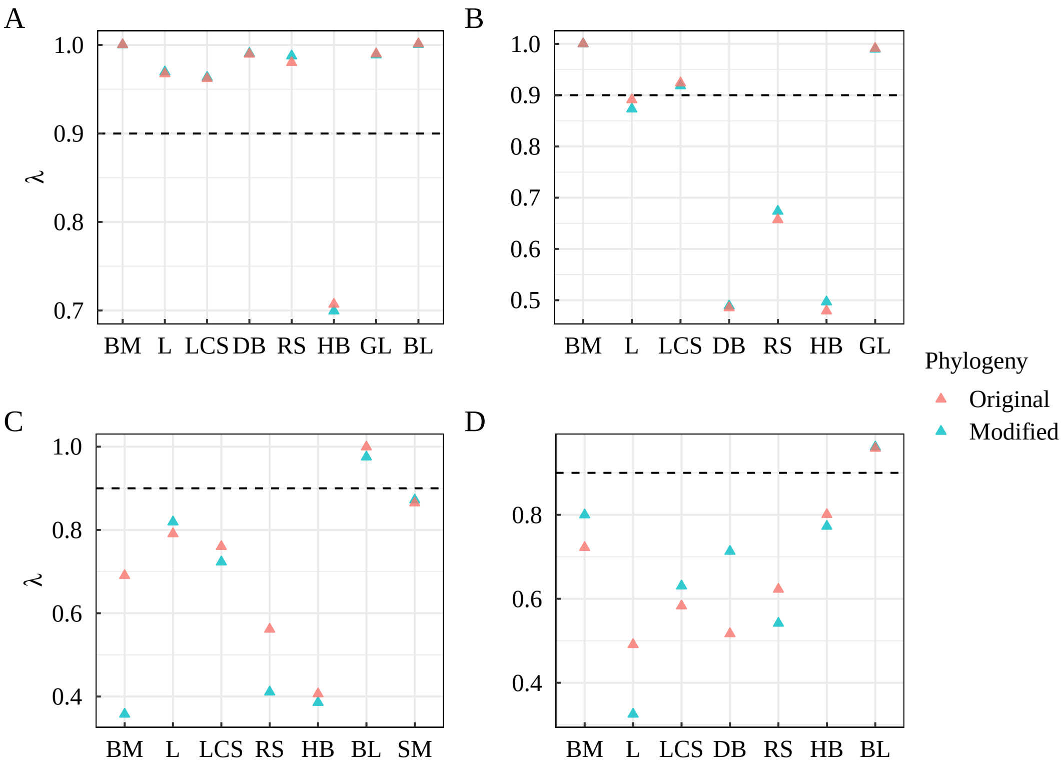

Figure 3.4: Phylogenetic signal in continuous traits(Pagel's λ) estimated with both original phylogenies and modified phylogenies.

ing the first 10 eigenvectors minimised the imputation error. As not all species were represented in the phylogenies (Figure 3.3), phylogenetic eigenvectors presented some missing values. I added taxonomic orders as a predictor variable. All traits in Figure 3.4 were included in the imputations, except for primary diet and diet breadth in reptiles.

Imputation error and robustness

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

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

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

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

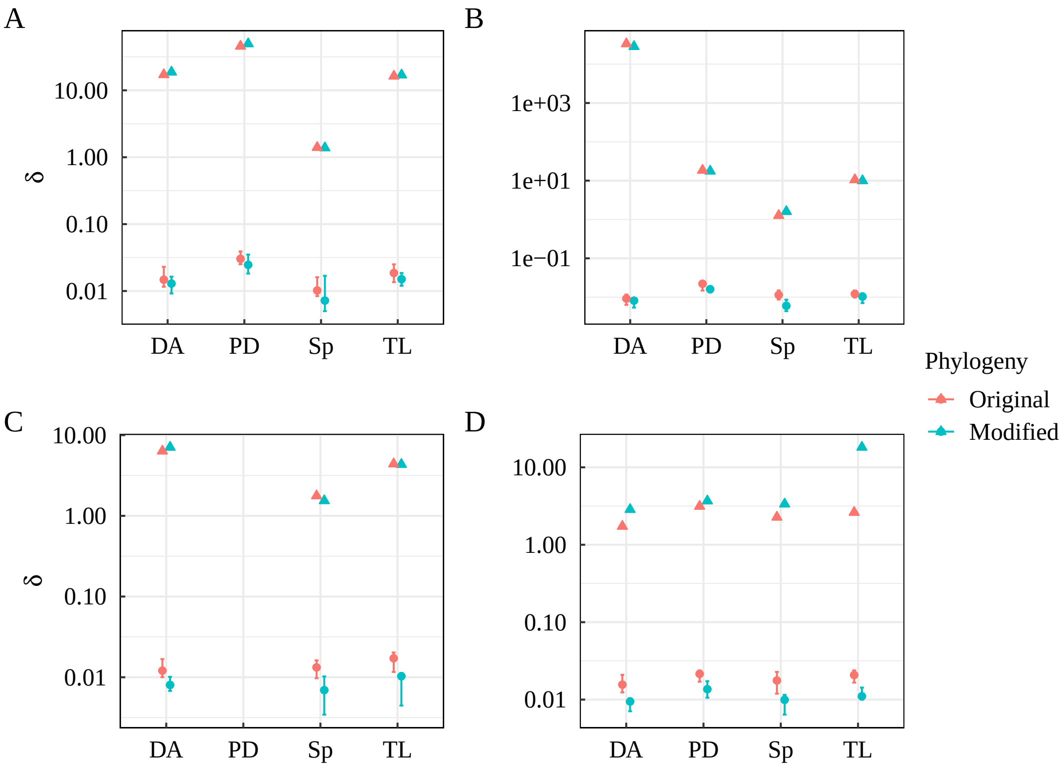

Figure 3.5: Phylogenetic signal in categorical traits (δ) estimated with both original phylogenies and modified phylogenies.

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

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

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

where X_t is a vector of the complete trait values and X_i a vector of the imputed trait values (Stekhoven 2011). For categorical traits, the is calculated as the PFC (Δ_{cat}, Equation 3.2). Imputation performance improves with decreasing error values.

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

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 3.4 shows the trait coverage within each class and for each trait, before and after correcting for taxonomy. Figure 3.5 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 3.4). 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 3.5). Wilcoxon rank sum tests, testing the null hypothesis that uncorrected and corrected completeness distributions came from the same population, rejected this hypothesis across all classes (alternative hypothesis: uncorrected medians were lower than corrected medians; mammals: p-value=$1.2{\cdot}10^{-9}$; birds: p-value<$2.2{\cdot}10^{-16}$; reptiles: p-value=0.025; amphibians: p-value<$2.2{\cdot}10^{-16}$). To conclude, correcting for taxonomy had a significant impact on trait completeness and increased coverage in most cases.

Among-class biases in the availability of trait information

Trait coverage. Trait coverage was highly variable across classes and traits. Trait coverage was initially good for most mammalian and avian traits, which had more than 50% coverage (Figure 3.4 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 3.4 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).

Trait completeness. Trait coverage revealed taxonomic biases, with higher resolution of trait information across mammals and birds. Trait completeness reflected similar biases. (Figure 3.5). 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.

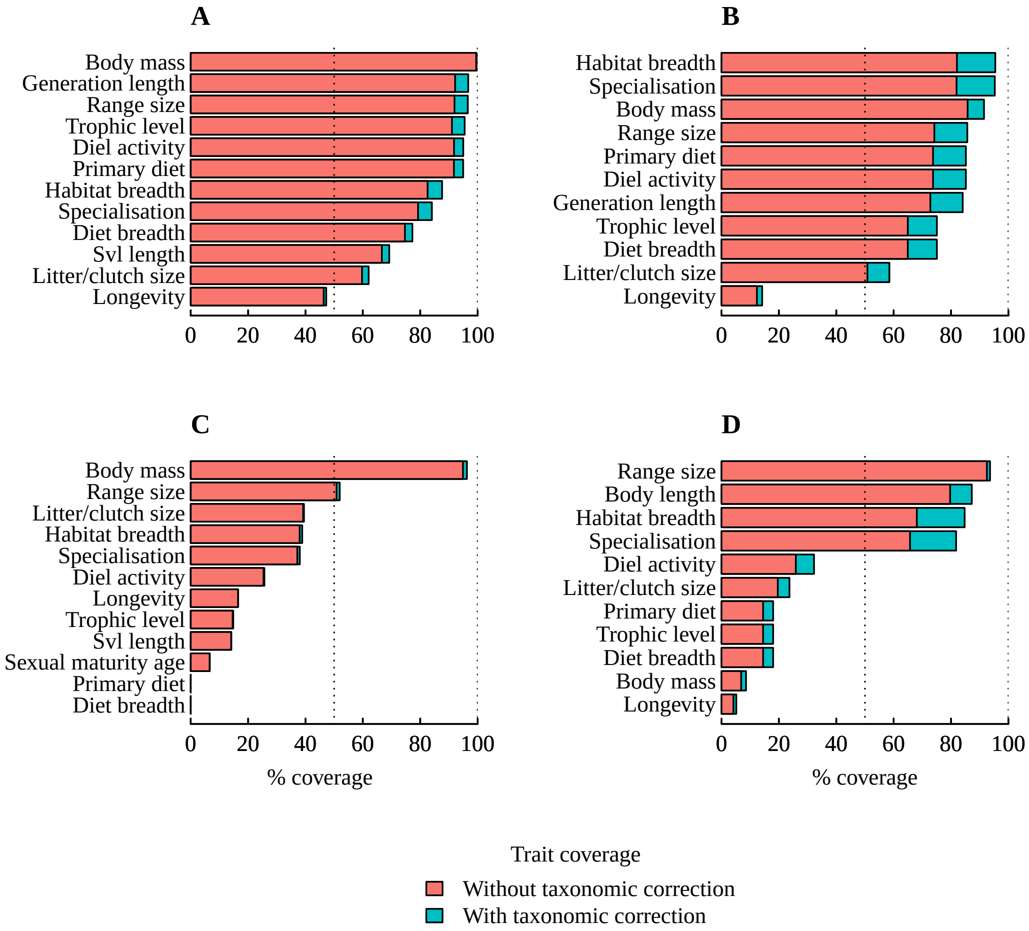

Figure 3.6: Trait coverage across all species before and after taxonomic correction. Here are shown all targeted traits as well as a few other traits used in imputations, as additional predictors (such as generation length for mammals and birds or body length for amphibians). **(A)** Mammals (5885 species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

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

Beyond cross-class biases in the availability of trait information, within-class patterns of missing values showed that certain families were less sampled than others.

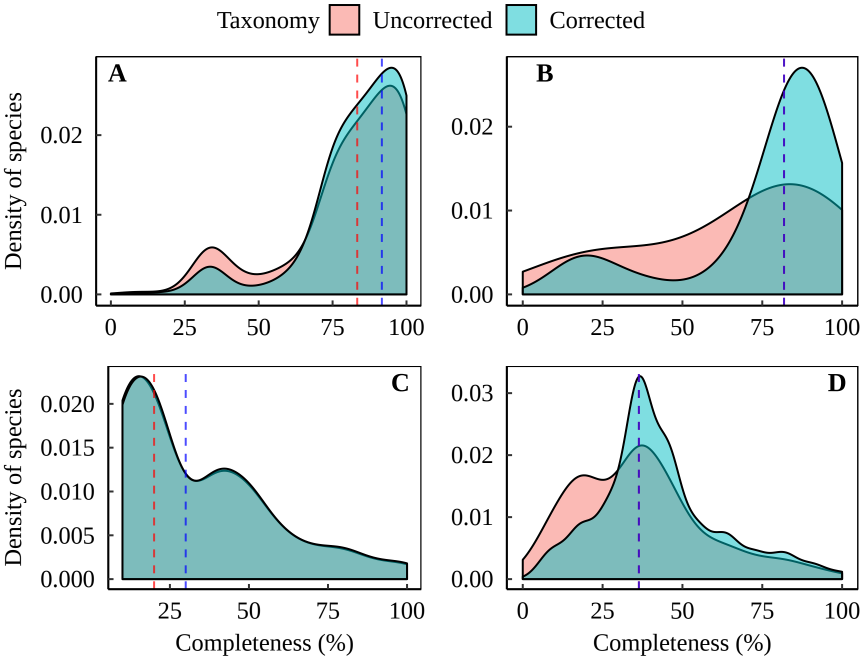

Figure 3.7: Distribution of completeness of trait information across species. **(A)** Mammals; **(B)** birds; **(C)** reptiles and **(D)** amphibians. Completeness was calculated here for the same set of traits shown in Figure 3.4 (all predictor traits). Correcting for taxonomy affected completeness, 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$<2\cdot10^{-16}$ in all cases)).

Within-class patterns of trait coverage

Within-class patterns of trait completeness

3.3.3 Imputation performance and robustness

Out-of-bag imputation errors

Congruence of imputed values among 8 imputed datasets

Comparison with another collected and imputed datasets for mammals and birds

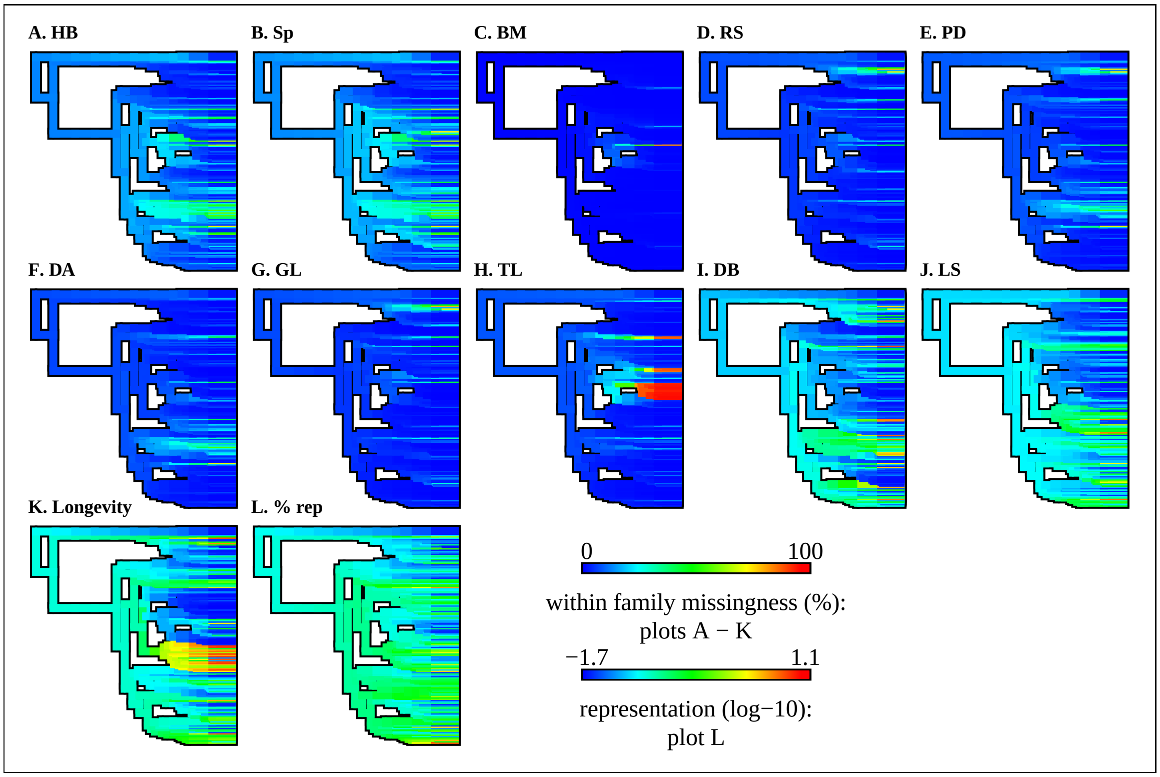

Figure 3.8:

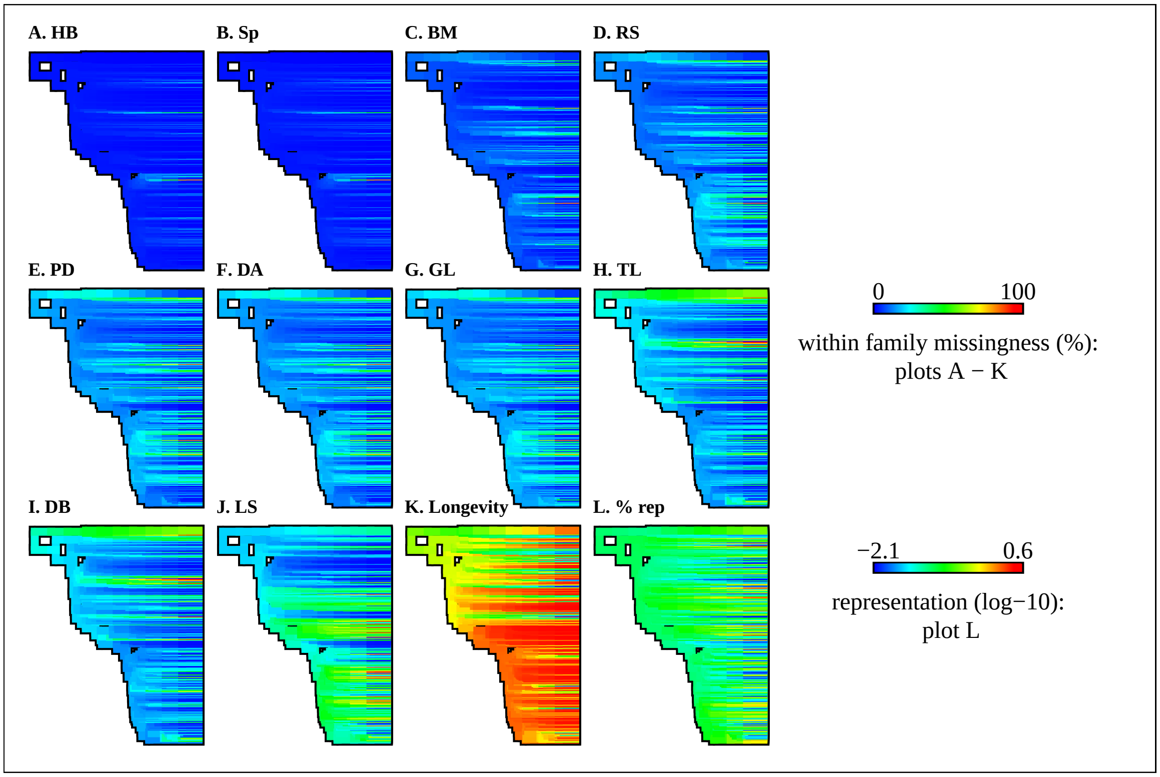

Figure 3.9:

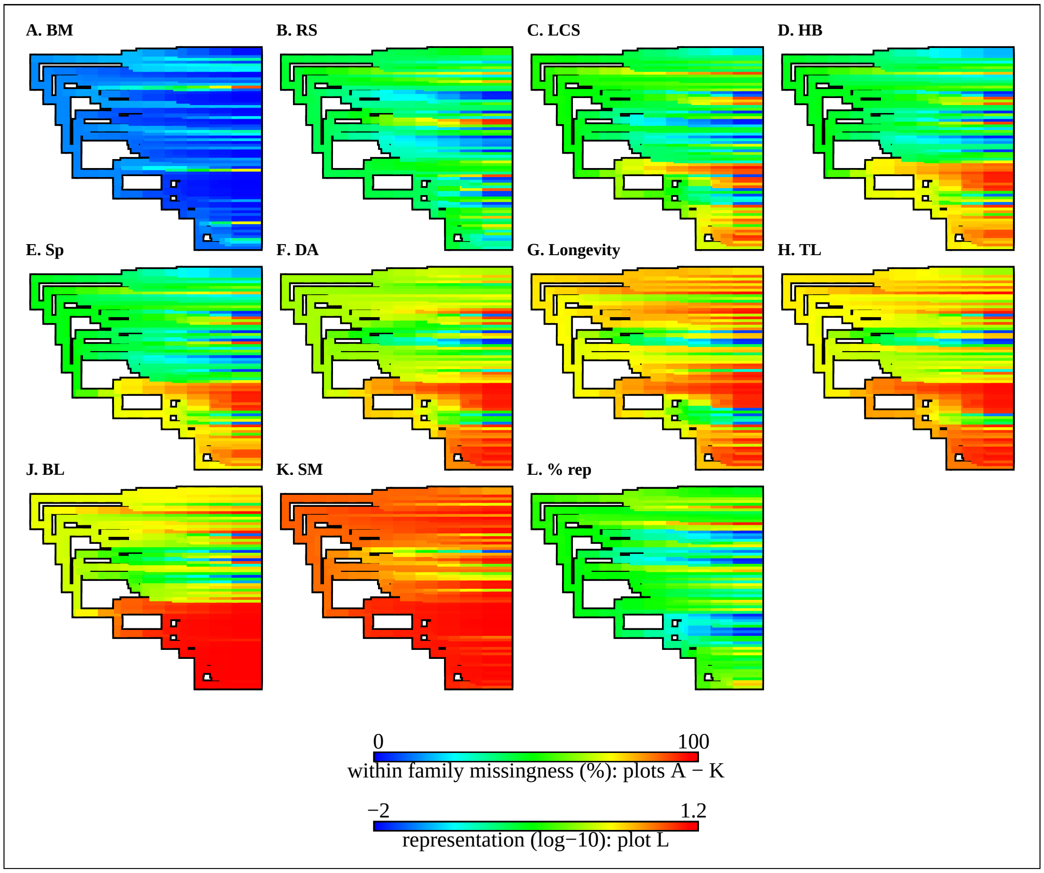

Figure 3.10:

3.4 Discussion

- Taxonomic challenges
- Manipulation of the phylogenies Adding species randomly => justified if traits have a strong phylogenetic signal even with the uncorrected phylogeny. Then adding species makes sense (because strong phylogenetic signal, "trade-off" between the quality of the imputations versus the quality of the phylogenies)
- Biases in availability of trait information across classes, Raunkier shortfall
- Imputation robustness

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

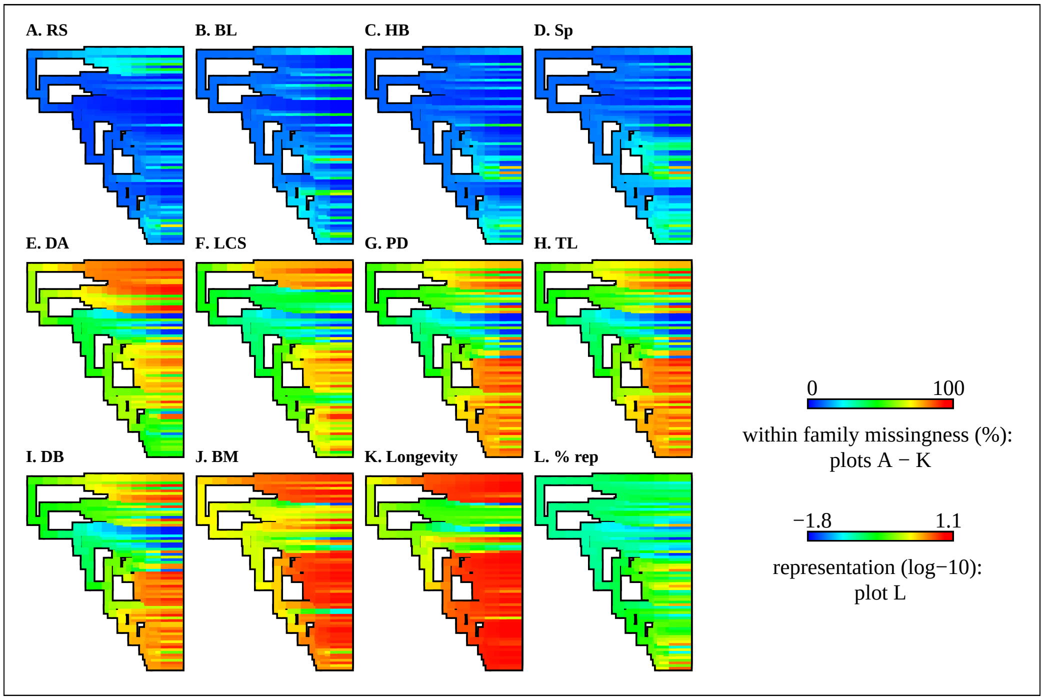

Figure 3.11:

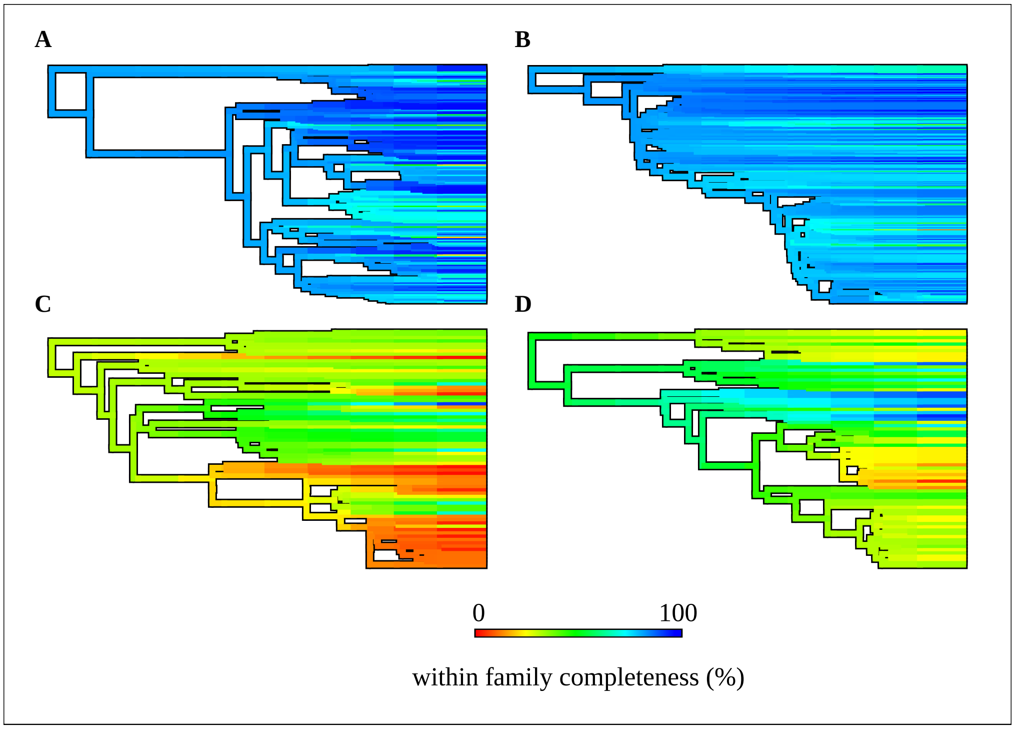

Figure 3.12: Median completeness across families. Tips labels are not shown here for better visualisation of the results; the same figures with tip labels are provided in the SI (zooming into the figure is necessary for mammals and birds); tip label information includes order and family. **(A)** Mammalian 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.

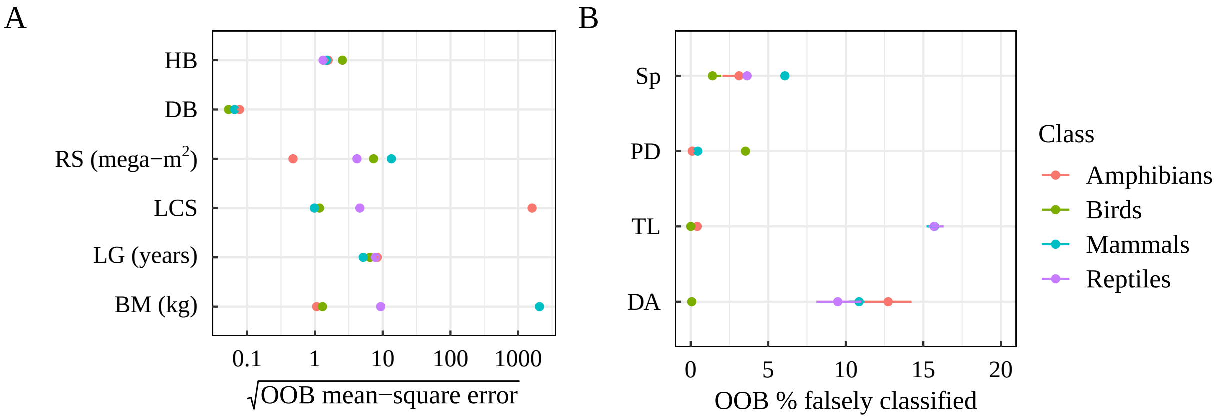

Figure 3.13: missForest mean-square errors and proportion of falsely classified values. (A) Mean-square errors for continuous traits. **(B)** Proportion of falsely classified values.

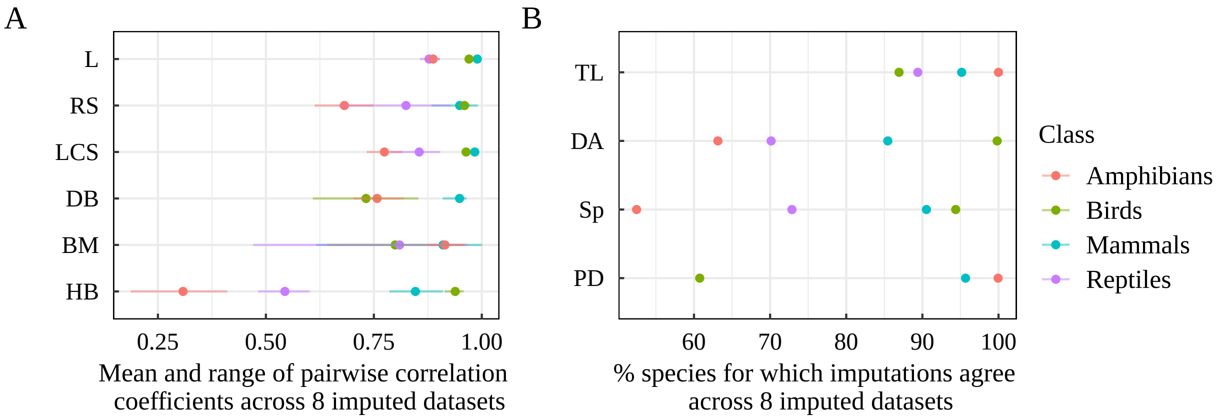

Figure 3.14:

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

5 | Outline and research questions for the next years

6 | Conclusion