Metadata S1. Soria, C.D., Pacifici, M. Di Marco, M. Stephen, S.M., and C. Rondinini. 2021. COMBINE: a coalesced mammal database of intrinsic and extrinsic traits. Ecology.

Metadata

Class I. Data set descriptors

A. Data set identity: COMBINE: A Coalesced Mammal Database of Intrinsic and Extrinsic traits

B. Data set identification code:

1. trait_databases.csv
2. trait_data_reported.csv
3. trait_data_imputed.csv
4. trait_data_sources.csv
5. taxonomy_crosswalk.csv
6. imputation_phylo_1.csv
7. imputation_phylo_30.csv
8. imputation_phylo_83.csv
9. imputation_phylo_181.csv
10. imputation_phylo_209.csv
11. imputation_phylo_219.csv
12. imputation_phylo_729.csv
13. imputation_phylo_756.csv
14. imputation_phylo_825.csv
15. imputation_phylo_979.csv
16. error_imputation.csv
17. SD_validation.csv
18. mean_error_validation.csv

C. Data set description:

Originators:

Carmen D. Soria; Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy
Email: carmen.soria@uniroma1.it

Abstract: The use of species' traits in macroecological analyses has gained popularity in the last decade, becoming an important tool to understand global biodiversity patterns. Currently, trait data can be found across a wide variety of data sets included

in websites, articles, and books, each one with its own taxonomic classification, set of traits and data management methodology. Mammals, in particular, are among the most studied taxa, with large sources of trait information readily available. To facilitate the use of these data, we did an extensive review of published mammal trait data sources between 1999 and May 2020 and produced COMBINE: a COalesced Mammal dataBase of INtrinsic and Extrinsic traits. Our aim was to create a taxonomically integrated database of mammal traits that maximized trait number and coverage without compromising data quality. COMBINE contains information on 54 traits for 6,234 extant and recently extinct mammal species, including information on morphology, reproduction, diet, biogeography, life-habit, phenology, behavior, home range and density. Additionally, we calculated other relevant traits such as habitat and altitudinal breadths for all species and dispersal for terrestrial non-volant species. All data are compatible with the taxonomies of the IUCN Red List v. 2020-2 and PHYLACINE v. 1.2. Missing data were adequately flagged and imputed for non-biogeographical traits with 20% or more data available. We obtained full data sets for 21 traits such as female maturity, litter size, maximum longevity, trophic level, and dispersal, providing imputation performance statistics for all. This data set will be especially useful for those interested in including species' traits in large-scale ecological and conservation analyses. There are no copyright or proprietary restrictions; we request citation of this publication and all relevant underlying data sources (found in Data S1: trait_data_sources.csv), upon using these data.

D. Key words: body size; comparative analyses; diet; hibernation; life history; longevity; macroecology; mammals; mass; sexual maturity; traits

Class II. Research origin descriptors

A. **Overall project description**

Identity: COMBINE: a COalesced Mammal dataBase of INtrinsic and Extrinsic traits

Originators:

Carmen D. Soria; Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy
Email: carmen.soria@uniroma1.it

Period of study: Not applicable

Objectives: Compile a wide variety of publicly available species' traits into a common repository with an updated taxonomy that could be used for macroecological analyses.

Abstract: Same as above. These data are not part of a larger program of study.

Sources of funding: CS and CR have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766417. MDM acknowledges support from MIUR Rita Levi Montalcini programme.

B. Specific subproject description

Site description: Global trait data compilation for extant and recently extinct mammals from already published sources.

Taxonomy: This database includes two different taxonomies at the species level: PHYLACINE v. 1.2 (Faurby et al. 2018; 5,831 species) and IUCN Red List (Version 2020-2; 5961 species). We included both taxonomies in our data set, as they can be used for different purposes. PHYLACINE v. 1.2 taxonomy covers species that have lived since the last interglacial period (around 130,000 years ago until present) and allows phylogenetic trait analyses. IUCN Red List taxonomy ensures compatibility with all IUCN Red List products such as current species' conservation status, range maps and information contained in species' assessments. COMBINE is also mostly compatible with other large-scale mammal data compilation initiatives such as the ASM Mammal Diversity database (https://www.mammaldiversity.org/), focused on taxonomy; and the VertLife initiative (http://vertlife.org/data/), focused on vertebrate phylogenetic data.

Research methods: We conducted a bibliographical search for mammal trait databases and data included in peer-reviewed articles published between 1999 and May 2020. We searched for relevant data combining keywords relevant to the target group ("vertebrate*", "mammal*"), the type of source ("database*", "dataset*", "data") and the target information ("trait*", "life-history trait*") in Web of Science and Google Scholar. In addition to the sources identified in the initial search, we included others discovered by snowball principle, i.e. papers and databases cited in the selected sources. We found and reviewed 43 of these data sources (trait_databases.csv). To maximize efficiency, we kept a subset of 14 sources that focused on all mammal species globally, and we selected traits that had over 10% data coverage. For sources that had raw data, we computed the mean value of the trait per species. For every species, the variable-specific source is referenced in a separate source data set (trait_data_sources.csv). To ensure compatibility between sources and avoid data loss, we aggregated all of them under a common taxonomy (IUCN 2020 version 2020-2). Nomenclatural mismatches were due to taxonomic changes, misspellings, and

formatting inconsistencies. Mismatches were resolved following a two-step procedure. We first extracted synonyms from the IUCN Red List website (https://www.iucnredlist.org/) using the IUCN API in R environment (R Core Team 2020) and verified potential matches. Those that could not be resolved automatically were checked manually and assigned to the target nomenclatures. The following species were not recognized under IUCN (2020) or PHYLACINE v. 1.2 taxonomies and were not included in the analysis: *Brotomys contractus*, *Bubalus bubalis*, *Cavia porcellus*, *Cercopithecus albogularis*, *Cervalces scotti*, *Clidomys osborni*, *Elephas antiquus*, *Felis catus*, *Gazella erlangeri*, *Ictidomys parvidens*, *Melomys spechti*, *Natalus lanatus*, *Nesophontes longirostris*, *Nesophontes submicrus*, *Nesophontes superstes*, *Plagiodontia araeum*, *Pseudopotto martini*, *Pteronotus pristinus*, *Saiga borealis*, *Solenodon aredondoi*, *Spirocerus kiakhtensis*, *Vicugna pacos* and *Vombatus hacketti*. The nomenclature conversion table between the taxonomies of all data sources and IUCN v. 2020-2 can be found in taxonomic_crosswalk.csv. Nomenclatural and taxonomic changes fell into 4 categories, and we derived data for these species based on the description below:

a. Genus and/or specific epithet change: Data was kept unchanged.
b. New species discovered: No data.
c. Species to subspecies: For all species included in the databases, that have been recently moved to the subspecies level in the last IUCN Red List taxonomic revision or PHYLACYNE 1.2 classification, we took the mean value of continuous traits between these formerly considered species and assigned it to the new formal species (except maximum longevity, upper and lower elevation limits). For maximum longevity and maximum elevation limit we kept the maximum value, and for minimum elevation limit we kept the minimum, to reflect the maximum and minimum values. For categorical traits, we used the value that better captured the variation between subspecies (e.g. if one subspecies is herbivorous and the other omnivorous, we considered the species to be omnivorous).
d. Subspecies to species: Data were kept unchanged. If there were no trait data for the new species in any of the data sources, we used the data of the species from which it was split. These data are flagged in trait_data_sources.csv as “data split from (species name)”.

To select only species-specific reported data, i.e. those coming from direct observations, we did not consider calculated values (e.g. imputed, mean value of congenerics or confamilials…) included in the data sources.

Many of our data sources take their information from the same 7 databases and data sets (Table 1). To avoid pseudoreplication bias, we decided to sequentially include trait values from one source rather than taking measures of central tendencies (e.g. mean or median) from all sources for that value. The order of data inclusion (Class IV, Section

B, Table 4) was trait-specific and based on relevance of the source to the considered trait (e.g. trophic level data is probably more accurate in MammalDIET2 (Kissling et al. 2014, Gainsbury et al. 2018), a diet focused source, than in PanTHERIA (Jones et al. 2009), a more general source), presence of data verification or data quality checks and time since publication.

Table 1. Data sources used by two or more of the databases and data sets included in COMBINE

Common sources	**Data sources included in COMBINE**
Smith et al. (2003)	PanTHERIA, EltonTraits, PHYLACINE, Pacifici et al. (2013)
Nowak et al. (1999)	PanTHERIA, EltonTraits, AnAge, Pacifici et al. (2013), Turbill et al. (2011)
Ernest (2003)	AnAge, Amniotes (Myhrvold et al. 2015)
Hayssen et al. (1993)	PanTHERIA, AnAge, Amniotes
PanTHERIA (Jones et al. 2009)	EltonTraits, AnAge, (Pacifici et al. 2013), Heldstab et al. (2018), Botero et al. (2013)
EltonTraits (Wilman et al. 2014)	PHYLACINE, Heldstab et al. (2018), Buckley et al. (2018)
AnAge (De Magalhães and Costa 2009)	Pacifici et al. (2013), Amniotes

We assembled data on a wide variety of traits, including morphology, reproduction, diet, biogeography, life-habit, phenology, behavior, home range and density; creating a common repository of already published trait data. Most traits provide information on a wide variety of orders within the class Mammalia, except forearm length which is almost exclusive to order Chiroptera (99.6% of the data).

We decided to keep two different diet classifications from PHYLACINE v. 1.2 and EltonTraits (coded with the prefix "dphy" and "det" in trait_data.csv, respectively). The first provides a proportional split of each species dietary preferences across plant, vertebrate, and invertebrate food items. The second gives more detailed information on the food item consumed: invertebrates, fish, reptiles and amphibians, mammals, and birds, general or unknown vertebrates, fruits, seeds, nectar and pollen, other plant materials or carrion.

Most traits were homogenous amongst sources and could be coalesced together with minimal or no transformations (changing measurement units). The following traits, that required more complex transformations, were combined from many other traits or were calculated:

a. Activity cycle: Defined as time of the day in which the species carries out most of its activities. Data came from EltonTraits and PanTHERIA. PanTHERIA classified species as nocturnal only, diurnal only or mixed, while EltonTraits had a non-exclusive binary measure of diurnal, nocturnal and crepuscular. We decided to follow PanTHERIA's classification, considering EltonTraits species that were not strictly diurnal or nocturnal as mixed.

b. Life-habit method: Life-habit traits indicate whether a species can be considered terrestrial, marine, or freshwater. Data came from IUCN (2020) and PHYLACINE v. 1.2. Following PHYLACINE v. 1.2, we decided to separate data from IUCN terrestrial mammals into terrestrial volant (those capable of powered flight, belonging to order Chiroptera) and terrestrial non-volant (the rest of terrestrial mammals).

c. Brain mass: Defined as weight of the adult brain in grams. Data came from Tsuboi et al. (2018) and Heldstab et al. (2018). We decided to use brain mass instead of volume, as it was the most used unit of measurement. We converted volume to mass using the known density of mammal brain tissue of 1.036 g/cm^3 (Blinkov and Glezer 1968).

d. Adult body length: Defined as the total length from the tip of the nose to the anus or base of the tail of an adult individual. Data came from Amniotes and PanTHERIA. Gaps were filled using information from male body length, female body length, female body length at maturity and undefined sex body length from Amniotes.

e. Sexual Maturity: Defined as the age at which individuals start being physically capable of reproducing. Data came from PanTHERIA and missing data was completed using female maturity and male maturity from Amniotes and AnAge, and undefined sex maturity from Amniotes.

f. Age of first reproduction: Defined as the age at which females give birth for the first time. Data came from Pacifici et al. (2013) and PanTHERIA. To fill in data gaps, we estimated the age of first reproduction as the sum of gestation length and age at female sexual maturity (Pacifici et al. 2013).

g. Dietary breadth: Dietary breadth can be used as an indicator of the number of different food elements a species consumes. Estimated as the number of different EltonTraits categories that constitute ≥20% of a species' diet (Usui et al. 2017).

h. Habitat breadth: Habitat breadth can be used as an indicator of a species' environmental tolerance. Estimated as the number of distinct level 1 IUCN habitats suitable for the species.

i. Dispersal: Defined as the distance travelled by a species between the birth site and the breeding site. Estimated for terrestrial non-volant species (bats, cetaceans, pinnipeds and sirenids were not considered) following Santini et al. (2013), using species' body mass and trophic level.

j. Altitude breadth: Calculated as the difference between the upper and lower elevation limit of a species. Data came from IUCN assessments.

k. Hibernation and/or torpor: Hibernation and torpor constitute adaptations that enable species to survive during adverse periods (such as cold temperatures, food shortages and droughts) by lowering their body temperature and metabolism. Torpor lasts less than 24 hours, while hibernation is defined by bouts of inactivity lasting from some days to several weeks (Ruf and Geiser 2015). Data came from Buckley et al. (2018), Heldstab et al. (2018), Botero et al. (2013) and Turbill et al. (2011). We grouped together both types of adaptations, considering them as an indicator of avoidance of adverse environmental conditions. Fully aquatic species (sirenids and cetaceans) were considered unable to hibernate (Heldstab et al. 2018).

The data set containing only reported data can be found in trait_data_reported.csv.

Project personnel:

Carmen D. Soria; Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy

Michela Pacifici; Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy

Moreno Di Marco; Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy

Sarah M. Stephen; Department of Conservation Biology, Centre for Nature Conservation, Georg-August-University Göttingen, Von-Slebold-Str. 2, 37075 Göttingen, Germany

Carlo Rondinini; Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy

Class III. Data set status and accessibility

A. Status

Latest update: August 2020.

Latest archive date: Not applicable.

Metadata status: Metadata updated as of September 2020.

Data verification: Data were checked for quality and consistency (see Class V, Section B) by CS and SS.

B. Accessibility

Storage location and medium: COMBINEv1.zip with this publication and on Figshare: https://doi.org/10.6084/m9.figshare.13028255.v4

Contact person: Carmen Soria, Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Rome 00185, Italy.

Copyright restrictions: None. We request citation of this publication in Ecology and all relevant underlying data sources (found in trait_data_sources.csv), upon using these data.

Proprietary restrictions: None.

Costs: None.

Class IV: Data structural descriptors

A. Data set file

Identity:

1. trait_databases.csv
2. trait_data_reported.csv
3. trait_data_imputed.csv
4. trait_data_sources.csv
5. taxonomy_crosswalk.csv
6. imputation_phylo_1.csv
7. imputation_phylo_30.csv
8. imputation_phylo_83.csv
9. imputation_phylo_181.csv
10. imputation_phylo_209.csv
11. imputation_phylo_219.csv
12. imputation_phylo_729.csv
13. imputation_phylo_756.csv
14. imputation_phylo_825.csv
15. imputation_phylo_979.csv
16. imputation_error.csv
17. SD_validation.csv

18. mean_error_validation.csv

Size:

1. 264 records (including header) and 6 fields. Total file size is 10.7 kB
2. 375840 records (including header) and 60 fields. Total file size is 1.6 MB
3. 375840 records (including header) and 60 fields. Total file size is 1.9 MB
4. 375840 records (including header) and 60 fields. Total file size is 3.6 MB
5. 14732 records (including header) and 2 fields. Total file size 285.4 kB
6. 177016 records (including header) and 29 fields. Total file size is 1.3 MB
7. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
8. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
9. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
10. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
11. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
12. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
13. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
14. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
15. 177016 records (including header) and 29 fields. Total file size is 1.6 MB
16. 364 records (including header) and 13 fields. Total file size is 6.4 kB
17. 79313 records (including header) and 13 fields. Total file size is 643.3 kB
18. 36 records (including header) and 3 fields. Total file size is 348 bytes

Format and storage mode:

Header information:

The first rows of all files contain variable names (see Class IV Section B).

Row information:

Each row represents data for a single database in 1, a single species in 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 17 and a single trait in 16 and 18.

Alphanumeric attributes: Mixed

Special characters/fields: missing fields are coded as NA.

Authentication procedures: Checksums for the data:

1. MD5: 7c841e36abc5125f179e9401fe5eda43
2. MD5: d894586c3858620723b1182d8b491d3e
3. MD5: 5c0af319124163830a8e7fa89c0abda5
4. MD5: f883e787cfed1a27342db4711eedaabd
5. MD5: 3eea52f3e79af6380b00915ed8535ddd
6. MD5: f98e1e12c56416b62fc2e55cd7139486
7. MD5: 74a9df9d3beeebf4bf8df47df58c9633
8. MD5: 0f24df5912f6aeda8574f9ece54d2bc6
9. MD5: 1813fe7ad5932a559b5c0134d5b5e62c
10. MD5: d6e4fbc1e9a08ad8b895eb571a057586

11. MD5: c32cfdd61ddfa53618f76def1212aae6
12. MD5: 720cacc3ba18c216af31be44f0712c87
13. MD5: db812945d61ee41c48492e55dee62ace
14. MD5: 2e794f03c5204f5900721ccf7db5c031
15. MD5: 8b13e1241edd1065df5a913175259723
16. MD5: ef8e361a43a8e103057466c3eb6809e0
17. MD5: cafc5e0493b66a22dabb710b5890c0fa
18. MD5: f099856d384904c8f7132cabd56f7731

B. Variable information:

Table 2. Variable information of trait_databases.csv

title

Definition: Title of the data set
Data type: Character
Values: 43 data source names

citation

Definition: Author and year of publishing the database or data set
Data type: Character
Values: 43 different citations

location

Definition: Geographic location of the database or data set
Data type: Character
Values: 8 different locations

taxonomic_scope

Definition: Taxonomical group studied
Data type: Character
Values: 15 different taxonomic groups

traits

Definition: Species' traits included in the source
Data type: Character
Values: 71 different traits

included

Definition: Data source was included in COMBINE
Data type: Character
Values: 0 (no), 1 (yes)

Table 3. Variable information of trait_data_reported.csv and trait_data_imputed.csv

order

Definition:	Order name of the species
Data type:	Character
Values:	29 order names
Completeness:	100%

family

Definition:	Family name of the species
Data type:	Character
Values:	175 family names
Completeness:	100%

genus

Definition:	Genus name of the species
Data type:	Character
Values:	1429 genus names
Completeness:	100%

species

Definition:	Specific epithet name of the species
Data type:	Character
Values:	4422 specific epithets
Completeness:	100%

iucn2020_binomial

Definition:	IUCN v. 2020-2 binomial name
Data type:	Character
Values:	5961 binomial names
Completeness:	100%

phylacine_binomial

Definition:	PHYLACINE v. 1.2 binomial name
Data type:	Character
Values:	5831 binomial names
Completeness:	100%

adult_mass_g

Definition:	Body mass of an adult individual in grams
Data type:	Numeric (float)
Values:	Estimates range from 1.6 g to $1.49*10^8$ g
Completeness:	96.33%

brain_mass_g

Definition: Weight of the brain of an adult individual in grams
Data type: Numeric (float)
Values: Estimates range from 0.071 g to 7,818 g
Completeness: 29.30% (97.5% with imputed data)

adult_body_length_mm

Definition: Total length from tip of the nose to anus or base of the tail of an adult individual in millimeters
Data type: Numeric (float)
Values: Estimates range from 30.99 mm to 30,490 mm
Completeness: 66.66% (97.5% with imputed data)

adult_forearm_length_mm

Definition: Total length from elbow to wrist of an adult individual in millimeters, specific to order Chiroptera
Data type: Numeric (float)
Values: Estimates range from 26 mm to 246 mm
Completeness: 16.27%

max_longevity_d

Definition: Maximum reported age at death for the species in days
Data type: Numeric (float)
Values: Estimates range from 30.42 days to 77,015 days
Completeness: 44.10% (97.5% with imputed data)

maturity_d

Definition: The amount of time needed to reach sexual maturity in days
Data type: Numeric (float)
Values: Estimates range from 13.57 days to 6,041.21 days
Completeness: 34.40%

female_maturity_d

Definition: The amount of time needed for a female to reach sexual maturity in days
Data type: Numeric (float)
Values: Estimates range from 23.81 days to 6,391.56 days
Completeness: 34.10% (97.5% with imputed data)

male_maturity_d

Definition: The amount of time needed for a male to reach sexual maturity in days
Data type: Numeric (float)

Values: Estimates range from 36 days to 8,212 days
Completeness: 17.84%

age_first_reproduction_d

Definition: Age at which females give birth to their first litter or their young attach to teats in days
Data type: Numeric (float)
Values: Estimates range from 39 days to 8,599.95 days
Completeness: 33.77% (97.5% with imputed data)

gestation_length_d

Definition: Length of time of fetal growth in days
Data type: Numeric (float)
Values: Estimates range from 10 days to 669.68 days
Completeness: 37.96% (97.5% with imputed data)

teat_number_n

Definition: Total number of teats present in an individual of the species
Data type: Numeric (integer)
Values: Estimates range from 1 teat to 26 teats
Completeness: 10.84%

litter_size_n

Definition: Number of offspring born per litter per female
Data type: Numeric (float)
Values: Estimates range from 0.9 offspring to 16.87 offspring
Completeness: 60.17% (97.5% with imputed data)

litters_per_year_n

Definition: Number of litters per female per year
Data type: Numeric (float)
Values: Estimates range from 0.12 litters to 10 litters
Completeness: 36.81% (97.5% with imputed data)

interbirth_interval_d

Definition: Time between reproduction events in days
Data type: Numeric (float)
Values: Estimates range from 55.58 days to 231 days
Completeness: 21.51% (97.5% with imputed data)

neonate_mass_g

Definition: Weight of an individual at birth in grams
Data type: Numeric (float)

Values: Estimates range from 0.0043 g to 2,250,000 g
Completeness: 32.40%

weaning_age_d

Definition: Age at which primary nutritional dependency on the mother ends and independent foraging begins in days
Data type: Numeric (float)
Values: Estimates range from 1.94 days to 1,826.25 days
Completeness: 35.00% (97.5% with imputed data)

weaning_mass_g

Definition: Weight at weaning in grams
Data type: Numeric (float)
Values: Estimates range from 0.7 g to 17,000,000 g
Completeness: 18.11%

generation_length_d

Definition: Average age of parents of the current cohort in days
Data type: Numeric (float)
Values: Estimates range from 128.98 days to 18980 days
Completeness: 22.91% (97.5% with imputed data)

dispersal_km

Definition: The distance an animal travels between its place of birth to the place where it reproduces in kilometers
Data type: Numeric (float)
Values: Estimates range from 0.040 km to 109.14 km
Completeness: 69.62% (75.09% with imputed data)

density_n_km2

Definition: Number of individuals of the species per squared kilometer
Data type: Numeric (float)
Values: Estimates range from 0.00026 ind/km^2 to 57,067.85 ind/km^2
Completeness: 20.15%

hibernation_torpor

Definition: Individuals of the species go through hibernation or torpor
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 49.46% (97.27% with imputed data)

fossoriality

Definition: The species is above ground dwelling or ground/fossorial dwelling

Data type: Binary

Values: Two levels:

- 1: fossorial and/or ground dwelling
- 2: above ground dwelling

Completeness: 46.20% (97.19% with imputed data)

home_range_km2

Definition: Size of the area within which everyday activities of individuals or groups of individuals are typically restricted in km^2

Data type: Numeric (float)

Values: Estimates range from $2.04*10^{-5}$ km^2 to 79244.75 km^2

Completeness: 12.78%

social_group_size_n

Definition: Number of individuals in a group that spends most of their daily time together

Data type: Numeric (float)

Values: Estimates range from 1 individual to 110 individuals

Completeness: 13.10%

dphy_invertebrate

Definition: Percentage of the diet composed of invertebrates

Data type: Numeric (float)

Values: Percentage values range from 0% to 100%

Completeness: 96.73% (97.5% with imputed data)

dphy_vertebrate

Definition: Percentage of the diet composed of vertebrates

Data type: Numeric (float)

Values: Percentage values range from 0% to 100%

Completeness: 96.73% (97.5% with imputed data)

dphy_plant

Definition: Percentage of the diet composed of plants and/or fungi

Data type: Numeric (float)

Values: Percentage values range from 0% to 100%

Completeness: 96.73% (97.5% with imputed data)

det_inv

Definition: Percentage of the diet composed of invertebrates

Data type: Numeric (float)

Values: Percentage values range from 0% to 100%

Completeness: 74.37%

det_vend

Definition: Percentage of the diet composed of mammals, birds
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_vect

Definition: Percentage of the diet composed of reptiles, snakes, amphibians, salamanders
Data type: Numeric (float)
Values: Percentage values range from 0% to 60%
Completeness: 74.37%

det_vfish

Definition: Percentage of the diet composed of fish
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_vunk

Definition: Percentage of the diet composed of vertebrates – general or unknown
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_scav

Definition: Percentage of the diet composed of scavenge, garbage, offal, carcasses, trawlers, carrion
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_fruit

Definition: Percentage of the diet composed of fruit, drupes
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_nect

Definition: Percentage of the diet composed of nectar, pollen, plant exudates, gums

Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_seed

Definition: Percentage of the diet composed of seed, maize, nuts, spores, wheat, grains
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_plantother

Definition: Percentage of the diet composed of other plant elements
Data type: Numeric (float)
Values: Percentage values range from 0% to 100%
Completeness: 74.37%

det_diet_breadth_n

Definition: Number of prevalent (≥ 20%) EltonTraits dietary categories consumed
Data type: Numeric (integer)
Values: Values range from 1 dietary category to 5 dietary categories
Completeness: 74.37% (97.5% with imputed data)

trophic_level

Definition: Trophic level of the species
Data type: Ordinal
Values: Three levels:
1: herbivore
2: omnivore
3: carnivore
Completeness: 91.28% (97.5% with imputed data)

foraging_stratum

Definition: Assignment to one of five foraging stratum categories
Data type: Ordinal
Values: Five levels:

- M: marine
- G: ground level, including aquatic foraging
- S: scansorial
- Ar: arboreal
- A: aerial

Completeness: 90.15% (97.6% with imputed data)

activity_cycle

Definition: Activity cycle of each species
Data type: Ordinal
Values: Three levels:
1: nocturnal only
2: nocturnal/crepuscular, cathemeral, crepuscular or diurnal/crepuscular
3: diurnal only
Completeness: 80.96% (97.4% with imputed data)

freshwater

Definition: The species spends a significant amount of time in freshwater bodies
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 96.68%

marine

Definition: The species spends a significant amount of time in oceans and/or seas
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 96.68%

terrestrial_non-volant

Definition: The species spends a significant amount of time on land
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 96.68%

terrestrial_volant

Definition: The species is capable of powered flight and spends a significant amount of time flying in the air
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 96.68%

upper_elevation_m

Definition: Upper elevation limit at which the species can be found in meters
Data type: Numeric (float)
Values: Estimates range from 0 m to 6,700 m
Completeness: 50.67%

lower_elevation_m

Definition: Lower elevation limit at which the species can be found in meters
Data type: Numeric (float)
Values: Estimates range from -100 m to 4,500 m
Completeness: 46.53%

altitude_breadth_m

Definition: Difference between the upper and lower elevation limits of a species in meters
Data type: Numeric (float)
Values: Estimates range from 0 to 6200 m
Completeness: 43.31%

island_dwelling

Definition: 20% or more of the breeding range occurs on an island
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 50.38%

island_endemicity

Definition: Score of island endemicity obtained from species' ranges and historical and fossil occurrence records
Data type: Ordinal
Values: Four levels:

- Exclusively marine
- Occurs on mainland
- Occurs on large land bridge islands: the species occurs on islands greater than 1,000 km2 that are separated from the mainland by water no more than 110 m deep. The islands would have been part of the mainland during the last glacial maximum.
- Occurs on small land bridge islands: the species occurs on islands smaller than 1,000 km2 that are separated from the mainland by water no more than 110 m deep. The islands would have been part of the mainland during the last glacial maximum.
- Occurs only on isolated islands: the species occurs on islands separated from the mainland by water deeper than 110 m.

Completeness: 93.10%

dissected_by_mountains

Definition: Range dissected by mountains (based on elevation gradients with slopes equal or higher than 5 degrees)
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 50.38%

glaciation

Definition: Historical exposure to glaciation, considered as more than 20% range overlap with areas glaciated in the last 21000 years
Data type: Binary
Values: 0 (no), 1 (yes)
Completeness: 50.38%

biogeographical_realm

Definition: Biogeographical realms in which the species can be encountered
Data type: Ordinal
Values: Eight biogeographical realms:

- Afrotropical
- Antarctic
- Australasian
- Indomalayan
- Nearctic
- Neotropical
- Oceanian
- Palearctic

Completeness: 95.66%

habitat_breadth_n

Definition: Number of distinct suitable level 1 IUCN habitats
Data type: Numeric (integer)
Values: Estimates range from 1 habitat to 9 habitats
Completeness: 90.30%

Table 4. Variable information of trait_data_sources.csv with data inclusion order

order

Definition: Order name source
Data type: Character
Values: 29 order names

family

Definition: Family name source
Data type: Character

Values:	175 family names

genus

Definition:	Genus name source
Data type:	Character
Values:	1429 genus names

species

Definition:	Specific epithet name source
Data type:	Character
Values:	4422 specific epithets

iucn2020_binomial

Definition:	IUCN v. 2020-2 binomial name source
Data type:	Character
Values:	5961 binomial names

phylacine_binomial

Definition:	PHYLACINE v. 1.2 binomial name source
Data type:	Character
Values:	5831 binomial names

adult_mass_g

Definition:	Adult body mass source
Data type:	Character
Values/inclusion order:	Amniotes, Pacifici et al. (2013), Smith et al. (2003) (EltonTraits), AnAge, PHYLACINE, split from (species name)

brain_mass_g

Definition:	Adult brain mass source
Data type:	Character
Values/inclusion order:	Tsuboi et al. (2018), Heldstab et al. (2018), split from (species name), imputed

adult_body_length_mm

Definition:	Adult body length source
Data type:	Character
Values/inclusion order:	Amniotes, PanTHERIA, mean of female and female head body length (Amniotes), female head body length (Amniotes), undefined sex head body length (Amniotes), split from (species name), imputed

adult_forearm_length_mm

Definition:	Adult forearm length source
Data type:	Character
Values/inclusion order:	PanTHERIA, split from (species name)

max_longevity_d

Definition:	Maximum longevity source
Data type:	Character
Values/inclusion order:	Amniotes, Pacifici et al. (2013), AnAge, split from (species name), imputed

maturity_d

Definition:	Maturity source
Data type:	Character
Values/inclusion order:	PanTHERIA, female maturity (Amniotes), mean of female (Amniotes) and male (AnAge), mean of female and male (AnAge), mean of female and undefined sex (Amniotes), mean of female, male and undefined sex (Amniotes), mean of female and male (Amniotes), mean of male and undefined (Amniotes), undefined sex maturity (Amniotes), split from (species name)

female_maturity_d

Definition:	Female maturity source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, split from (species name), imputed

male_maturity_d

Definition:	Male maturity source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, split from (species name)

age_first_reproduction_d

Definition:	Age of first reproduction source
Data type:	Character
Values/inclusion order:	Pacifici et al. (2013), PanTHERIA, split from (species name), calculated, imputed

gestation_length_d

Definition:	Gestation length source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name), imputed

teat_number_n

Definition:	Teat number source
Data type:	Character
Values/inclusion order:	PanTHERIA, split from (species name)

litter_size_n

Definition:	Litter size source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name), imputed

litters_per_year_n

Definition:	Litters per year source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name), imputed

interbirth_interval_d

Definition:	Interbirth interval source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name)

neonate_mass_g

Definition:	Neonate mass source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name)

weaning_age_d

Definition:	Weaning age source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name), imputed

weaning_mass_g

Definition:	Weaning mass source
Data type:	Character
Values/inclusion order:	Amniotes, AnAge, PanTHERIA, split from (species name)

generation_length_d

Definition: Generation length source
Data type: Character
Values/inclusion order: Pacifici et al. (2013), IUCN, split from (species name), imputed

dispersal_km

Definition: Dispersal source
Data type: Character
Values/inclusion order: calculated, split from (species name), imputed

density_n_km2

Definition: Density source
Data type: Character
Values/inclusion order: TetraDENSITY, PanTHERIA, split from (species name)

hibernation_torpor

Definition: Hibernation and torpor source
Data type: Character
Values/inclusion order: Buckley et al. (2018), Heldstab et al. (2018), Botero et al. (2013), Turbill et al. (2011), marine, split from (species name), imputed

fossoriality

Definition: Fossoriality source
Data type: Character
Values/inclusion order: PanTHERIA, split from (species name), imputed

home_range_km2

Definition: Home range source
Data type: Character
Values/inclusion order: PanTHERIA, split from (species name)

social_group_size_n

Definition: Social group size source
Data type: Character
Values/inclusion order: PanTHERIA, split from (species name)

dphy_invertebrate

Definition: Diet invertebrate source
Data type: Character
Values/inclusion order: PHYLACINE, split from (species name), imputed

dphy_vertebrate

Definition: Diet vertebrate source
Data type: Character
Values/inclusion order: PHYLACINE, split from (species name), imputed

dphy_invertebrate

Definition: Diet plant source
Data type: Character
Values/inclusion order: PHYLACINE, split from (species name), imputed

det_inv

Definition: Diet inv source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_vend

Definition: Diet vend source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_vect

Definition: Diet vect source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_vfish

Definition: Diet vfish source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_vunk

Definition: Diet vunk source

Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_scav
Definition: Diet scav source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_fruit
Definition: Diet fruit source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_nect
Definition: Diet nect source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_seed
Definition: Diet seed source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_plantother
Definition: Diet plantother source
Data type: Character
Values/inclusion order: EltonTraits, split from (species name)

det_diet_breadth_n
Definition: Diet breadth source
Data type: Character
Values/inclusion order: Calculated (EltonTraits), split from (species name), imputed

trophic_level
Definition: Trophic level source
Data type: Character

Values/inclusion order:	MammalDIET2, PanTHERIA, split from (species name), imputed

foraging_stratum

Definition:	Foraging stratum source
Data type:	Character
Values/inclusion order:	EltonTraits, split from (species name), imputed

activity_cycle

Definition:	Activity cycle source
Data type:	Character
Values/inclusion order:	EltonTraits, PanTHERIA, split from (species name), imputed

freshwater

Definition:	Freshwater source
Data type:	Character
Values/inclusion order:	IUCN, PHYLACINE, split from (species name)

marine

Definition:	Marine source
Data type:	Character
Values/inclusion order:	IUCN, PHYLACINE, split from (species name)

terrestrial_non-volant

Definition:	Terrestrial non-volant source
Data type:	Character
Values/inclusion order:	IUCN, PHYLACINE, split from (species name)

terrestrial_volant

Definition:	Terrestrial volant source
Data type:	Character
Values/inclusion order:	IUCN, PHYLACINE, split from (species name)

upper_elevation_m

Definition:	Upper elevation limit source
Data type:	Character

Values/inclusion order: IUCN, split from (species name)

lower_elevation_m

Definition: Lower elevation limit source
Data type: Character
Values/inclusion order: IUCN, split from (species name)

altitude_breadth_m

Definition: Altitude breadth source
Data type: Character
Values/inclusion order: calculated, split from (species name)

island_dwelling

Definition: Island dwelling source
Data type: Character
Values/inclusion order: Botero et al. (2013)

island_endemicity

Definition: Island endemicity source
Data type: Character
Values/inclusion order: PHYLACINE

dissected_by_mountains

Definition: Dissected by mountains source
Data type: Character
Values/inclusion order: Botero et al. (2013)

glaciation

Definition: Glaciation source
Data type: Character
Source order: Botero et al. (2013)

biogeographical_realm

Definition: Biogeographical realm source
Data type: Character
Values/inclusion order: IUCN

habitat_breadth_n

Definition:	Number of distinct suitable level 1 IUCN habitats
Data type:	Character
Values/inclusion order:	calculated, split from (species name)

Table 5. Variable information of taxonomy_crosswalk.csv

iucn2020_binomial

Definition:	IUCN v. 2020-2 binomial name
Data type:	Character
Values:	5961 binomial names

input_name

Definition:	Binomial name from other data sources
Data type:	Character
Values:	6823 binomial names

Table 6. Variable information of imputation data for all phylogenies (archives 6 – 15)

iucn2020_binomial

Definition:	IUCN v. 2020-2 binomial name
Data type:	Character
Values:	5802 binomial names

phylacine_binomial

Definition:	PHYLACINE v. 1.2 binomial name
Data type:	Character
Values:	5831 binomial names

order

Definition:	Order name of the species
Data type:	Character
Values:	29 order names

adult_mass_g

Definition:	Input and imputed adult body mass values
Data type:	Numeric (float)
Values:	Estimates range from 1.6 g to $1.49*10^8$ g

neonate_mass_g

Definition: Input and imputed neonate body mass values
Data type: Character
Values: Estimates range from 0.0043 g to 2,250,000 g

brain_mass_g

Definition: Input and imputed brain mass values
Data type: Numeric (float)
Values: Estimates range from 0.071 g to 7,818 g

adult_body_length_mm

Definition: Input and imputed adult body length values
Data type: Numeric (float)
Values: Estimates range from 30.99 mm to 30,490 mm

dphy_plant

Definition: Input and imputed percentage of the diet composed by plants and/or fungi
Data type: Numeric (float)
Values: Estimates range from 0% to 100%

dphy_vertebrate

Definition: Input and imputed percentage of the diet composed by vertebrates
Data type: Numeric (float)
Values: Estimates range from 0% to 100%

dphy_invertebrate

Definition: Input and imputed percentage of the diet composed by invertebrates
Data type: Numeric (float)
Values: Estimates range from 0% to 100%

trophic_level

Definition: Input and imputed trophic level values
Data type: Ordinal
Values: Three levels:
1: herbivore
2: omnivore
3: carnivore

det_diet_breadth_n

Definition: Input and imputed number of prevalent (≥ 20%) EltonTraits dietary categories consumed
Data type: Numeric (integer)
Values: Values range from 1 dietary category to 5 dietary categories

foraging_stratum

Definition: Input and imputed foraging stratum values

Data type: Ordinal

Values: Five levels:

- M: marine
- G: ground level, including aquatic foraging
- S: scansorial
- Ar: arboreal
- A: aerial

habitat_breadth_n

Definition: Input and imputed habitat breadth values

Data type: Numeric (integer)

Values: Estimates range from 1 habitat to 9 habitats

activity_cycle

Definition: Input and imputed activity cycle values

Data type: Ordinal

Values: Three levels:

1: nocturnal only

2: nocturnal/crepuscular, cathemeral, crepuscular or diurnal/crepuscular

3: diurnal only

fossoriality

Definition: Input and imputed fossoriality values

Data type: Ordinal

Values: Two levels:

- 1: fossorial and/or ground dwelling
- 2: above ground dwelling

hibernation_torpor

Definition: Input and imputed hibernation or torpor values

Data type: Binary

Values: 0 (no), 1 (yes)

max_longevity_d

Definition: Input and imputed maximum longevity values

Data type: Numeric (float)

Values: Estimates range from 30.42 days to 77,015 days

litter_size_n

Definition: Input and imputed litter size values
Data type: Numeric (float)
Values: Estimates range from 0.9 individuals to 16.89 individuals

litters_per_year_n

Definition: Input and imputed litters per year values
Data type: Numeric (float)
Values: Estimates range from 0.12 litters to 10 litters

interbirth_interval_d

Definition: Input and imputed interbirth interval values
Data type: Numeric (float)
Values: Estimates range from 17 days to 1769 days

gestation_length_d

Definition: Input and imputed gestation length values
Data type: Numeric (float)
Values: Estimates range from 10 days to 669.68 days

weaning_age_d

Definition: Input and imputed weaning age values
Data type: Numeric (float)
Values: Estimates range from 1.94 days to 1826.25 days

female_maturity_d

Definition: Input and imputed female maturity values
Data type: Numeric (float)
Values: Estimates range from 23.81 days to 6391.56 days

age_first_reproduction_d

Definition: Input and imputed age of first reproduction values
Data type: Numeric (float)
Values: Estimates range from 39 days to 8599.95 days

generation_length_d

Definition: Input and imputed generation length values
Data type: Numeric (float)
Values: Estimates range from 128.98 days to 18980 days

density_n_km2

Definition: Input and imputed density values
Data type: Numeric (float)
Values: Estimates range from 0.0002585 days to 57067.85 days

altitude_breadth

Definition: Input and imputed altitude breadth values
Data type: Numeric (float)
Values: Estimates range from 0 m to 6200 m

dispersal_km

Definition: Input and imputed dispersal values
Data type: Numeric (float)
Values: Estimates range from 0.040 km to 109.14 km

Table 7. Variable information of imputation_error.csv

trait_name

Definition: Name of the trait used for imputation
Data type: Character
Values: 27 trait names

missingness

Definition: Proportion of missing data
Data type: Numeric (float)
Values: Estimates range from 0% to 79.37%

nrmse_pfc

Definition: If the error is NRMSE or PFC
Data type: Ordinal
Values: NRMSE, PFC

mean_error

Definition: Mean NRMSE or PFC of all 10 phylogenies
Data type: Numeric (float)
Values: Estimates range from 0.035 to 0.912

error_phylo (1, 30, 83, 181, 209, 219, 729, 756, 825, 979)

Definition: Error of each of the ten phylogenies
Data type: Numeric (float)
Values: Estimates range from 0.031 to 0.935

Table 8. Variable information of SD_validation.csv

iucn2020_binomial

Definition: IUCN v. 2020-2 binomial name
Data type: Character

Values: 5802 binomial names

phylacine_binomial

Definition: PHYLACINE v. 1.2 binomial name
Data type: Character
Values: 5831 binomial names

SD_age_first_reproduction

Definition: Standard deviation of reported and 10 imputed values for age of first reproduction
Data type: Numeric (float)
Values: Estimates range from 0 to 810.614

SD_adult_body_length

Definition: Standard deviation of reported and 10 imputed values for adult body length
Data type: Numeric (float)
Values: Estimates range from 0 to 1943.613

SD_brain_mass

Definition: Standard deviation of reported and 10 imputed values for brain mass
Data type: Numeric (float)
Values: Estimates range from 0 to 983.419

SD_female_maturity

Definition: Standard deviation of reported and 10 imputed values for female maturity
Data type: Numeric (float)
Values: Estimates range from 0 to 815.359

SD_generation_length

Definition: Standard deviation of reported and 10 imputed values for generation length
Data type: Numeric (float)
Values: Estimates range from 0 to 1611.883

SD_gestation_length

Definition: Standard deviation of reported and 10 imputed values for gestation length
Data type: Numeric (float)
Values: Estimates range from 0 to 59.161

SD_interbirth_interval

Definition: Standard deviation of reported and 10 imputed values for interbirth interval
Data type: Numeric (float)
Values: Estimates range from 0 to 254.318

SD_litter_size
Definition: Standard deviation of reported and 10 imputed values for litter size
Data type: Numeric (float)
Values: Estimates range from 0 to 1.161

SD_litters_per_year
Definition: Standard deviation of reported and 10 imputed values for litters per year
Data type: Numeric (float)
Values: Estimates range from 0 to 1.495

SD_max_longevity
Definition: Standard deviation of reported and 10 imputed values for maximum longevity
Data type: Numeric (float)
Values: Estimates range from 0 to 810.614

SD_weaning_age
Definition: Standard deviation of reported and 10 imputed values for weaning age
Data type: Numeric (float)
Values: Estimates range from 0 to 157.6

Table 9. Variable information of mean_error_validation.csv

trait_name
Definition: Name of the trait used for imputation
Data type: Character
Values: 11 trait names

mean_missingness
Definition: Mean proportion of missing data
Data type: Numeric (float)
Values: Estimates range from 0.351 to 0.79

mean_nrmse
Definition: Mean NRMSE of the validation
Data type: Numeric (float)
Values: Estimates range from 0.211 to 0.379

C. **Data anomalies:**

Missing data: Each species and trait we analyzed were characterized by the presence of missing data. The percentage of missing data per trait ranged between 3.67% for body mass to 89.16% for teat number. To fill these data gaps, making the data set ready to use for analyses, missing values were imputed for a subset of 27 traits (Table 6) with the missForest algorithm in R (Nonparametric Missing Value Imputation using Random Forest; Stekhoven and Buhlmann, 2012). This algorithm allows the imputation of categorical and continuous variables and does not need tuning parameters or assumptions of the distribution of the data (Breiman, 2001). Our subset of traits was composed of those with more than 20% data completeness that could be adequately imputed phylogenetically (i.e. we did not include the following biogeographical traits: upper and lower elevation limits, island dwelling, island endemicity, glaciation, dissected by mountains and biogeographical realm). For diet, we only used PHYLACINE's classification (plant, invertebrate and vertebrate) as it can be accurately imputed phylogenetically (Faurby et al. 2018, Gainsbury et al. 2018). To include phylogenetic information, we randomly selected 10 phylogenies from PHYLACINE v. 1.2 (IDs: 1, 30, 83, 181, 209, 219, 729, 756, 825, 979), and extracted 10 eigenvectors for each phylogeny to be included as variables in the imputation routine. We thus obtained a total of 10 phylogeny-specific data sets. To select the optimum number of eigenvectors, we ran an imputation routine with 5, 10, 15 and 20 eigenvectors and selected the number of eigenvectors with the highest accuracy of imputation. We ran an imputation routine over each data set, with 10 iterations per imputation, obtaining 10 complete data sets (imputation_phylo archives), with imputed data filling in data gaps in the observed data. To estimate the accuracy of the imputation, we used the out of bag (OOB) error provided by the algorithm to calculate the normalized root mean squared error (NRMSE; Oba et al 2003) and the proportion of falsely classified entries (PFC), for continuous and categorical variables respectively (found in imputation_error.csv). Both estimates range from 0 (highest accuracy) to 1 (lowest accuracy). We calculated the NRMSE and PFC for each phylogeny and the mean across all of them, which can be found in imputation_error.csv. We retained all variables for which the NRMSE or PFC values were < 0.4. Imputed values are flagged in the source table, to allow easy separation from observed values (trait_data_sources.csv). The methodology we used to calculate dispersal cannot be applied to marine or terrestrial volant mammals (Santini et al., 2013). We thus estimated dispersal missing values from a separate imputation routine, including only terrestrial non-volant species. We also produced a "combined" imputed data set, by calculating the mean of the imputation value across all ten imputed data sets (for continuous variables) or the most repeated imputed value (for categorical variables). If there was a tie between the most repeated imputed values, we

assigned an NA. The data set containing reported and imputed data can be found in trait_data_imputed.csv

Even though adequate imputation methodologies are useful to temporarily fill in data gaps, we acknowledge that it is not the ideal and want to remark the importance of the collection and digitization of primary natural history data.

Class V: Supplemental descriptors

A. **Data acquisition**: See Class II, Section B

B. **Data assurance/quality control procedures:**

We developed a two-step data validation process to identify potentially erroneous data pre-imputation.

We first checked for inconsistencies in mass and longevity related traits. For each species, adult body mass had to be higher than weaning body mass and neonate body mass; and maximum longevity higher than maturity (female, male and combined), age at first reproduction, weaning age, gestation length and interbirth interval. Nineteen longevity related values did not meet these criteria and were not included.

To check the data quality, we calculated the standard deviation (SD) between the original value and those obtained from the jackknifed multiple imputation of 10 phylogenies (same IDs and methodology as in Class IV Section C) for continuous traits included in the imputation with a NRMSE < 0.4 (brain mass, adult body length, maximum longevity, litter size, litters per year, interbirth interval, gestation length, weaning age, female maturity, age of first reproduction, and generation length). These jackknifed imputations were done by sequentially extracting 5% of the trait data, obtaining 20 tables per trait. To avoid phylogenetic bias in the extraction, species were ordered randomly. Standard deviation values can be found in SD_validation.csv. Imputation NRMSE can be found in mean_error_validation.csv.

For estimations of global model parameters that account for uncertainty in data imputation (e.g. estimate of body mass vs gestation time relationship), we recommend repeating all analyses independently across the 10 data sets and calculate the mean parameter and the overall variance (Nakagawa and Freckleton, 2008). For global analyses not associated to the estimate of statistical parameters (e.g. global-scale mapping of functional diversity) the use of the combined (already averaged) data set might be sufficient to users' needs. For species and site-specific studies, we encourage using more specialized data sources. Despite our efforts in gathering as much good quality data, the scope of the project implies that there were uncertainties beyond our control and intraspecific variability is not represented.

C. **Related material**: Not applicable

D. **Computer programs and data processing algorithms:** Trait data was assembled using the statistical language R, version 3.6.3 (R Development Core Team 2020) and the packages missForest, PVR and ape.

E. **Archiving:** Not applicable

F. **Publications and results:** Not applicable

G. **History of data set usage:** Not applicable

Acknowledgments

We would like to thank all the researchers that collected all the original field data and compiled and analyzed all the data sources used in this work. Without their efforts, this work would not have been possible. CS and CR have received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766417. MDM acknowledges support from MIUR Rita Levi Montalcini programme.

Literature cited

Blinkov, S. M. and I. A. I. Glezer. 1968. The Human Brain in Figures and Tables: A Quantitative Handbook. Basic Books, New York, USA.

Botero, C. A., R. Dor, C. M. McCain, and R. J. Safran. 2013. Environmental harshness is positively correlated with intraspecific divergence in mammals and birds. Molecular Ecology 23:259-268.

Breiman, L. 2001. Random Forests. Machine Learning 45: 5–32.

Buckley, L. B., I. Khaliq, D. L. Swanson, and C. Hof. 2018. Does metabolism constrain bird and mammal ranges and predict shifts in response to climate change? Ecology and Evolution 8:12375–12385.

Ernest, S. K. M, S. 2003. Life history characteristics of placental nonvolant mammals. Ecology 84:3402.

Faurby, S., M. Davis, R. Pedersen, S. D. Schowanek, A. Antonelli, and J. C. Svenning. 2018. PHYLACINE 1.2: The Phylogenetic Atlas of Mammal Macroecology. Ecology 99:2626.

Gainsbury, A. M., O. J. S. Tallowin, and S. Meiri. 2018. An updated global data set for diet preferences in terrestrial mammals: testing the validity of extrapolation. Mammal Review 48:160–167.

Hayssen, V., A. V. Tienhoven, and A. V. Tienhoven. 1993. Asdell's Patterns of Mammalian Reproduction: A Compendium of Species-Specific Data. Cornell University Press, Ithaca, New York, USA.

Heldstab, S. A., K. Isler, and C. P. van Schaik. 2018. Hibernation constrains brain size evolution in mammals. Journal of Evolutionary Biology 31:1582–1588.

IUCN. 2020. IUCN Red List of threatened species. Version 2020-2. http://www.iucnredlist.org.

Jones, K. E., J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, W. Sechrest, E. H. Boakes, C. Carbone, C. Connolly, M. J. Cutts, J. K. Foster, R. Grenyer, M. Habib, C. A. Plaster, S. A. Price, E. A. Rigby, J. Rist, A. Teacher, O. R. P. Bininda-Emonds, J. L. Gittleman, G. M. Mace, and A. Purvis. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology 90:2648.

Kissling, W. D., L. Dalby, C. Fløjgaard, J. Lenoir, B. Sandel, C. Sandom, K. Trøjelsgaard, and J. C. Svenning. 2014. Establishing macroecological trait datasets: Digitalization, extrapolation, and validation of diet preferences in terrestrial mammals worldwide. Ecology and Evolution 4:2913–2930.

De Magalhães, J. P., and J. Costa. 2009. A database of vertebrate longevity records and their relation to other life-history traits 22:1770–1774.

Myhrvold, N. P., E. Baldridge, B. Chan, D. Sivam, D. L. Freeman, and S. K. M. Ernest. 2015. An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles. Ecology 96:3109–000.

Nakagawa, S., R. P. Freckleton. 2008. Missing inaction: the dangers of ignoring missing data. Trends in Ecology and Evolution 23:592–596

Nowak, R. M. 1999. Walker's Mammals of the World. 6 edition. Johns Hopkins University Press, Baltimore, Maryland, USA.

Oba, S., M. Sato, I. Takemasa, M. Monden, K. Matsubara, S. Ishii. 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19:2088–2096.

Pacifici, M., L. Santini, M. Di Marco, D. Baisero, L. Francucci, G. Grottolo Marasini, P. Visconti, and C. Rondinini. 2013. Generation length for mammals. Nature Conservation 5:89–94.

Paradis, E., S. Blomberg, B. Bolker, J. Brown, S. Claramunt, J. Claude, H. S. Cuong, R. Desper, G. Didier, B. Durand, J. Dutheil, R. J. Ewing, O. Gascuel, T. Guillerme, C. Heibl, A. Ives, B. Jones, F. Krah, D. Lawson, V. Lefort, P. Legendre, J. Lemon, G. Louvel, E. Marcon, R. McCloskey, J. Nylander, R. Opgen-Rhein, A. A. Popescu, M. Royer-Carenzi, K. Schliep, K. Strimmer and D. de Vienne. 2020. ape: Analyses of Phylogenetics and Evolution. R package version 5.4.

R Development Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ruf, T. and F. Geiser. 2015. Daily torpor and hibernation in birds and mammals. Biological Reviews 90:891–926.

Santini, L., M. Di Marco, P. Visconti, D. Baisero, L. Boitani, and C. Rondinini. 2013. Ecological correlates of dispersal distance in terrestrial mammals. Hystrix 24:181–186.

Santini L., N. J. B. Isaac and G. F. Ficetola. 2018. TetraDENSITY: a database of population density estimates in terrestrial vertebrates. Global Ecology and Biogeography 27:787-791.

Santos, T., J. A. F. Diniz-Filho, T. F. Rangel and L. M. Bini. 2018. PVR: Phylogenetic Eigenvectors Regression and Phylogenetic Signal-Representation Curve. R package version 1.0.

Smith, F. A., S. K. Lyons, S. K. M. Ernest, K. E. Jones, D. M. Kaufman, T. Dayan, P. Marquet, J. H. Brown, and J. P. Haskell. 2003. Body size of late Quaternary mammals. Ecology 84:3403.

Stekhoven, D. J., and P. Buehlmann. 2012. MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics 28: 112-118.

Stekhoven, D. J. 2016. missForest – package: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4.

Tsuboi, M., W. van der Bijl, B. T. Kopperud, J. Erritzøe, K. L. Voje, A. Kotrschal, K. E. Yopak, S. P. Collin, A. N. Iwaniuk, and N. Kolm. 2018. Breakdown of brain–body allometry and the encephalization of birds and mammals. Nature Ecology and Evolution 2:1492–1500.

Turbill, C., C. Bieber, and T. Ruf. 2011. Hibernation is associated with increased survival and the evolution of slow life histories among mammals. Proceedings of the Royal Society B: Biological Sciences 278:3355–3363.

Usui, T., S. H. M. Butchart, and A. B. Phillimore. 2017. Temporal shifts and temperature sensitivity of avian spring migratory phenology: a phylogenetic meta-analysis. Journal of Animal Ecology 86:250–261.

Wilman, H., J. Belmaker, J. Simpson, C. de la Rosa, M. M. Rivadeneira, and W. Jetz. 2014. EltonTraits 1.0: species-level foraging attributes of the world's birds and mammals. Ecology 95:2027.