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

February 26, 2019

Submitted for Upgrade

Contents

List of Tables

List of Figures

1 Introduction

2 Trait data collection and imputation of missing values

2.1 Introduction

In this chapter, I collected and imputed ecological trait data for terrestrial vertebrates. Although terrestrial vertebrates have been extensively studied, the amount of trait information available in the literature is highly variable across classes and traits (Newbold, manuscript). Moreover, to my knowledge, there exist no comprehensive database of vertebrate traits encompassing all classes at the same time. Nevertheless, past and recent efforts to release data, either on freely accessible platforms or alongside publications, allowed us to compile an extensive list of sources from which to collate trait information.

Here, I collated trait information for as many terrestrial vertebrate species as possible. Targeted traits related to species life history (body mass, longevity, litter or clutch size, trophic level, diet) and to their habitat preferences (habitat breadth, habitat specialisation). Traits were selected for three main reasons: (1) estimates were available for many species; (2) estimates were available across all four terrestrial vertebrate classes, allowing cross-classes comparative analyses (true for all traits except diet); (3) selected traits have been shown to be response or effect traits in other studies or are related to response or effect traits (Table in Chapter 1). After assessing the initial trait coverage for each class, I imputed missing trait values using random forests algorithms.

The present chapter details the methodology employed to collate and impute trait information. I elaborate on some of the challenges met when compiling information across a large number of species, such as inconsistency of taxonomy across sources. I examine the robustness of trait imputations and, for mammals and birds, compare results with another collated dataset (Cooke et al). Indeed, in October 2018, Cooke et al released extensive information for six mammalian and avian traits, presenting us with a good opportunity for comparison.

2.2 Methods

2.2.1 Ecological trait data collection

Sources and compilation methods.

I collated ecological trait data for terrestrial vertebrates from published databases and unpublished sources (Table 2.1).

Table 2.1: Data sources for trait compilation. 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. In bold, the traits of interest. Other traits were added for potential correlations in further imputations.

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

Compilation methods

Continuous traits. All continuous traits (body mass, litter or clutch size, longevity) were averaged within species when different sources provided estimates. Longevity and maximum longevity were assumed to provide the same information and were also averaged within species. In addition, I 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. Finally, species geographical range sizes were calculated from distribution data, extracted from the IUCN Red List.

Categorical traits. Diet was available for all classes except reptiles. Species diet was described as a binary variable recording whether food items were known to be consumed by a species or not. For amphibians and birds, trophic levels were partly inferred from the diet. 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 used these compiled traits to design three other ecological variables. Diet breadth was calculated as the number of food items a species was recorded to ingest. Similarly, habitat breadth was calculated as the number of habitats a species was known to use. Weights were assigned to each habitat in this calculation depending on whether the habitat was recorded to be suitable or marginal for each species; outcomes were not sensitive to different weight choices (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 how diet and habitat variables were compiled are provided in the SI.

2.2.2 Phylogenies

I obtained phylogenetic trees for birds, amphibians, mammals and squamates from Hedges et al (2015) (available at http://www.biodiversitycenter.org/ttol, downloaded 06/07/2018).

2.2.3 Taxonomic synonymy

Extracting synonyms and harmonising taxonomy in trait datasets.

Procedure Across the different sources, similar species could appear under different binomial names. 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. As such, taxonomic synonymy was 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. I attempted to correct for taxonomy by developping 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; when possible, I best assigned binomial names to species common names.

The automated procedure consisted in extracting species accepted and synonymic binomial names from the IUCN Red List or from the Integrated Taxonomic Information System database (ITIS), using the rredlist and taxize R packages. For each class, I started by generating a list of all binomial names figuring across datasets. These 'original' binomial names were corrected for typos using gnr_resolve (taxize R package). For each of these corrected names, the IUCN RedList was queried and synonyms and accepted names were stored. 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). 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 (Table XXX).

Nevertheless, additional manual checks were required to make sure that all vertebrate species appearing in PREDICTS were represented in the trait datasets. Taxonomic synonymy was resolved manually for PREDICTS species that did not match any species in the trait datasets. Information was extracted from other diverse sources (such as The Reptile Database; Avibase; AmphibiaWeb). The need to apply additional manual inputs underlined the fact that the automated procedure to correct for taxonomic synonymy was not optimal. The Red List and ITIS were not comprehensive taxonomic sources for accepted names and synonyms. Species 'pseudoreplication' is likely to have persisted to a degree.

Increase in trait coverage due to taxonomic correction Across all classes, correcting for taxonomy overall increased trait coverage, measured as the percentage of species for which information was not missing. For mammals and birds, intial trait coverage was high. On the other hand, the variability in coverage was much higher for reptiles and amphibians.

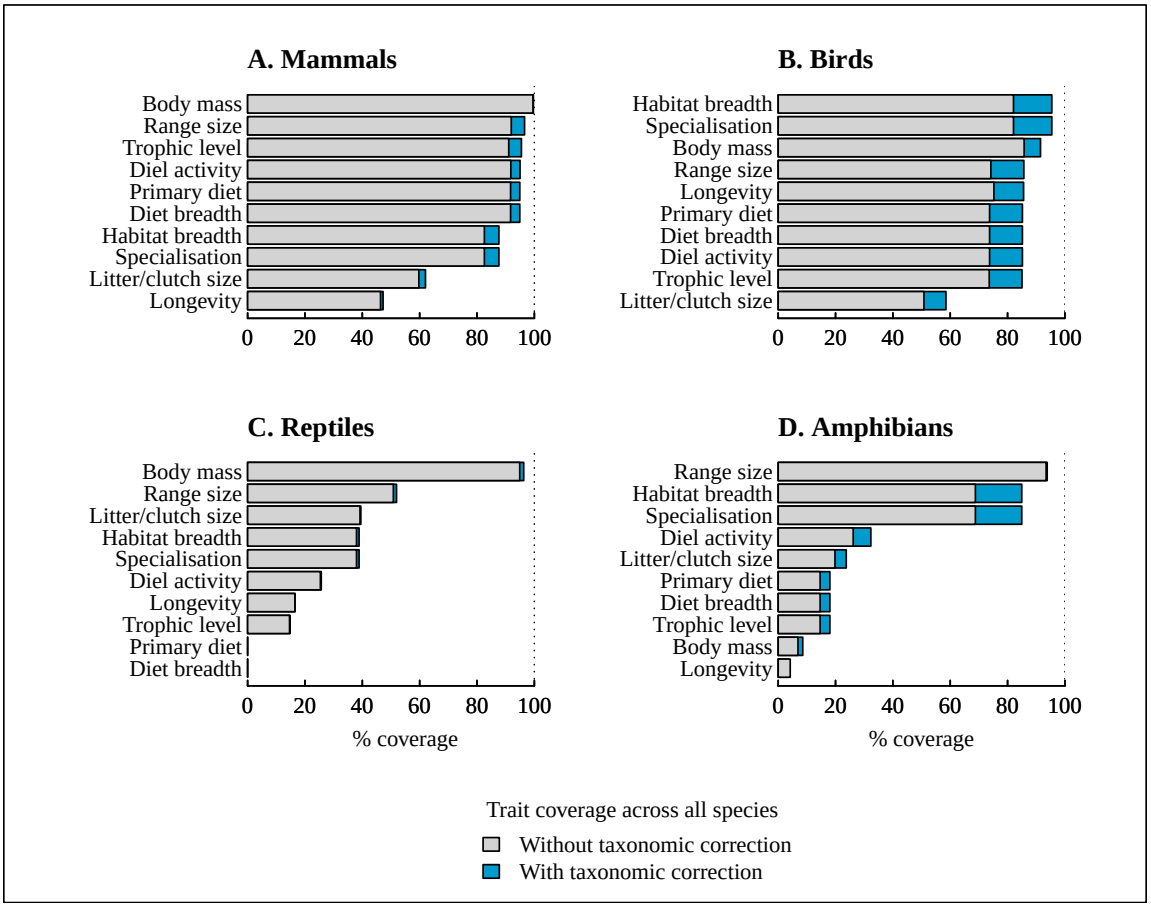

Figure 2.1: Trait coverage across all species before and after taxonomic correction. Trait coverage is defined here as the percentage of species for which trait information is available. Correcting for taxonomic synonymy improved trait coverage in most cases.

For species figuring in PREDICTS, trait coverage disproportionally increased for reptiles and amphibians.

Harmonising taxonomy in phylogenetic trees and increasing species representation.

Taxonomic correction across tip labels. To correct for taxonomy across phylogenies, I applied the same method as above, replacing synonyms by their identified accepted names in trees' tip labels. In some cases, this procedure assigned the same accepted name to different phylogenetic tips. This was the case for 2.6% of mammalian, 1.5% of avian, 1% of amphibian and 1.5% of reptilian species, which then had multiple phylogenetic positions. In other words, replicated tips appeared. Most of these replicated species had two different positions (see table). For each replicated tip, I selected one tip to conserve and dropped other tips from the phylogenies (Figure 2.3). 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

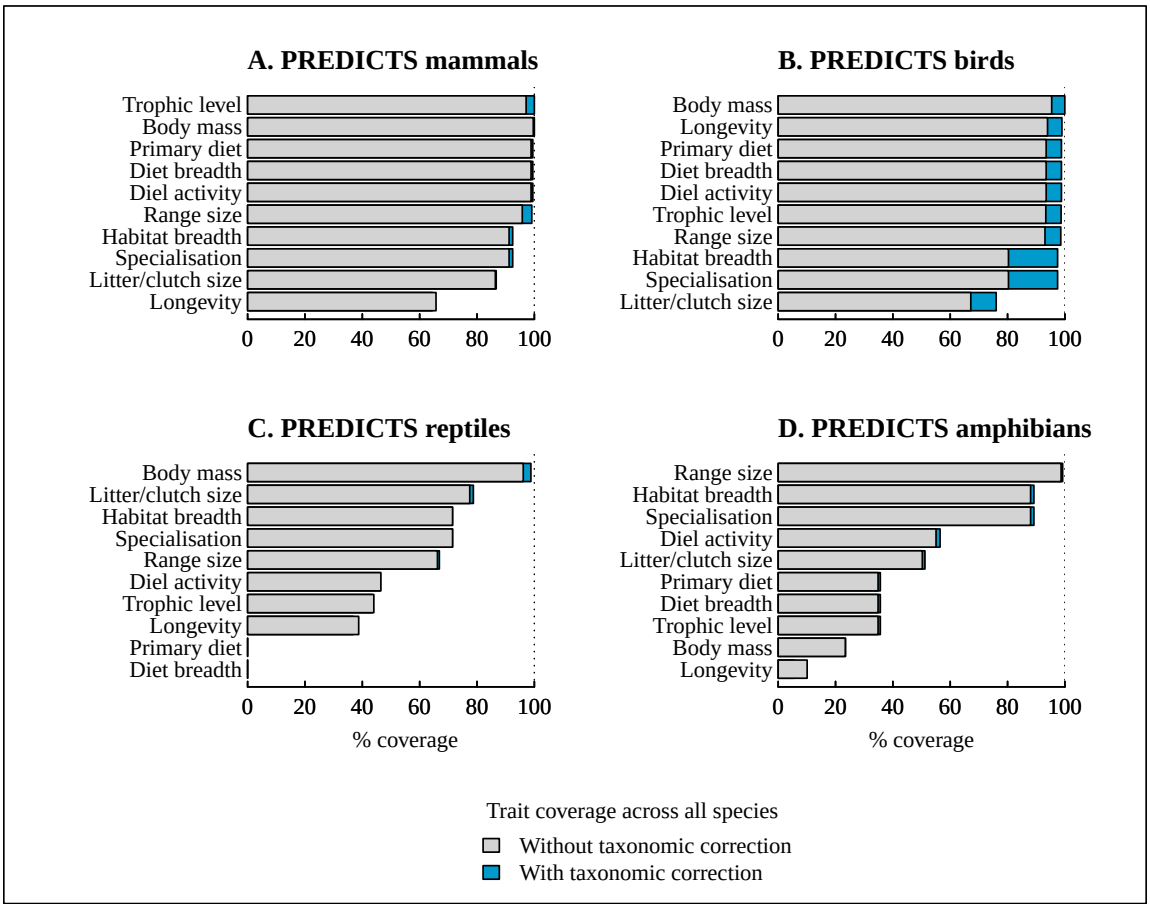

Figure 2.2: Trait coverage across all species before and after taxonomic correction. Trait coverage is defined here as the percentage of species for which trait information is available. Correcting for taxonomic synonymy improved trait coverage in most cases.

how replicated tips were dropped are available in the SI.

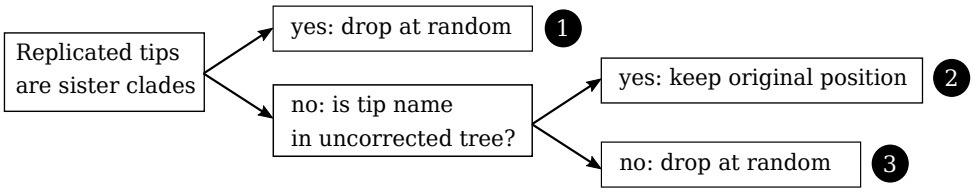

Figure 2.3: Procedure followed to drop replicated tips from phylogenies. Most of these were replicated twice, and were sister clades. In that case, tips to drop were chosen randomly, as it did not affect the 'true' phylogenetic position of the species (1). When replicated were not sister clades, I kept 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 table). See SI for more case examples and more details on the procedure.

Random species attachments. Some species in the trait datasets were not represented in the phylogenies. When applicable, and to increase representation, these species were attached to their genera in the trees at a random position (phytools). Only a small fraction of species that had

no initial phylogenetic representation were randomly attached to their genera (Table 2.2). Overall, correcting for taxonomy improved species representation in the phylogenies (Figure 2.4). For amphibian and reptilian PREDICTS species, representation disporprotionally increased (minimum representation: 73%,for PREDICTS amphibians).

Table 2.2: Species representation in phylogenetic trees (corrected taxonomy). The number of species randomly attached to their genera anged from 17 (amphibians) to 611 (reptiles). Finally, most avian and mammalian species were placed in the phylogenies, whereas more than half reptilian and amphibian species were not.

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

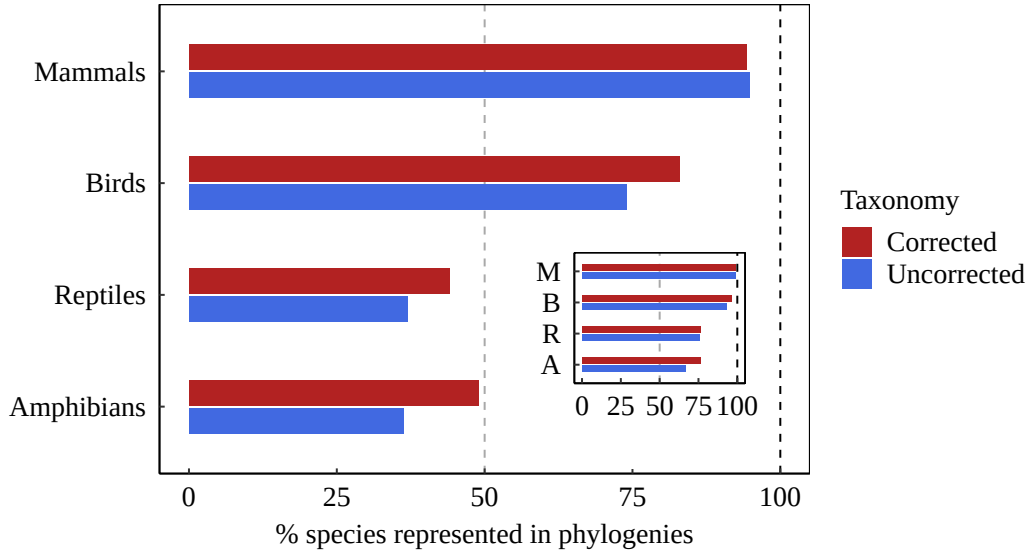

Figure 2.4: Percentage of species represented in the phylogenies for both corrected and uncorrected trait datasets. Overall, taxonomic correction increased species representation in phylogenetic trees. Representation for mammals and birds was high (after taxonomic correction: 83% of avian and 94% of mammalian species had a phylogenetic position). On the other hand, reptiles and amphibians were poorly represented (after taxonomic correction: only 44% of reptilian and 49% of amphibian species were placed in phylogenetic trees). The inset barplot shows representation for species figuring in PREDICTS. For these, species presence in phylogenetic trees after correction was high across all classes, with a minimum representation of 76% for amphibians.

2.2.4 Trait transformations

Species for which all trait values were missing were filtered out. All continuous traits were log10 (except habitat, square root) and standardised to zero-mean and unit variance.

2.2.5 Imputation of missing trait values

Randomness in missing trait values

Phylogenetic signal

The phylogenetic signal in all continuous trait was assessed using Pagel's λ (phytools package). For categorical traits, model of evolution fitted (different models), selection of the model that best fits the data. Some traits showed a strong phylogenetic signal (Table). As such, phylogenies could explain trait values and necessity to incorporate phylogenetic information in further imputations. Are traits evolutionary conserved?

Imputations of missing trait values

Penone et al (2014) assessed the performance of four different imputation approaches (K-nearest neighbour (kNN), multivariate imputation by chained equations (mice), random forest algorithms implemented with missForest and phylogenetic imputations implemented with phylopars). They summarised the advantages and disadvantages of each method. 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 impute on continuous traits. missForest was faster and could deal with both continuous and categorical data. Based on these results, I imputed missing trait values using random forest algorithms, as implemented by missForest. Phylogenetic relationships were including by extracting the first 10 phylogenetic eigenvectors for the phylogenies (PVR package Santos 2018) and adding them as predictor variables. Penone et al showed that 10 phylogenetic eigenvectors minimised the imputation error.

Robustness of imputations

2.3 Results

2.3.1 Comparing data collation outputs for mammals and birds

Collected traits

Comparison of initial coverage

Comparison of collected trait values

Imputed traits

Imputed VS collected

2.3.2 Imputation robustness

2.3.3 Congruence of several imputations

2.4 Discussion

Discuss taxonomy and robustness of imputations

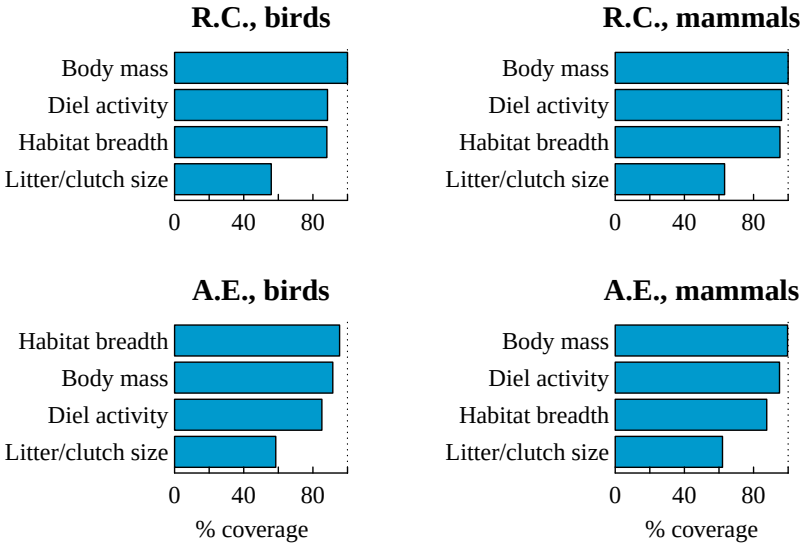

Figure 2.5

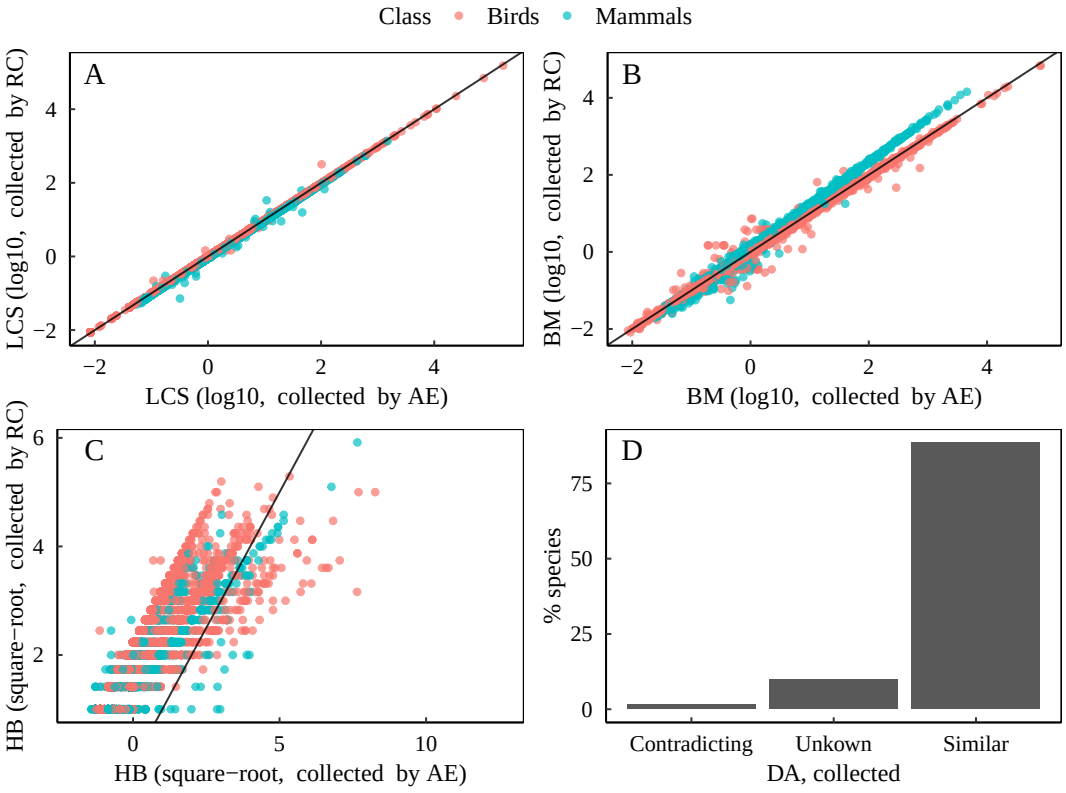

Figure 2.6

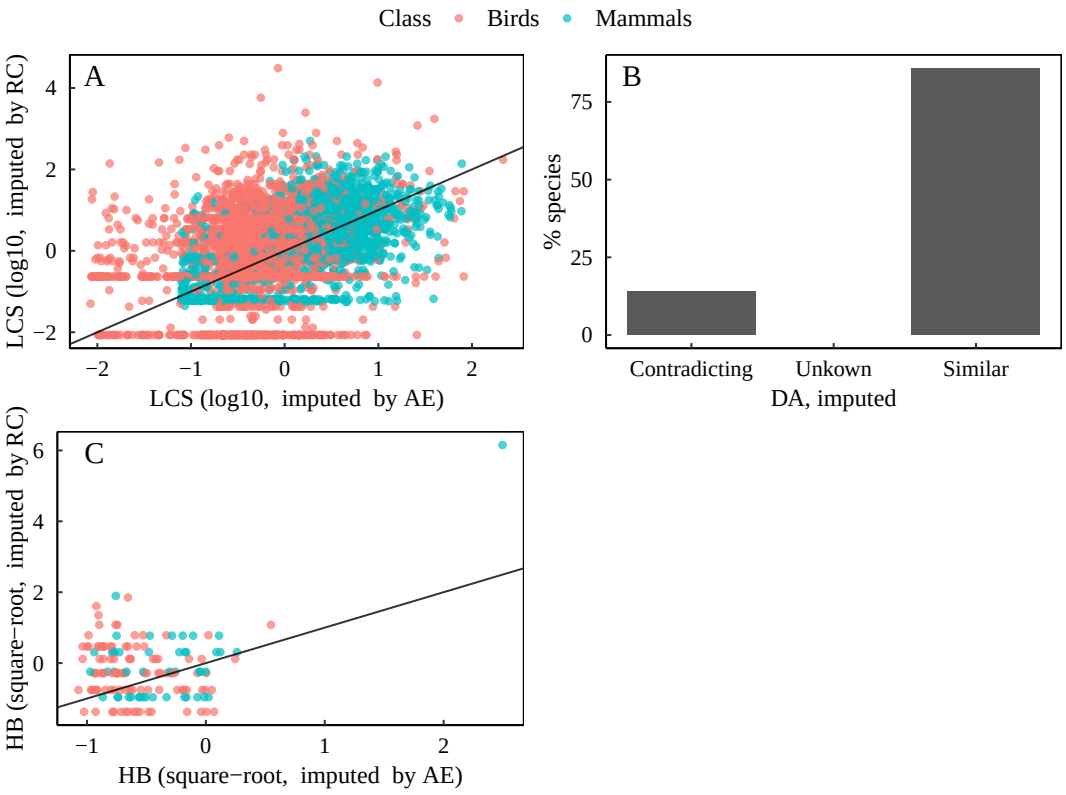
Class
Birds
Mammals
A
LCS (log10, imputed by RC)
LCS (log10, imputed by AE)
B
% species
Contradicting
Unkown
Similar
DA, imputed
C
HB (square−root, imputed by RC)
HB (square−root, imputed by AE)

Figure 2.7

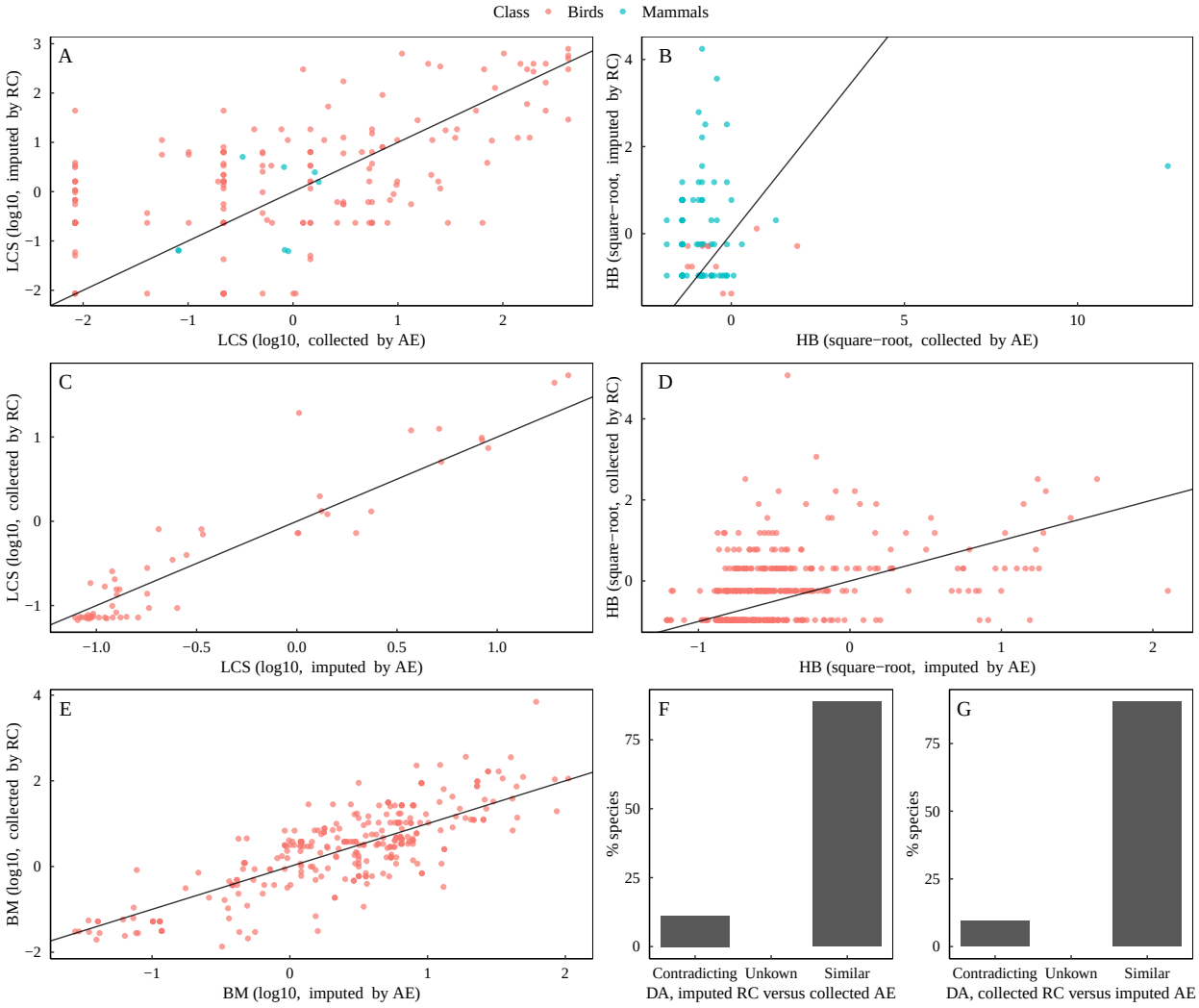
Class
Birds
Mammals
A
LCS (log10, imputed by RC)
LCS (log10, collected by AE)
B
HB (square−root, imputed by RC)
HB (square−root, collected by AE)
C
LCS (log10, collected by RC)
LCS (log10, imputed by AE)
D
HB (square−root, collected by RC)
HB (square−root, imputed by AE)
E
BM (log10, collected by RC)
BM (log10, imputed by AE)
F
% species
Contradicting
Unkown
Similar
DA, imputed RC versus collected AE
G
% species
Contradicting
Unkown
Similar
DA, collected RC versus imputed AE

Figure 2.8

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

4 Outline and research questions for the next years