Journal of Animal Ecology

British Ecological Society

Journal of Animal Ecology 2012, **81**, 1211–1222 doi: 10.1111/j.1365-2656.2012.01999.x

Biases in comparative analyses of extinction risk: mind the gap

Manuela González-Suárez*, Pablo M. Lucas and Eloy Revilla

Department of Conservation Biology, Estación Biológica de Doñana (EBD-CSIC) Calle Américo Vespucio s/n, 41092 Sevilla, Spain

Summary

1. Comparative analyses are used to address the key question of what makes a species more prone to extinction by exploring the links between vulnerability and intrinsic species' traits and/or extrinsic factors. This approach requires comprehensive species data but information is rarely available for all species of interest. As a result comparative analyses often rely on subsets of relatively few species that are assumed to be representative samples of the overall studied group.

2. Our study challenges this assumption and quantifies the taxonomic, spatial, and data type biases associated with the quantity of data available for 5415 mammalian species using the freely available life-history database PanTHERIA.

3. Moreover, we explore how existing biases influence results of comparative analyses of extinction risk by using subsets of data that attempt to correct for detected biases. In particular, we focus on links between four species' traits commonly linked to vulnerability (distribution range area, adult body mass, population density and gestation length) and conduct univariate and multivariate analyses to understand how biases affect model predictions.

4. Our results show important biases in data availability with *c*.22% of mammals completely lacking data. Missing data, which appear to be not missing at random, occur frequently in all traits (14–99% of cases missing). Data availability is explained by intrinsic traits, with larger mammals occupying bigger range areas being the best studied. Importantly, we find that existing biases affect the results of comparative analyses by overestimating the risk of extinction and changing which traits are identified as important predictors.

5. Our results raise concerns over our ability to draw general conclusions regarding what makes a species more prone to extinction. Missing data represent a prevalent problem in comparative analyses, and unfortunately, because data are not missing at random, conventional approaches to fill data gaps, are not valid or present important challenges. These results show the importance of making appropriate inferences from comparative analyses by focusing on the subset of species for which data are available. Ultimately, addressing the data bias problem requires greater investment in data collection and dissemination, as well as the development of methodological approaches to effectively correct existing biases.

Key-words: data imputation, extinction risk, life-history traits, phylogenetic generalized linear models, PHYLOPARS

Introduction

An important priority for conservation biology is to understand what makes a species or population more likely to become extinct. A popular and appealing answer is based on comparative analyses that explore the links between species vulnerability to extinction and intrinsic ecological and life-history species' traits (Purvis *et al*. 2000; Fisher & Owens 2004; Cardillo *et al*. 2008; Fritz, Bininda-Emonds & Purvis 2009; Pinsky *et al*. 2011) or extrinsic factors (Kerr & Currie 1995; Forester & Machlis 1996; Cardillo *et al*. 2004). This approach requires large databases describing species traits (or extrinsic factors) in a format suitable for comparative analyses. Compiling such databases takes considerable effort from multiple dedicated researchers, who ideally make their complete databases publicly available allowing future research (Jones *et al*. 2009). However, any efforts to gather information are limited by the fact that data are not

*Correspondence author. E-mail: manuela.gonzalez@ebd.csic.es

available for all species in all locations, something that has been previously recognized by other authors (Fisher, Blomberg & Owens 2003; Luck 2007; Nakagawa & Freckleton 2008; Matthews *et al.* 2011). As a result, gathered data represent only a subset of species and locations, which traditional comparative analyses implicitly assume are a representative sample of the taxon or group of interest (but see Matthews *et al.* 2011). Our study challenges this assumption testing the hypothesis that studied species, those for which data are available, are not a random sample of the global biodiversity and that this bias affects results from comparative studies. In particular, we address three objectives: (i) to describe existing biases associated with the number and type of data available in a mammalian comparative data set; (ii) to test the hypothesis that life-history, ecological and behavioural traits are associated with greater data availability, because some traits can facilitate, or complicate, research and make species more or less appealing as study subjects (Matthews *et al.* 2011); and (iii) to investigate whether the existing biases affect the results and conclusions of comparative analyses linking intrinsic species' traits and vulnerability to extinction. Specifically, we compare results from standard phylogenetically informed comparative analyses based on different subsets of species, some of which attempt to control biases.

Currently available tools for comparative analyses can be broadly classified into phylogenetic and non-phylogenetic regressions (Bielby *et al.* 2010). Phylogenetic methods are more commonly used and include regressions using phylogenetic independent contrasts (Felsenstein 1985), a popular approach despite its unrealistic assumptions about Brownian trait evolution (Blomberg, Garland & Ives 2003), and generalized regressions, such as phylogenetic generalized least square models (PGLSs, Martins & Hansen 1997), which provide a flexible alternative with fewer assumptions. Non-phylogenetic methods include regression trees (Breiman 1984) that have fewer data requirements but can be unstable and fail to account for phylogenetic relationships (Bielby *et al.* 2010). All of these tools are limited by data availability because they generally require complete data for all predictors. Therefore, exploring patterns with multiple predictors requires either interpolating missing data, which may introduce biases if data are not missed at random (Little & Rubin 2002), or eliminating all species with any missing data, which can bias estimates and reduces the sample size considerably (Nakagawa & Freckleton 2008). For example, a well-cited study by Cardillo *et al.* (2006) drew inferences from $< 20\%$ of the extant species in some analyses, while analyses for this study were in some cases limited to $< 12\%$ of the species of interest. If those species with available data are not a random sample, conclusions may not apply to the broad group of interest and inferences need to be made carefully.

In recent years, many authors have contributed to develop large databases suitable for comparative analyses, which describe life-history traits in diverse taxa including birds, mammals, amphibians, fish and angiosperms (Froese & Pauly 2000; Sekercioglu, Daily & Ehrlich 2004; Bielby *et al.* 2008; Sodhi *et al.* 2008; Jones *et al.* 2009). For this study, we decided to focus on mammals for several reasons. First, mammals are arguably the best-studied group with many species of conservation, economic and social interest. Second, the links between species traits and vulnerability have been extensively investigated in mammals with multiple comparative studies published showing how traits such as adult body mass, distribution range area, gestation length and population density are linked to vulnerability to extinction (Purvis *et al.* 2000; Fagan *et al.* 2001; Brashares 2003; Cardillo 2003; Cardillo *et al.* 2004, 2005, 2006, 2008; Davidson *et al.* 2009; Fritz, Bininda-Emonds & Purvis 2009). Finally, we had free online access (http://www.utheria.org/) to a large mammalian life-history data set, PanTHERIA (Jones *et al.* 2009), which was also used in several recent comparative studies (e.g., Bininda-Emonds *et al.* 2007; Cardillo *et al.* 2008; Davies *et al.* 2008; Fritz, Bininda-Emonds & Purvis 2009).

In this study, we show that species' ecology, life history and morphology explain variation in the quantity of data collected. Data appear to be not missing at random, and thus applying imputations techniques to fill data gaps may be difficult. Moreover, existing biases affect estimates obtained from comparative analyses suggesting the predictive ability of currently used models may be limited. Although our results are limited to mammalian species, the existence of data biases that can affect comparative analyses is likely common to other taxa. Overall, these findings highlight the importance of explicitly considering data biases in comparative analyses and ultimately, the need for gathering and publishing basic natural history data even if currently deemed 'old-fashion'.

Materials and methods

DATA BASE

PanTHERIA is comprised of two files: the median data set and the raw data file. The median data set includes an entry for each of the 5415 mammalian species recognized by the Wilson & Reeder's (2005) taxonomy with calculated median values for 30 variables describing morphology, development, reproduction, ecology and spatial data (Jones *et al.* 2009). These median values were calculated from a varying number of estimates gathered from the literature (Jones *et al.* 2009). The raw data file includes these individual literature estimates, which we used to estimate the total number and the type of data available for each mammalian species. Because we expect entries in the raw data file to represent a reasonably random sample of the literature, we equate more data entries with more available published data. Certainly this relationship is likely not exact because some published sources are easier to access than others and, as Jones *et al.* (2009) discuss, the database may include some duplicate entries. However, we assume any bias associated with finding data in the literature and duplicate entries is minor compared with the bias in data collection and publication.

In the raw data files, species names were tracked onto the Wilson & Reeder mammalian taxonomy (2005) based on the synonyms file provided by Jones *et al.* (2009). The final file includes all 5415 extant mammalian species, but 1211 species have zero data entries (i.e. no literature records were available for those species).

ARE THERE BIASES IN DATA AVAILABILITY?

Taxonomic and phylogenetic bias

Phylogenies capture the evolutionary history of a group of species better than taxonomy, but generally there is also more uncertainty associated with phylogenies because of unresolved or inconsistent relationships among taxa. For that reason, we explored the potential for biases, that is, related species having greater similarity in the total number of data entries than expected by chance, using both taxonomy, as defined by Wilson & Reeder (2005), and phylogeny, based on the best date estimates of the mammalian supertree (Fritz, Bininda-Emonds & Purvis 2009). Because tip branches in the phylogeny were not fully resolved, we generated 10 trees with randomly resolved polytomies using the procedure MULTI2DI (package APE in R, Development Core Team 2011). Parameter estimates were identical for all trees indicating that how polytomies were resolved did not influence results.

We used nonparametric Kruskal–Wallis tests to compare data counts among orders, families and genera. We addressed the question of phylogenetic bias using the parameter λ (Pagel 1999a), which characterizes the phylogenetic correlation in the total number of data entries available per species (log-transformed). We used the procedures CORPAGEL (package APE in R) and GLS (NLME package in R) to define an intercept-only model with data availability as the dependent variable. Following Freckleton, Harvey & Pagel (2002), we compared log likelihood estimates to determine whether the estimate of λ was significantly different from 0 (0 indicates no phylogenetic correlation in the data).

Other biases

To assess biases in the type of data, we grouped the original 25 variables listed by the raw data file into five data groups: Ecology, Morphology, Development, Reproduction and Spatial (Table S1 in Supporting Information), and compared the average number of entries per variable per species among groups. We explored biases in data availability related to threat category using the 2008 Red List classification (International Union for Conservation of Nature 2010). Threat classification was available for 5288 species in our data base, including 731 listed as Data Deficient. Finally, to explore spatial biases, we obtained global distribution maps of terrestrial mammals from the IUCN spatial database (International Union for Conservation of Nature 2010). We used data from the 4847 species recognized by Wilson & Reeder (2005) and with range areas defined as presence 'extant' or 'probably extant'. Maps were projected in the cylindrical equal area projection and onto a grid equivalent to $2° \times 2°$ near the equator (Hurlbert & Jetz 2007). For each grid cell, we calculated the following: species richness, as the total number of distinct species' ranges overlapping any area of the cell; data richness, as the mean number of data entries per species occupying the cell; and the coefficient of variation in the number of data entries among all species occupying the cell. Data richness reflects the average data availability expected for any species occupying a cell, whereas the coefficient of variation indicates the difference in data availability among species within the same cell.

DO SPECIES TRAITS EXPLAIN THE BIAS IN DATA AVAILABILITY?

To explore if intrinsic species traits could explain data availability, we used 28 variables describing life-history, behavioural and ecological traits, and an estimate of distribution range area provided in the median data set (Jones *et al.* 2009). Analysing these data presented us with a series of challenges. First, because of abundant missing data in the 29 variables considered, using AIC model selection approaches was not possible as models based on different data sets are not comparable (Burnham & Anderson 2002). Therefore, we initially defined univariate models to explore relationships between data availability and species' traits using all available data for each trait. We then defined a multivariate model based on a reduced data set, which included only data-rich (<950 missing cases. Table 1) and non-highly correlated variables to reduce the effects of collinearity ($r < 0{\cdot}80$, Variance Inflation Factor, VIF < 5. Table S2). The reduced data set includes 10 quantitative and three categorical variables that describe intrinsic species traits but represents only a limited number of species (as many data are missing). Thus, multivariate results may not reflect patterns common to all mammals.

A second challenge was whether, and how, to incorporate non-independence of species data because of evolutionary relationships (McNab 2003; Purvis 2008). We followed three different approaches using PGLSs, taxonomically corrected generalized linear mixed models (GLMMs) and non-corrected regression trees. Phylogenetic correction is generally preferable to taxonomic correction (if a good phylogeny exists); however, our estimate of data availability best fits a negative binomial distribution which, to our knowledge, cannot be modelled using frequentist phylogenetic models. Thus, we defined PGLSs log-transforming our dependent variable (using procedures CORPAGEL and GLS in R). However, analyses based on the transformed dependent variables are problematic (O'Hara & Kotze 2010), thus we also fitted GLMMs including taxonomic random effects (nested effects of order, family and genus) with a negative binomial distribution (in SAS 9·2 SAS Institute Inc., Cary, NC, USA). We fitted univariate and multivariate PGLSs and GLMMs. Finally, using the complete data set, we built a regression tree with the procedure RPART (package RPART in R), log-transforming the number of data available to meet the normality requirement. Missing data were handled with surrogate splits as described by Breiman (1984). The tree was pruned using 40 sets of 10-fold cross-validations to produce an optimal tree based on the modal number of splits corresponding to the 1 SE rule (Breiman 1984; De'ath & Fabricius 2000). By comparing the results from all three alternative, imperfect methods, we aimed to assess the overall agreement and ideally identify some general (non-method dependent) patterns to explain data availability.

HOW DO DATA BIASES AFFECT COMPARATIVE ANALYSES?

To understand how data limitations affect our understanding of the links between intrinsic species traits and vulnerability to extinction, we explored how results from a standard comparative approach, PGLSs, differed among distinct data sets. For simplicity, we selected a priori four species' traits that have been consistently linked to vulnerability to extinction in mammals: adult body mass, distribution range area, gestation length and population density (see Introduction for a reference list). We defined a group including all species with estimates available for all four traits in the median data set of PanTHERIA ($N = 636$). This group represents the total sample available for multivariate regression analyses based on the four traits. We compared this group with two other data sets: the PanTHERIA data set, which includes all data available for each trait (see Table 1 for number of species per trait), and an imputed data set, with data on all four traits for 5016 mammalian species. The imputed data set was populated using the phylogenetic data imputation technique

Table 1. Coefficient estimates for univariate GLMMs and phylogenetic generalized least square models (PGLSs) describing the number of data entries available in the mammalian database PanTHERIA as a function of intrinsic species traits. *N* indicates number (and percentage) of species with available data for each trait from the total 5415 mammals studied

		Coefficient (SE)	
Variable	*N* (%)	GLMMs	PGLSs
Activity cycle	1657 (30·6)		
Nocturnal	732	–	–
Crepuscular, cathemeral	486	0·57 (0·067)**	0·45 (0·071)**
Diurnal	439	0·32 (0·081)**	0·13 (0·094)
Terrestriality	2634 (48·6)		
Fossorial	1144	–	–
Above-ground	1490	−0·29 (0·062)**	0·06 (0·075)
Trophic level	2159 (39·9)		
Herbivore	781	–	–
Omnivore	739	0·20 (0·058)**	0·23 (0·061)**
Carnivore	639	−0·29 (0·069)**	−0·26 (0·088)*
Neonate body mass[a]	1083 (20·0)	0·09 (0·027)**	0·13 (0·031)**
Weanling body mass[a]	487 (9·0)	0·13 (0·012)**	0·09 (0·045)[†]
Adult body mass[a]	3539 (65·4)	0·30 (0·020)**	0·27 (0·036)**
Neonate head–body length[a]	226 (4·2)	0·24 (0·090)*	0·31 (0·123)*
Weanling head–body length[a]	47 (0·9)	0·40 (0·222)	0·28 (0·264)
Adult head–body length[a]	1939 (35·8)	0·64 (0·065)**	0·70 (0·123)**
Adult forearm length[a]	903 (16·7)	1·80 (0·283)**	1·29 (0·300)**
Teat number[a]	639 (11·8)	0·22 (0·224)	0·57 (0·243)*
Age at eye opening[a]	474 (8·8)	−0·01 (0·028)	−0·22 (0·139)
Weaning age[a]	1161 (21·4)	0·21 (0·057)**	−0·06 (0·084)
Sexual maturity age[a]	1049 (19·4)	0·22(0·050)**	−0·06 (0·082)
Age at first birth[a]	445 (8·2)	0·22 (0·073)*	0·15 (0·114)
Dispersal age[a]	143 (2·6)	0·09 (0·116)	0·06 (0·138)
Maximum longevity[a]	1011 (18·7)	0·42 (0·068)**	0·81 (0·102)**
Gestation length[a]	1359 (25·1)	0·19 (0·058)*	0·07 (0·115)
Interbirth interval[a]	695 (12·8)	0·18 (0·066)*	0·07 (0·111)
Litter size[a]	2498 (46·1)	0·08 (0·091)	0·93 (0·122)**
Litters per year[a]	893 (16·5)	0·10 (0·122)	0·40 (0·154)*
Diet breadth	2159 (39·9)	0·16 (0·015)**	0·14 (0·014)**
Habitat breadth	2722 (50·3)	0·30 (0·048)**	0·30 (0·045)**
Group size[a]	388 (7·2)	0·12 (0·044)*	0·19 (0·050)**
Social group size[a]	705 (13·0)	0·42 (0·072)**	0·46 (0·089)**
Population density[a]	954 (17·6)	−0·09 (0·021)**	0·02 (0·034)
Home range[a]	705 (13·0)	0·07 (0·018)**	0·11 (0·029)**
Home range individual[a]	624 (11·5)	0·07 (0·019)**	0·10 (0·030)**
Distribution range area[a]	4664 (86·1)	0·64 (0·016)**	0·22 (0·006)**

[†] $P < 0{\cdot}10$, * $P < 0{\cdot}05$, ** $P < 0{\cdot}001$.
[a] Log_{10}-transformed.

implemented in the program PHYLOPARS (Bruggeman, Heringa & Brandt 2009). We used available data from PanTHERIA assuming allometric relationship among traits and the phylogeny supertree mentioned earlier (Fritz, Bininda-Emonds & Purvis 2009). The supertree describes phylogenetic relationships for 5016 species listed by Wilson & Reeder (2005), and thus data for the remaining 399 species could not be imputed using this approach. When running PHYLOPARS, we assumed no phenotypic variation, that is, no measurement error, to avoid re-estimation of already available data. Leave-one-out cross-validation analyses were used to estimate bias (mean differences between observed and estimated values) and absolute error (mean of the absolute differences between observed and estimated) for each trait. Both bias (−0·005 to 0·002) and absolute errors (0·06–0·88) were generally low.

To compare data sets, we first plotted the distribution of values in each of the four traits for the subset group, the PanTHERIA data set and the imputed data set. Second, using PGLSs, we estimated the relationship between the four traits and vulnerability to extinction as defined by the IUCN Red List category (International Union for Conservation of Nature 2010). We used the same phylogenetic supertree with randomly resolved polytomies. For each trait, we defined univariate PGLSs for the sample of 622 species with data available for all four traits, with a Red List category (not including Data Deficient) and represented in the phylogenetic tree (henceforth the 'multivariate subset'). To explore the range of expected values, univariate PGLSs were also defined for 500 samples of 622 species each. Species were selected at random from the imputed data set. This analysis was repeated drawing random samples from the PanTHERIA data set. Following Purvis *et al.* (2000), the Red List categories were converted to a continuous index: Least Concern = 0, Near Threatened = 1, Vulnerable = 2, Endangered = 3, Critically Endangered = 4 and Extinct in the Wild/Extinct = 5. For analyses

that included data on range area, we removed species listed under IUCN Red List criteria B (small geographic range or area of occupancy) to avoid circularity. This resulted in a subset of 584 species which were compared to 500 random samples of 584 species each.

We also defined multivariate PGLSs including all four traits for different subgroups of the multivariate subset (drawing 300 random replicates per subgroup). Each subgroup was defined to conform to the distribution of values observed for a given trait in the imputed data set, while trying to maximize the number of species per subgroup (the number was limited when the distributions were very dissimilar). Defining a subgroup that conformed to all four trait distributions at once was not possible; thus, we defined four subgroups each conforming to a single trait and ran four separate multivariate PGLSs. We compared parameter estimates of these subgroups with those obtained from a multivariate PGLS based on the entire multivariate subset. Sampling to conform to the distribution of a given trait often altered the distribution of the other traits and in some cases increased the deviation from the general distribution observed with all data. This analysis was repeated conforming to the distribution of values in the PanTHERIA data set (non-imputed data).

Results

ARE THERE BIASES IN DATA AVAILABILITY?

We found no data for 1211 (22·4%) of the 5415 extant mammalian species recognized by Wilson & Reeder (2005), whereas 438 species have a single entry in the raw data file, and 401 have two entries. On the other hand, the highest number of data entries is 443 for the deer mouse (*Peromyscus maniculatus*, Rodentia). Missing data are prevalent, occurring in all studied traits, often at high frequencies (14–99% of missing cases, Table 1), and with no species having complete data for all 29 studied traits (all species are missing information for at least one trait).

Taxonomic and phylogenetic bias

The number of data entries per species differs among orders (Kruskal–Wallis test, $\chi^2 = 880{\cdot}1$, d.f. = 28, $P < 0{\cdot}0001$. Fig. S1), families ($\chi^2 = 1351{\cdot}1$, d.f. = 152, $P < 0{\cdot}0001$) and genera ($\chi^2 = 2463{\cdot}0$, d.f. = 1229, $P < 0{\cdot}0001$). Species without data belong primarily to the orders Rodentia and Soricomorpha (no data for 29 and 46% of their species, respectively). On the other hand, among carnivores and ungulates < 7% of species have no data. Analyses based on phylogenetic relationships also indicate similar data availability among closely related species. The estimate of the parameter λ (0·510), which characterizes the degree of phylogenetic correlation, is significantly different from 0 ($\chi^2 = 1379{\cdot}8$, d.f. = 5015, $P < 0{\cdot}001$).

Other biases

More information is available for some types of data than others (Kruskal–Wallis test, $\chi^2 = 3851{\cdot}6$, d.f. = 4, $P < 0{\cdot}0001$. Fig. S2). Generally, morphological data are the most abundant (1·86 ± 2·34, mean ± SD number of entries per species), followed by reproduction (1·27 ± 2·44), ecology (0·94 ± 1·61), development (0·66 ± 1·40) and spatial data (0·33 ± 1·19). However, there is some variation among taxa in the relative abundance of each data type (e.g. spatial data are not the least abundant for the Erinaceomorpha; Fig. S2). We also find differences in data availability among the different threat categories defined by the IUCN Red List (Kruskal–Wallis test, $\chi^2 > 502{\cdot}5$, d.f. = 3, $P < 0{\cdot}0001$. Fig. S3). In particular, the number of entries per species is higher for non-threatened species (classified as Least Concern or Near Threatened, $N = 3420$ species) than for threatened species (classified as Critically Endangered, Endangered or Vulnerable, $N = 1070$). In fact, there is no data available for 29% of the mammalian species classified as Endangered or Critically Endangered, whereas only 16% of the Least Concern and 14% of the Near Threatened species lack data (23% of Vulnerable species lack data). Not surprisingly, there are fewer data for species classified as Data Deficient ($N = 731$) or Extinct ($N = 67$), with many species in these categories having no data at all (39% and 48%, respectively).

Finally, there is evidence of spatial biases in data availability with species living in the northern hemisphere, particularly in North America and Europe, being considerably better studied than those in tropical or southern regions (Fig. 1b). This contrasts with the pattern of species richness (Fig. 1a), so that, on average, there are fewer studies per species in those areas with the highest diversity of mammals. Moreover, less-studied areas often have high variation in data availability among species (Fig. 1c) and include relatively high percentages (20–60%) of species with < 3 data entries (very poorly studied species).

DO SPECIES TRAITS EXPLAIN THE BIAS IN DATA AVAILABILITY?

Univariate analyses identify many traits as significant in explaining data availability (Table 1). Results from GLMMs and PGLSs are generally similar, although some traits are identified as significant under one approach but not the other (Table 1). Nevertheless, there are no contradictory results, that is, no traits identified as significant have opposite estimated effects. Results from the multivariate models, based on data from 266 species, are very similar using taxonomic and phylogenetic correction, although significance is marginal for some traits using phylogenetic correction (Table 2). Both approaches suggest that the number of data entries is generally greater for diurnal or crepuscular mammals with larger body mass, bigger litter sizes, earlier sexual maturation age and longer life spans. Species with more data also have a wider distribution range area and live above-ground at higher population densities.

Regression coefficients are generally similar in univariate and multivariate analyses, although as expected, some variables are significant in the univariate analyses but not in the multivariate models (Table 1). In addition, there is a qualitative change in the estimated effect based on GLMMs for

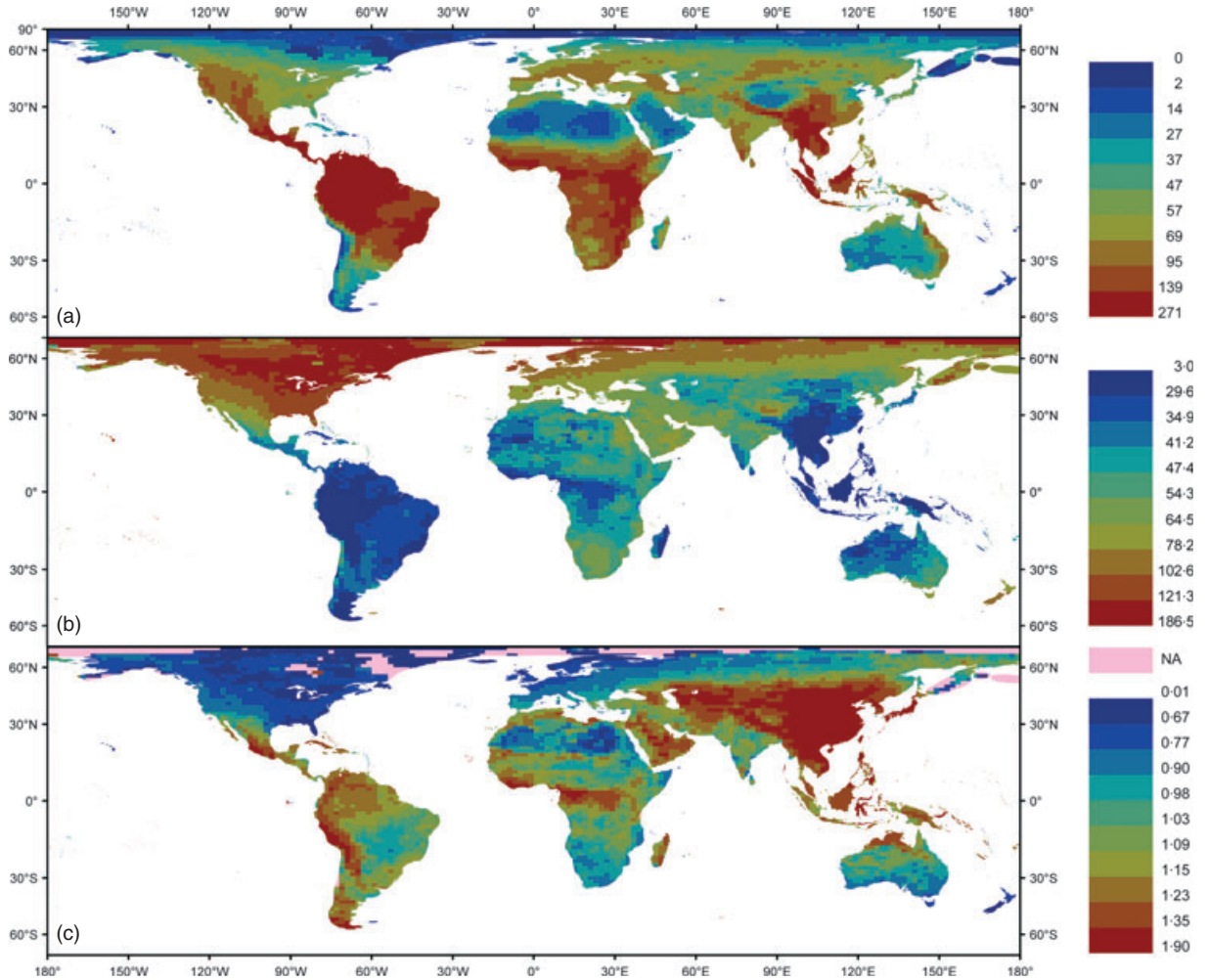

Fig. 1. Global distribution of terrestrial mammalian species and data availability. Panels: (a) species richness calculated as the total number of terrestrial mammals per cell; (b) data richness calculated as the mean number of data entries per terrestrial mammal per cell; (c) coefficient of variation in the number of data entries per terrestrial mammal per cell. Species range distribution maps were obtained from the IUCN spatial data base. Grid size is equivalent to 2° × 2° near the equator.

three traits (terrestriality, sexual maturity age and population density). Noticeably, these traits are not significant in the univariate PGLSs. To test whether the changes are simply because of changes in sample size and composition (i.e. which species are represented), we re-ran univariate GLMMs analyses using only the 266 species included in the multivariate models. Using these data, all three coefficients are not significantly different from zero suggesting there is an effect of sample size and composition, but also a possible interaction among variables in the multivariate model that leads to significant coefficients. The non-corrected approach based on regression trees gives similar results, although fewer traits are associated with data availability. The final tree reveals that the highest mean number of data entries is associated with species having larger distribution areas and greater body mass (Fig. 2). On the other hand, species with smaller distribution area, small body mass and reduced habitat breadth have the fewest data entries. The tree explains 34·9% of the variance in the number of entries (calculated as 1-relative error). Overall, based on the three different approaches, two variables appear as clearly linked with data availability: adult body mass and distribution range area. Other traits such as activity pattern, terrestriality, sexual maturity age, maximum longevity, litter size, population density and habitat breadth are also likely relevant predictors.

Although these analyses had to be limited to species with trait data, the main observed relationships appear to extend to those species missing data. For example, there is a strong and positive relationship between the median adult body mass in each taxonomic family (calculated from available species data) and the proportion of species with body mass data in that family (Spearman correlation $r = 0{\cdot}39$, $P < 0{\cdot}001$, $N = 153$ families; for families with 10 or more species $r = 0{\cdot}45$, $P < 0{\cdot}001$, $N = 69$ families). In other words, families with the smaller, on average, species such as rodents or shrews have fewer species with body mass estimates available. Therefore, missing data appear to not be missing at random, but rather the likelihood of having information on a given trait is affected by the trait value itself (smaller species are less likely to have body mass estimates).

Table 2. Coefficient estimates for multivariate GLMMs and phylogenetic generalized least square models (PGLSs) describing the number of data entries available in the mammalian database PanTHERIA as a function of intrinsic species traits. Models are based on 266 species for which data were available (from the total 5415 mammals studied)

	Coefficient (SE)	
Variable	GLMMs	PGLSs
Activity cycle		
Nocturnal	–	–
Crepuscular, cathemeral	0·21 (0·072)*	0·13 (0·075)†
Diurnal	0·24 (0·084)*	0·18 (0·094)†
Terrestriality		
Fossorial	–	–
Above-ground	0·22 (0·088)*	0·17 (0·094)†
Trophic level		
Herbivore	–	–
Omnivore	0·05 (0·094)	0·06 (0·095)
Carnivore	−0·08 (0·132)	0·05 (0·131)
Adult body mass[a]	0·18 (0·061)*	0·19 (0·069)*
Weaning age[a]	0·01 (0·126)	−0·07 (0·145)
Sexual maturity age[a]	−0·33 (0·146)*	−0·29 (0·154)†
Maximum longevity[a]	0·50 (0·19)*	0·66 (0·190)**
Gestation length[a]	−0·05 (0·130)	−0·14 (0·189)
Litter size[a]	0·63 (0·176)**	0·90 (0·195)**
Diet breadth	−0·03 (0·025)	−0·02 (0·023)
Habitat breadth	−0·01 (0·053)	0·03 (0·056)
Population density[a]	0·16 (0·040)**	0·11 (0·042)*
Distribution range area[a]	0·16 (0·039)**	0·19 (0·042)**

†$P < 0{\cdot}10$, *$P < 0{\cdot}05$, **$P < 0{\cdot}001$.
[a]Log_{10}-transformed.

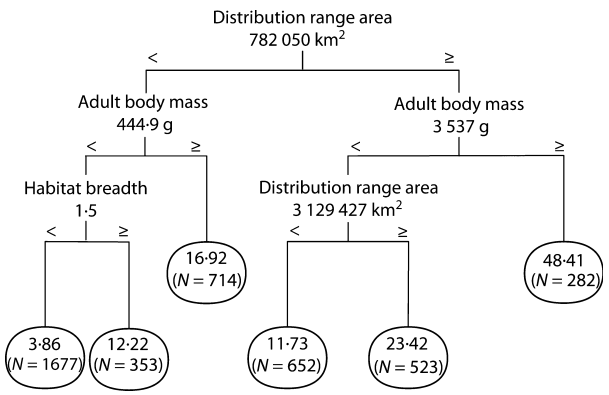

Fig. 2. Regression tree showing the number of data entries available in the mammalian database PanTHERIA based on diverse intrinsic species traits. On each node, the threshold value and name of the splitting trait are indicated. Data on the leaves (represented by circles) provide the average number of data entries and the number of species in the group.

The relationship is not as clear for distribution range area (Spearman correlation $r = 0{\cdot}15$, $P = 0{\cdot}06$, $N = 153$ families), likely because range area is available for many more species (considering only families with species missing data, $r = 0{\cdot}32$, $P = 0{\cdot}02$, $N = 79$ families) and because range is a more flexible 'trait' (less constrained by evolution). In fact, estimates of range area are more variable among species within a family (median CV within families = 0·17) than estimates of adult body mass (0·10).

HOW DO DATA BIASES AFFECT COMPARATIVE ANALYSES?

From the 5415 extant mammals, adult body mass data are only available for 65·4% of the species, range area for 86·1%, population density for 17·6% and gestation length for 25·1%. However, the subset of species with data on all four traits is much smaller, including only 636 species (11·7%). From this subgroup, 622 species (the multivariate subset) also have Red List status and defined phylogenetic relationships; thus, the subset of species available for multivariate PGLSs is quite small compared to the overall mammalian biodiversity. In addition, species with data on all four traits are not a representative sample of all mammals (Fig. 3). These species represent 22 of the 29 mammalian orders (notably excluding all 84 species from the order Cetacea) and 91 of the 153 families. The multivariate subset includes <7% of the species from the most populous families (Muridae, $n = 730$, and Cricetidae $n = 681$) but >51% of the 35 canids (Canidae). In general, the subset includes mammals with higher body mass (following a bimodal distribution), larger range areas, lower population densities and longer gestation periods (Fig. 3).

Univariate PGLSs associating Red List status with each of the four traits show that the parameter estimates obtained for the multivariate subset are generally not representative of the relationships expected for all mammals (Fig. 4) and they would be rarely, if ever, observed when using representative (random) samples. Intercept values estimated using the multivariate subset are higher than those based on all data. Therefore, analyses based on the subset appear to overestimate the baseline extinction risk (measured by Red List status). In addition, although there are no changes in slope sign (the relationships between each trait and Red List status are qualitatively the same), the slope values, which estimate the strength of the relationship, vary. In particular, the increase in Red List status predicted in response to a reduction in range area or population density is more pronounced for the multivariate subset than when considering all data, suggesting that the subset overestimates the influence of these traits on the extinction risk. On the other hand, the multivariate subset appears to underestimate the rate of increase in Red List status associated with longer gestation periods. The estimates of the relationship between body mass and Red List status are similar for all data and the multivariate subset. Results are qualitatively the same when the multivariate subset results are compared with random samples from the PanTHERIA data set (non-imputed data, Fig. S4).

Parameter estimates from multivariate PGLSs defined for subgroups conforming to all mammal distributions differ from those calculated for the multivariate subset (Fig. 5, S5). In particular, conforming to the distributions of all available data for any trait affects the expected relationship between body mass and Red List status. This relationship is

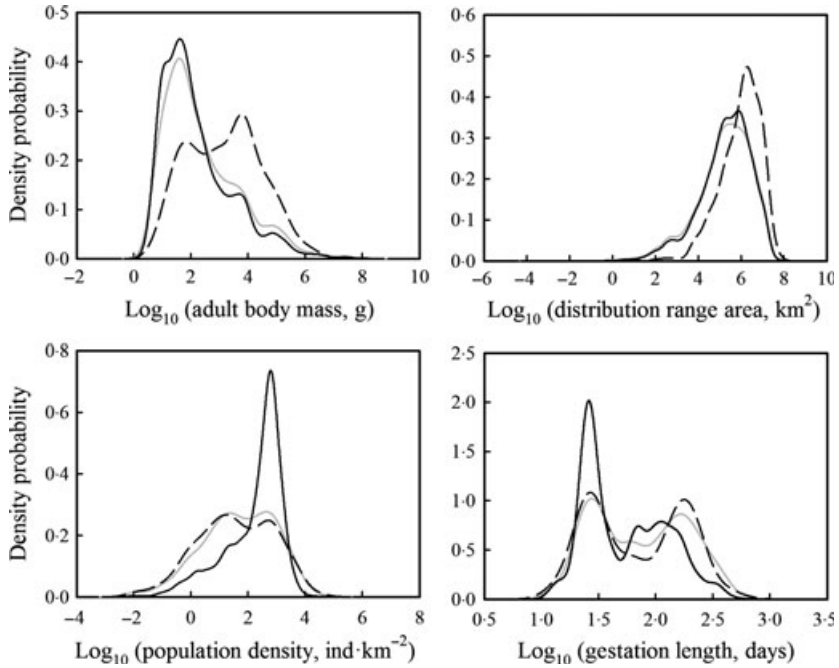

Fig. 3. Distribution of values for four traits consistently linked to vulnerability to extinction in mammals for all species (imputed data set and black solid line), for the PanTHERIA data set (grey solid line), and for the 636 species with data on all four traits (dotted line). Sample sizes for the PanTHERIA data set are as follows: adult body mass $N = 3539$, distribution range area $N = 4664$, population density $N = 954$, gestation length $N = 1359$. Imputed data are available for 5016 species.

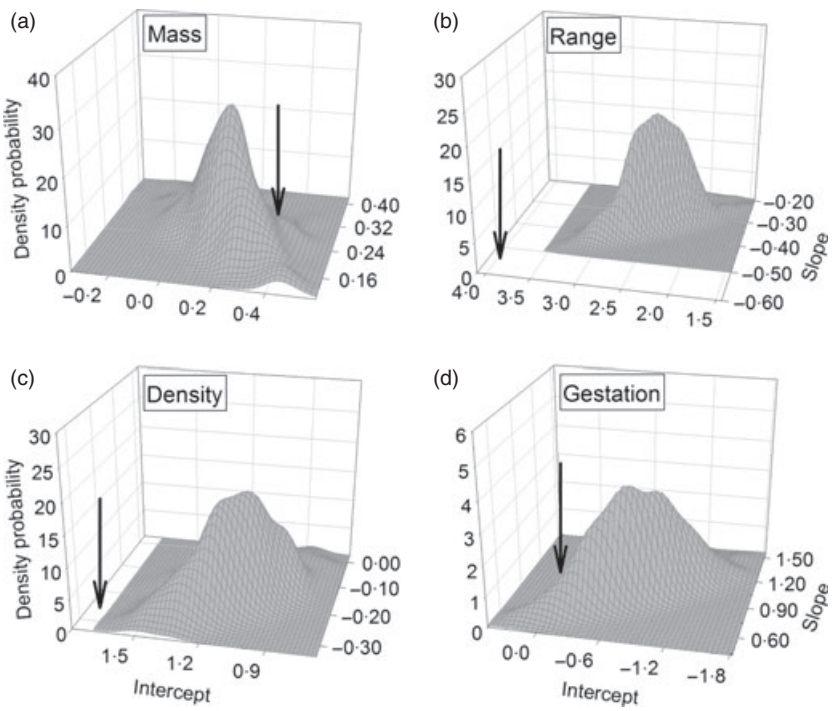

Fig. 4. Values of the intercept and slope coefficients estimated in univariate phylogenetic generalized least square models explaining Red List status as a function of (a) adult body mass, (b) distribution range area, (c) population density and (d) gestation length. The grey surfaces show the distribution of parameter estimates obtained for all mammals (imputed data set) calculated using 500 subsets of 622 species each drawn at random (for range area subsets had 584 species after excluding those listed under the IUCN small range criteria). The black arrows indicate the parameter estimates obtained for the multivariate subset (species with Red List status and data available for all four traits. See Fig. 3). For illustration purposes, the arrow points are placed along the *z*-axis at the point of intersection with the grey surface. Note, the values in the intercept axis in panel (a) are reversed.

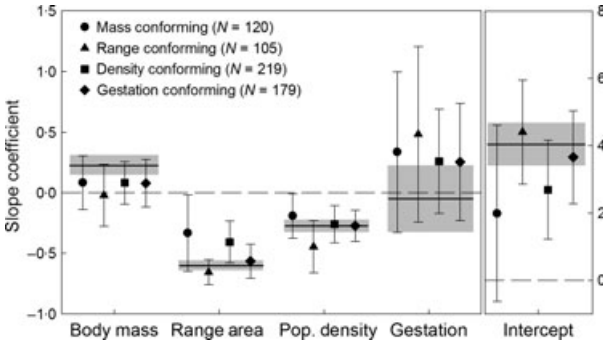

Fig. 5. Multivariate phylogenetic generalized least square models parameter estimates for subgroups of mammals with Red List status and data available for the four traits listed but selected to conform to the distribution of all mammalian data for each trait (imputed data set. See Fig. 3). Symbols are the mean estimate with error bars representing ±2 SD from 300 independent random replicates (*N* in the legend indicates the number of species in each replicate. The sample size was determined to maximize the number of species while conforming to the distribution of all data). Error bars overlapping with zero indicate the relationship between the trait and the Red List status is ambiguous. Horizontal black lines represent the mean parameter estimates of a single PGLS based on the entire multivariate subset (622 species with data on all four traits). Grey bands represent ± 1 SE. Grey bands overlapping with zero indicate a trait is not linked to vulnerability to extinction.

significantly positive for the multivariate subset but not for the conforming subgroups, indicating body mass may not be a good predictor of extinction risk in mammals after all. Similarly, conforming to the distribution of body mass or population density generally reduces the estimated effect of range area on vulnerability to extinction, so that a reduction in range area is not predicted to increase Red List status as rapidly. On the other hand, when conforming to the distribution of range area, the rate at which Red List status increases with a reduction in population density is greater. Finally, although gestation length has been identified as important by several previous studies and our own univariate analyses, the coefficient estimate is not significantly different from zero in any of our multivariate models, suggesting that gestation length may not be strongly associated to Red List status when other factors are taken into account. For other traits, the univariate and multivariate coefficients predict the same general relationships. We found no evidence of collinearity in the multivariate models (VIF < 1·55).

Discussion

ARE THERE BIASES IN DATA AVAILABILITY?

Our results show an important bias in data availability for life-history, ecological and behavioural traits in mammalian species, arguably the best-studied group of organisms. We found that some species are better studied than others and that biases have taxonomic and phylogenetic signals, so that related species have similar data availability (Amori & Gippoliti 2000). As a result, some groups tend to be data-poor, for example, Rodentia, while others are generally well-studied, for example, Artiodactyla (Fig. S1). Surprisingly, we found that non-threatened species are better studied than mammals of conservation concern (Fig. S3), with data completely lacking for nearly one-third of the most threatened species (Endangered and Critically Endangered). Although conservation biologists have been at work for some time, we still know more about common species, possibly those of direct economic importance such as pests or game, than species at risk of extinction.

In addition to biases in the amount of data available, we also identify biases in the type of data gathered. For example, morphological data, e.g., adult body mass, are more often available than spatial data, e.g., home range size (Fig. S2). These biases are likely due to technological constraints, which have limited our ability to track small species such as bats (Holland & Wikelski 2009), and/or financial or logistic limitations associated with obtaining different types of data. Importantly, scarcity of data for some traits likely affects the results of comparative studies because a lack of power in the analyses may limit our ability to recognize traits as important. In fact, we find that traits linked most often to vulnerability to extinction are also those with data for more species and which appear to explain data availability (i.e. body mass and range area).

We also identify important biases in the spatial distribution of data availability (Fig. 1, Amori & Gippoliti 2000). Regions of higher species richness, where mean data availability per species is lowest, largely correspond to tropical areas, where the highest abundance of threatened species also occurs (Schipper *et al*. 2008). In these regions, there is also a greater disparity in data availability among species, so that the limited number of studies conducted concentrate on a few of the species present, likely those easier to study or more attractive, while many species remain poorly known. In contrast, areas with higher data availability and where species are more uniformly studied (lower variation in data availability among them) are predominantly in developed countries where fewer endangered species are found (Schipper *et al*. 2008), but more resources are invested in research leading to more data collection and publication (World Bank 2011).

DO SPECIES TRAITS EXPLAIN THE BIAS IN DATA AVAILABILITY?

Our analyses show that existing biases in data availability are in part explained by intrinsic species traits, which presumably influence the ease and attractiveness of a species as a study organism (Matthews *et al*. 2011). Interestingly, traits associated with higher data availability (Tables 1, 2 and Fig. 2) do not appear to define a single group of species but may represent two general types of well-studied mammals: the big, long-lived mammals occupying large range areas and the smaller mammals with an early maturation age and large litter sizes. The former, for which there is overall the largest amount of data, probably correspond to charismatic

megafauna, such as the giraffe (*Giraffa camelopardalis* with 275 data entries), whereas the second group includes common species with a fast life history more suitable for ecological experiments and manipulations such as the deer mouse (*P. maniculatus* with 443 entries). The bimodal distribution of body mass values in Fig. 3 also supports the existence of two types of best-studied mammals. Interestingly, a bimodal pattern in body mass has been also reported regarding vulnerability to extinction (Cardillo *et al.* 2005; Davidson *et al.* 2009). In fact, these studies describe a threshold around 3–5 kg of body mass, which coincides well with the *c.*3·5 kg split in our regression tree explaining data availability.

HOW DO DATA BIASES AFFECT COMPARATIVE ANALYSES?

We find that the existence of biases in data availability has worrying consequences for comparative analyses. Multivariate analyses that consider several traits associated with vulnerability to extinction are likely limited to a skewed subset of species that are not a representative sample of all mammals. For example, we show that well-studied species appear to be larger, have bigger range areas, longer gestation periods and live at lower population densities than those less studied. Comparative analyses that partly correct these biases give different results than analyses based on the skewed subset, indicating that ignoring existing biases in the data available has consequences for our understanding of how species traits influence vulnerability to extinction. Our study does not imply that previous conclusions are necessarily mistaken or erroneous, but rather raises concern over our ability to accurately make broad inferences with the available data. For example, large body size is perhaps identified as a trait associated with higher risk of extinction because we have data primarily for the big and rare vs. the small and common.

POTENTIAL SOLUTIONS AND RECOMMENDATIONS

We have identified important data biases in the mammalian life-history literature, which appear to reflect a pattern of data 'not missing at random'. That is, the probability of not having information for a trait depends on the unobserved values of that trait (Little & Rubin 2002). This presents a great challenge for analysing these data because as we have seen here deleting species with missing data greatly reduces the available sample size and introduces biases in model estimates. However, conventional techniques to fill gaps (such as multiple imputation) generally assume that data are missing at random or completely at random (Little & Rubin 2002; Nakagawa & Freckleton 2008). For data 'not missing at random', it is possible to use imputation but a clear understanding of the mechanism causing the missing data is generally necessary. However, missing data in PanTHERIA are likely missing as a result of multiple mechanisms. For example, some species may be harder to study because of their life history, while others may simply live in areas where research is complicated by the topology or political situation. In addition, basic research in some areas may be published in non-English journals or in publication formats not as readily available to researchers compiling databases. Therefore, filling data gaps in PanTHERIA using conventional approaches may be challenging.

Alternatively, missing data may be inferred based on expected relationships among traits and phylogenetic relationships (Pagel 1999b; Bruggeman, Heringa & Brandt 2009). We applied this approach in this study, but the method is not without challenges. First, one must have a complete phylogeny, yet phylogenies are rarely complete. Even for well-studied species such as mammals, we found *c.*9% of extant species are not reflected in the most updated phylogeny. Second, these methods assume relatively simple relationships among traits (e.g. allometric) and evolutionary models (i.e. Brownian evolution), which may not be realistic for many ecological and behavioural traits (Blomberg, Garland & Ives 2003). Finally, interpolation based on a skewed sampled may generate biased data sets, so inferences should be made with caution. For example, exploratory analyses with PHYLOPARS (M. González-Suárez, unpublished data) suggest that estimates of body mass for species with missing mass in the order Carnivora ($N = 34$) can differ up to 2 order of magnitude when imputation is performed using only data for small carnivores (≤3 kg) vs. only data for large carnivores (> 3 kg). Imputed values are always larger when estimated from the large carnivore data set. Interestingly, biasing the data set by body mass (imputing data based on large vs. small carnivores) also changes the estimates of missing data for range size, population density and gestation length. The implications of these challenges for our own analyses are that we cannot accurately quantify biases because we cannot truly know mammalian diversity. However, our results are consistent using imputed data or only the available data in PanTHERIA, thus we feel there is strong evidence that biases exist and that we can show their general direction.

In conclusion, our results highlight the need for gathering additional data because many species, even within well-known taxa, are poorly studied and imputing missing data is very challenging. In addition, obtained data must be made available to others (Costello 2009). Efforts such as PanTHERIA (Jones *et al.* 2009), which make published data readily available in a convenient format, are key to understanding general patterns because individual researchers are limited in their ability to gather the large amounts of data needed for broad comparative analyses. As we see it, reducing data biases requires both augmenting data collection and encouraging data dissemination. In addition, comparative analyses need to acknowledge and explicitly address the bias in data availability, making inferences that are appropriate to the available data (i.e. restricted to the subset of species used in the analyses). We suggest that comparative analyses incorporate approaches to explore the consequences of existing biases. For example, as done here, authors may use resampling techniques to incorporate uncertainty, or compare results from univariate and multivariate models as the former may include considerably more species. In addition,

sensitivity analyses of imputed data sets based on diverse missingness-causing mechanisms may be conducted (Little & Rubin 2002). The long-term solution is an increase in data availability, but meanwhile comparative studies should acknowledge the limitations in the existing information by implementing approaches that account for data biases, and authors should be cautious with their conclusions. Analyses may also focus on the best-studied groups, such as ungulates and carnivores, for which more data are available and for which conclusions, even if not as general, may not be as biased. Finally, incorporating extrinsic factors associated with extinction risk is essential to fully understand why some species are more vulnerable than others. However, data on extrinsic factors influencing vulnerability are also likely biased and should be also explored with caution.

Acknowledgements

Our work would not have been possible without all the researchers that contributed to PanTHERIA by conducting and publishing mammalian research and by defining and populating the database. Kate Jones kindly provided us with access to the raw data file. Assistance from Jorn Bruggeman was essential to generate the imputed data set with PHYLOPARS. We would also like to thank Tim Coulson, Miguel Clavero, Miguel Delibes, Peter Thrall and two anonymous reviewers for valuable suggestions to improve earlier versions of this manuscript. This work was funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 235897, the Spanish Ministry of Science and Innovation (CGL2009-07301/BOS and BES-2010-034151).

References

Amori, G. & Gippoliti, S. (2000) What do mammalogists want to save? Ten years of mammalian conservation biology *Biodiver and Conservation*, **9**, 785–793.

Bielby, J., Cooper, N., Cunningham, A.A., Garner, T.W.J. & Purvis, A. (2008) Predicting susceptibility to future declines in the world's frogs. *Conservation Letters*, **1**, 82–90.

Bielby, J., Cardillo, M., Cooper, N. & Purvis, A. (2010) Modelling extinction risk in multispecies data sets: phylogenetically independent contrasts versus decision trees. *Biodiversity and Conservation*, **19**, 113–127.

Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., MacPhee, R.D.E., Beck, R.M.D., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. & Purvis, A. (2007) The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Blomberg, S.P., Garland, T. & Ives, A.R. (2003) Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, **57**, 717–745.

Brashares, J.S. (2003) Ecological, behavioral, and life-history correlates of mammal extinctions in West Africa. *Conservation Biology*, **17**, 733–743.

Breiman, L. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, California, USA.

Bruggeman, J., Heringa, J. & Brandt, B.W. (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, **37**, W179–W184.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York, USA.

Cardillo, M. (2003) Biological determinants of extinction risk: why are smaller species less vulnerable? *Animal Conservation*, **6**, 63–69.

Cardillo, M., Purvis, A., Sechrest, W., Gittleman, J.L., Bielby, J. & Mace, G.M. (2004) Human population density and extinction risk in the world's carnivores. *PLoS Biology*, **2**, 909–914.

Cardillo, M., Mace, G.M., Jones, K.E., Bielby, J., Bininda-Emonds, O.R.P., Sechrest, W., Orme, C.D.L. & Purvis, A. (2005) Multiple causes of high extinction risk in large mammal species. *Science*, **309**, 1239–1241.

Cardillo, M., Mace, G.M., Gittleman, J.L. & Purvis, A. (2006) Latent extinction risk and the future battlegrounds of mammal conservation. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 4157–4161.

Cardillo, M., Mace, G.M., Gittleman, J.L., Jones, K.E., Bielby, J. & Purvis, A. (2008) The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of the Royal Society of London B Biological Sciences*, **275**, 1441–1448.

Costello, M.J. (2009) Motivating online publication of data. *BioScience*, **59**, 418–427.

Davidson, A.D., Hamilton, M.J., Boyer, A.G., Brown, J.H. & Ceballos, G. (2009) Multiple ecological pathways to extinction in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 10702–10705.

Davies, T.J., Fritz, S.A., Grenyer, R., Orme, C.D.L., Bielby, J., Bininda-Emonds, O.R.P., Cardillo, M., Jones, K.E., Gittleman, J.L., Mace, G.M. & Purvis, A. (2008) Phylogenetic trees and the future of mammalian biodiversity. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 11556–11563.

De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Fagan, W.F., Meir, E., Prendergast, J., Folarin, A. & Karieva, P. (2001) Characterizing population vulnerability for 758 species. *Ecology Letters*, **4**, 132–138.

Felsenstein, J. (1985) Phylogenies and the comparative method. *American Naturalist*, **125**, 1–15.

Fisher, D.O., Blomberg, S.P. & Owens, I.P.F. (2003) Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials. *Proceedings of the Royal Society of London B Biological Sciences*, **270**, 1801–1808.

Fisher, D.O. & Owens, I.P.F. (2004) The comparative method in conservation biology. *Trends in Ecology & Evolution*, **19**, 391–398.

Forester, D.J. & Machlis, G.E. (1996) Modeling human factors that affect the loss of biodiversity. *Conservation Biology*, **10**, 1253–1263.

Freckleton, R.P., Harvey, P.H. & Pagel, M. (2002) Phylogenetic analysis and comparative data: a test and review of evidence. *American Naturalist*, **160**, 712–726.

Fritz, S.A., Bininda-Emonds, O.R.P. & Purvis, A. (2009) Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology Letters*, **12**, 538–549.

Froese, R. & Pauly, D. (2000) *FishBase 2000: concepts, design and data sources*. ICLARM, Los Baños, Philippines.

Holland, R.A. & Wikelski, M. (2009) Studying the migratory behavior of individual bats: current techniques and future directions. *Journal of Mammalogy*, **90**, 1324–1329.

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 13384–13389.

International Union for Conservation of Nature (2010) IUCN Red list of threatened species. Version 2010.4. http://www.iucnredlist.org/ (accessed February 08, 2010).

Jones, K.E., Bielby, J., Cardillo, M., Fritz, S.A., O'Dell, J., Orme, C.D.L., Safi, K., Sechrest, W., Boakes, E.H., Carbone, C. *et al*. (2009) PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, **90**, 2648.

Kerr, J.T. & Currie, D.J. (1995) Effects of human activity on global extinction risk. *Conservation Biology*, **9**, 1528–1538.

Little, R.J.A. & Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Wiley, Hoboken, New Jersy, USA.

Luck, G.W. (2007) A review of the relationships between human population density and biodiversity. *Biological Reviews*, **82**, 607–645.

Martins, E.P. & Hansen, T.F. (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, **149**, 646–667.

Matthews, L.J., Arnold, C., Machanda, Z. & Nunn, C.L. (2011) Primate extinction risk and historical patterns of speciation and extinction in relation to body mass. *Proceedings of the Royal Society of London B Biological Sciences*, **278**, 1256–1263.

McNab, B.K. (2003) Standard energetics of phyllostomid bats: the inadequacies of phylogenetic-contrast analyses. *Comparative Biochemistry and Physiology Part A Molecular & Integrative Physiology*, **135A**, 357–368.

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

O'Hara, R.B. & Kotze, D.J. (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Pagel, M. (1999a) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.
Pagel, M. (1999b) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, **48**, 612–622.
Pinsky, M.L., Jensen, O.P., Ricard, D. & Palumbi, S.R. (2011) Unexpected patterns of fisheries collapse in the world's oceans. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 8317–8322.
Purvis, A. (2008) Phylogenetic approaches to the study of extinction. *Annual Review of Ecology and Systematics*, **39**, 301–319.
Purvis, A., Gittleman, J.L., Cowlishaw, G. & Mace, G.M. (2000) Predicting extinction risk in declining species. *Proceedings of the Royal Society of London B Biological Sciences*, **267**, 1947–1952.
Schipper, J., Chanson, J.S., Chiozza, F., Cox, N.A., Hoffmann, M., Katariya, V., Lamoreux, J., Rodrigues, A.S.L., Stuart, S.N., Temple, H.J. *et al.* (2008) The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science*, **322**, 225–230.
Sekercioglu, C.H., Daily, G.C. & Ehrlich, P.R. (2004) Ecosystem consequences of bird declines. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 18042–18047.
Sodhi, N.S., Koh, L.P., Peh, K.S.H., Tan, H.T.W., Chazdon, R.L., Corlett, R.T., Lee, T.M., Colwell, R.K., Brook, B.W., Sekercioglu, C.H. *et al.* (2008) Correlates of extinction proneness in tropical angiosperms. *Diversity and Distributions*, **14**, 1–10.
Wilson, D.E. & Reeder, D.M. (2005) *Mammal Species of the World: A Taxonomic and Geographic Reference*, 3rd edn. Johns Hopkins University Press, Baltimore.
World Bank (2011) Research and development expenditure (% of GDP) http://data.worldbank.org/. accessed: 25/04/2011.

Received 20 February 2012; accepted 9 April 2012
Handling Editor: Tim Coulson

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1. Central tendency (mean and median) and maximum number of data entries per species for each mammalian taxonomic order.

Fig. S2. Availability of different types of data (mean and SD number of data entries per species) for mammalian orders with at least 10 extant species (excluding the 1211 species for which no data was available in any category).

Fig. S3. Number of data entries (mean and SD) for the different categories of threat defined by the IUCN Red List for 5288 mammalian species.

Fig. S4. Values of the intercept and slope estimated in univariate GLMMs explaining Red List status as a function of (A) adult body mass, (B) distribution range area, (C) population density, and (D) gestation length.

Fig. S5. Multivariate PGLSs parameter estimates for subgroups of mammals with Red List status and data available for the four traits listed but selected to conform to the distribution of all available data in PanTHERIA for each trait.

Table S1. Grouping of the original PanTHERIA variables as listed in the raw data file into five groups for our analyses.

Table S2. Correlations among species median estimates of life-history, ecological and behavioural traits obtained from PanTHERIA. All variables were log10-transformed except diet breadth and habitat breadth.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.