Upgrade report

Adrienne Etard

February 16, 2019

Introduction

1 | Literature review

2 | Trait data collection and imputation of missing values

2.1 Methods

2.1.1 Ecological trait collection

I collated ecological trait data for terrestrial vertebrates from published databases and unpublished sources (Table). Targeted traits related to species life history (body mass, longevity, litter or clutch size, trophic level, diet) and to their habitat preferences. 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 1, and additional refs). Selecting traits that were ecologically relevant was particularly important, and I will develop this point further down.

All continuous traits were averaged within species when different sources provided estimates. Species diet was described as a binary variable recording whether food items were known to be consumed by a species or not. Diet was available for all classes except reptiles. For amphibians and birds, trophic levels were partly derived 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. See Supporting Information for details on food items and recorded habitats.

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 (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.

Due to the lack of comprehensive diet information readily available for reptiles, and despite compilation efforts for other classes, diet was excluded from further analyses in all four classes.

In addition, I compiled traits that were potentially correlated to either body mass or age at sexual maturity, 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. Longevity and maximum longevity were assumed to provide the same information and were averaged within species.

Finally, species geographical range sizes were calculated from distribution data, extracted from the IUCN Red List. 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).

2.1.2 Tackling taxonomic synonymy

Extraction of synonyms and harmonisation of taxonomy in trait datasets.

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, which was complemented with some manual entries in certain cases. 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), before replacing recorded synonyms by their accepted names in target datasets. To this end, the rredlist and taxize R packages were employed. For each class, I started by generating a list of all binomial names figuring across datasets (trait datasets, phylogenies and PREDICTS species). 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 then homogenised by replacing recorded synonyms with their accepted scientific names. Overall this procedure reduced the total number of species by XXXXXXX. (Table)

Harmonisation of taxonomy in phylogenetic trees (tip labels).

Some species in the trait datasets were not represented in the phylogenies. When applicable, and to increase representation, these species were randomly attached to their genera in the trees (using phytools); see Table. To maximise the number of matches between species in trait datasets and in phylogenetic trees, taxonomy across tree tip labels was also standardised.

Filtering, manual corrections and transformations

The first trait compilation using corrected datasets harvested XXX species for... Species for which all trait values were missing were filtered out. I verified if all vertebrate species in PREDICTS matched a species in the trait datasets, which was only the case after additional manual taxonomic corrections were applied, despite the automated process exposed above.

I also filtered out species for which range sizes (hence distribution data) were missing, except if they appeared in PREDICTS.

2.1.3 Imputation of missing trait value

Trait choice

2.2 Results

2.2.1 Increase in trait coverage due to taxonomic correction

Across all species

Across PREDICTS species

2.2.2 Imputation results

Robustness

2.3 Discussion

Discuss taxonomy extensively!!