Predicting the conservation status of Data Deficient species

Journal:	*Conservation Biology*
Manuscript ID:	Draft
Wiley - Manuscript type:	Contributed Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Bland, Lucie; Zoological Society of London, Institute of Zoology; Imperial College London, Division of Biology Collen, Ben; University College London, Department of Genetics, Evolution and Environment Orme, C. David; Imperial College London, Division of Biology Bielby, Jon; Zoological Society of London, Institute of Zoology
Keywords:	Mammals < Animals, Indicators, Predictive modeling, Red lists, Threatened species, Statistics
Abstract:	We have no appreciation of the level of extinction risk faced by a sixth of the 65,000+ species currently on the IUCN Red List. Determining the status of these Data Deficient (DD) species is essential to developing an accurate picture of global biodiversity and protecting potentially threatened DD species. Using terrestrial mammals as our focal taxon, we compared the outcomes of seven Machine Learning (ML) tools in predicting threat for species of known conservation status using taxonomic, life-history, geographical and threat information. ML tools showed very high species classification accuracy (up to 92%) and ability to correctly identify centres of threatened species richness. Applying the best model to DD species, we predict 313 of 493 DD species (64%) to be at risk, increasing the estimated proportion of threatened terrestrial mammals from 22% to 27%. Regions predicted to contain large numbers of threatened DD species are already conservation priorities, but show considerably higher levels of risk than previously recognized. We conclude that unless directly targeted for monitoring, species classified as DD are likely to slide towards extinction unnoticed. Taking into account information on DD species may therefore help tackle data gaps in biodiversity indicators and conserve the earth’s poorly-known biodiversity.

Predicting the conservation status of Data Deficient species

Running head: Predicting extinction risk

Keywords: IUCN Red List, Data Deficient, Extinction risk, Machine Learning

Word count: 5,950 words

Lucie M. Bland[1,2*], Ben Collen[3], C. David L. Orme[2], and Jon Bielby[1]

[1]Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, UK

[2]Division of Biology, Imperial College London, Silwood Park, Ascot, SL5 7PY, UK

[3]Centre for Biodiversity and Environmental Research, University College London, Gower Street, London, WC1 E6BT, UK

* e-mail: lucie.bland@ioz.ac.uk

Abstract

We have no appreciation of the level of extinction risk faced by a sixth of the 65,000+ species assessed by the IUCN Red List. Determining the status of these Data Deficient (DD) species is essential to developing an accurate picture of global biodiversity and identifying potentially threatened DD species. To address this gap in our knowledge, we used predictive models incorporating species' life-history, geography and threat information to predict the conservation status of DD species within terrestrial mammals. We constructed the models using seven Machine Learning (ML) tools trained on species of known status. The resultant models showed very high species classification accuracy (up to 92%) and ability to correctly identify centres of threatened species richness. Applying the best model to DD species, we predict 313 of 493 DD species (64%) to be at risk, increasing the estimated proportion of threatened terrestrial mammals from 22% to 27%. Regions predicted to contain large numbers of threatened DD species are already conservation priorities, but show considerably higher levels of risk than previously recognized. We conclude that unless directly targeted for monitoring, species classified as DD are likely to slide towards extinction unnoticed. Taking into account information on DD species may therefore help tackle data gaps in biodiversity indicators and conserve the earth's poorly-known biodiversity.

Introduction

In light of global biodiversity change, the 12th target of the Strategic Plan of the Convention on Biological Diversity (CBD) states that by “2020 the extinction of known threatened species has been prevented” (Convention on Biological Diversity 2010). Understanding the level of extinction risk faced by different species, and why interspecific differences in risk arise are therefore some of the greatest challenges facing conservation biology. Assessment frameworks for threatened species are crucial to identifying risk and monitoring progress towards CBD targets (Jones et al. 2011), and one of the most widely used is the International Union for Conservation of Nature (IUCN) Red List (IUCN 2001; Butchart et al. 2010).

There has been much improvement in the taxonomic coverage of the Red List over recent years, resulting in a more comprehensive understanding of species’ extinction risk (Collen & Bailie 2010; Böhm et al. 2013). However, a sixth of the 65,000+ species assessed by the IUCN are classified as Data Deficient (DD) due to a lack of information on taxonomy, geographic distribution, population status or threats (IUCN 2010). To date 15% of mammals (Schipper et al. 2008), 25% of amphibians (Stuart et al. 2004), 19% of reptiles (Böhm et al. 2013) and 49% of freshwater crabs (Cumberlidge et al. 2009) are classified as DD. Uncertainty within many groups about the true level of extinction risk of DD species considerably influences our understanding of patterns of threat and risk (Butchart & Bird 2010; Bland et al. 2012), as the distribution of DD species is often taxonomically and spatially biased (Bielby et al. 2006; Bland et al. 2012). For example, 25% of data-sufficient mammals are threatened with extinction, but estimates range from 21% if all DD species were non-threatened to 36% if all DD species were threatened (Hilton-Taylor et al. 2009). In addition, genuinely threatened

DD species may be neglected by conservation programmes due to their uncertain extinction risk status.

Determining the true conservation status of DD species is essential in developing an accurate picture of global biodiversity and enabling the protection of threatened species. Re-assessment of the 10,673 species currently classified as DD to a data-sufficient category could be achieved through focused field surveys, but the prospect of this occurring is unlikely given the monetary and time costs of biodiversity surveys (Balmford & Gaston 1999) and current levels of investment in IUCN Red List assessments (Stuart et al. 2010). However, large amounts of life-history, ecological and phylogenetic information are available for DD species. The distribution of many DD species is known, allowing inference of species' geographical range size, environmental niche and exposure to anthropogenic threats. These data alone are insufficient for making a decision on formal Red List status, but could be used to help inform global estimates of risk. Comparative studies of extinction risk based on species trait data have previously yielded insight into the determinants of risk across taxa (Purvis 2008; Cardillo & Meijaard 2012), and could enable the preliminary re-assessment of DD species.

Comparative datasets frequently contain many variables, with non-linearities, complex interactions and missing values (Cutler et al. 2007), and as such traditional statistical methods may lack predictive ability. Machine Learning (ML) methods, derived from the artificial intelligence literature, are flexible and powerful tools for finding patterns in datasets (Webb 2002; Hastie et al. 2009). They rely on few assumptions and can utilize large amounts of data, which has made them increasingly popular with ecologists (Prasad et al.

2006; Ozesmi et al. 2006; Cutler et al. 2007; Olden et al. 2008). A wide range of ML algorithms are available, and their relative predictive performance depends on the study objectives and available data (No Free Lunch Theorem: see Webb 2002 and Hastie et al. 2009). A series of comparisons have been made to identify the strengths and weaknesses of different ML algorithms for ecological applications (Elith & Graham 2009; Kampichler et al. 2010; Keller et al. 2011), but only tree-based ML methods have been applied to threatened species classification (Jones et al. 2006; Boyer 2008; Davidson et al. 2009, 2012). The outputs of ML algorithms are probability estimates of a given outcome, which allow easy interpretation of levels of certainty in predicting complex processes such as extinction risk. As a result of these properties, ML algorithms represent a robust approach to identifying the complex pathways leading to observed patterns of extinction risk, and deriving rules-of-thumb to predict the true level of risk of DD species.

Here we investigate the performance of ML algorithms in predicting extinction risk and in estimating the prevalence of risk in DD terrestrial mammals. Terrestrial mammals are a well-suited model taxon for the purposes of our study: they contain a high proportion of species of known conservation status (85%) and previous studies (Purvis et al. 2000; Cardillo et al. 2005, 2008; Davidson et al. 2009) provide a benchmark against which to measure improvement in predictive accuracy. There is also a high amount of data available on the biology of the clade, even for Data Deficient species. We predict extinction risk from data on a range of intrinsic factors, including species' life history and ecology, and extrinsic factors, including environmental data and measures of threat intensity. Specifically, we address the following questions:

1) What are the relative powers of seven different ML methods (classification trees, random forests, boosted trees, k-nearest neighbours, support vector machines, neural networks and decision stumps) to predict extinction risk in terrestrial mammals?
2) How accurately can those methods predict current geographical patterns of extinction risk?
3) Using the models obtained, what is the predicted level of extinction risk faced by DD species?
4) How do our findings change current geographical patterns of extinction risk for terrestrial mammals?

Methods

Dataset

We collated a database for 4,461 terrestrial mammal species with threat status classified as non-threatened (LC, NT), threatened (VU, EN, CR) and Data Deficient (DD) (IUCN 2008). We treated species as threatened or non-threatened, as highly imbalanced categories (2,826 LC species versus 157 CR species) are difficult to discriminate using predictive models (Webb 2002) and uncertainty around classifications with multiple categories is difficult to interpret and communicate. In contrast, machine learning predictions from our binary classification provide a simple quantification of both the likely probability of threatened status for each species and the level of uncertainty around that prediction.

For each species, we collated the following life-history traits (IUCN 2008; Jones et al. 2009): body mass, litter size, habitat breadth, trophic level and number of IUCN-listed habitats. Each trait was available for at least 60% of species. Since some ML methods require complete data, missing data was either phylogenetically imputed (Fritz et al. 2009; Bruggeman et al. 2009), or assigned the genus or family median for species missing from the phylogeny. We used species' range maps to determine geographical range size (IUCN 2010), the latitude of range centroid (IUCN 2010), and extract summary statistics within ranges for a range of global variables: annual mean and seasonality of temperature and precipitation (Hijmans et al. 2005); minimum and range of elevation (Hijmans et al. 2005); mean and minimum human population density for the year 2000 (CIESIN 2005a); and averages for each of Net Primary Productivity (NPP) (Imhoff et al. 2004), Human Footprint (CIESIN 2005b), GDP for the year 1990 (CIESIN 2002) and human appropriation of NPP (Imhoff et al. 2004). Finally, we recorded biogeographical distribution (IUCN 2010), External Threat Index (Cardillo et al. 2004) and habitat suitability (Rondinini et al. 2011) for each species. See Appendix S1 for details.   Previous studies have reached inconsistent conclusions about the primary traits explaining variation in extinction risk across species (Cardillo & Meijaard 2012). In addition, uninformative explanatory variables are unlikely to affect predictive performance in problems with fewer variables than species (Webb 2002; Kuhn 2008). We therefore do not undertake variable selection, but instead focus on using all available traits implicated in determining extinction risk to make the best predictions.

Training of Machine Learning tools

Six ML tools were used to model risk status across all variables: classification trees, random forests, boosted trees, k-nearest neighbours, support vector machines and neural networks. We also computed decision stumps using geographical range size alone to assess the predictive power of that variable. We developed models for all mammals and separately for rodents, bats, primates and carnivores to explore the taxonomic transferability of ML predictive accuracy. ML tools cannot currently take into account phylogenetic relatedness between species, so we included taxonomic order, family and genus in all models to partially account for shared evolutionary history. For each taxonomic dataset, we removed highly correlated (r=0.9) and low variance variables, which can lead to colinearity and zero variance in cross-validation partitions. All numeric predictors were centred and scaled to a standard normal distribution before analysis (Kuhn 2008).

We set aside DD species and, within each taxonomic group, divided the remaining species into a 25% validation set and 75% training set. For each ML method, we used ten-fold cross-validation on the 75% training set to optimize model tuning parameters by maximizing the Area Under the Receiver Operating Characteristic Curve (AUROC), which is insensitive to class imbalance and does not require the specification of misclassification costs (Fawcett 2006). The best ML tool for each dataset for predicting threatened and non-threatened status was then found by comparing AUROC values of various tuned models on the 25% validation set.

In all models, we used Youden's index (Youden 1950) to identify a probability threshold above which species are identified as threatened. This lends equal weight to detecting threatened and non-threatened species, which does not reflect the true prevalence of threat but is reasonable given the importance of identifying threatened species (Vié et al. 2009). All analyses were conducted in R version 2.14.1, using the *caret* package (Kuhn 2008) to optimize model parameters. For further details see Appendix S1.

Spatial analysis of predictions

Using species' range maps (IUCN 2010), we then computed the observed and predicted proportion of threatened species from the 991 species in the 25% validation set across a global grid of 4,505 equal-area hexagons. We fitted a linear regression across cells of observed threat as a function of predicted threat, cell species richness and average range size of species, excluding cells with fewer than 10 species. We also fitted simultaneous autoregressive models to account for spatial autocorrelation (Appendix S1). We produced maps in ArcGIS 9.3 and conducted all analyses in R version 2.14.1.

Predictions for Data Deficient species

We predicted the status of 493 DD species from the best performing global model, using the same threshold as for the validation dataset (Appendix S2) and tabulated the number of DD species predicted to be threatened and non-threatened in 6,593 hexagons. We then compared the proportion of threatened species in cells with and without incorporating our predictions for DD species. Finally, we used linear regression and spatial autoregressive models of observed threat as a function of predicted threat to test for a regression slope different from one.

Results

Comparison of Machine Learning tools and taxonomic levels

Area Under Receiver Operator Characteristic Curve (AUROC) for best models ranged between 0.873 and 0.961 (Table 1), indicating that ML tools calibrated on species-specific information can accurately predict species threat. The best model for the global dataset identified correctly 93.5% of threatened species and 88.7% of non-threatened species

(Appendix S1). There were significant differences in performance across tools (Friedman test, χ^2=18.3, p=0.005, df=6). *Post hoc* symmetry tests showed that this difference was caused by the lack of power of decision stumps based on geographical range size alone, compared to boosted trees (p=0.05, df=1), neural networks (p=0.05, df=1) and support vector machines (p=0.05, df=1). Predictions from the global model for individual orders achieved higher AUROC than predictions from the order-specific models (Appendix S1), indicating that predictions are more reliable when information from all mammals is taken into account.

Spatial predictions

Observed and predicted proportions of threatened species in assemblages of the validation set were broadly consistent (Fig. 1), indicating that ML tools can correctly predict macroecological patterns of extinction risk. In both ordinary least squares (OLS) and spatial regression (SAR) models, we found a strong positive association between predicted assemblage threat on observed assemblage threat (OLS: slope=0.592, p<0.0001, $t_{1,4501}$= 79.03, AIC= -18182; SAR: slope= 0.596, p<0.0001, $t_{1,4499}$=5.457, AIC= -19050). The relationship is mediated by a significant interaction with assemblage species richness in both OLS and SAR models (OLS: slope=0.066, p-value<0.001, $t_{1,4501}$= 3.865; SAR: slope=0.096, p-value<0.0001, $t_{1,4499}$= 5.448), with model fit improving with larger assemblage size (Appendix S1). Mean assemblage risk was globally over-predicted (observed: 36.8%, predicted: 46.7%), mirroring over-predictions at the species level (observed: 22.1%, predicted: 26.7%).

Predictions for Data Deficient species

Our model outputs predict 313 of 493 DD species to be threatened with extinction, implying that underlying risk levels are much greater in DD species (63.5%) than data-sufficient species (22.1%) (Appendix S2). The spatial congruence between threat hotspots identified using only data-sufficient species and hotspots incorporating our DD species predictions was very high (Spearman rank correlation= 0.987, p< 0.001; Fig. 2 and 3). Additionally, the levels of threat in centres of threatened species richness may previously have been underestimated according to our regression model of observed vs. predicted threat (testing for slope≠1: OLS: slope=1.036, p<0.0001, $F_{1,6591}$=242.96; SAR: slope= 1.043, p<0.0001, $\chi^2_{1,6589}$=214.15).

Discussion

We have no appreciation of the true level of extinction risk faced by one in six species on the IUCN Red List. These Data Deficient species are of great conservation concern, as they contribute to considerable uncertainty in estimates of risk (Butchart & Bird 2010; Bland et al. 2012) and are neglected by conservation programmes due to their uncertain status. Accurate predictive models of risk based on species traits could therefore enhance our understanding of risk patterns, and enable the proactive conservation of threatened Data Deficient species.

Predictions for Data Deficient species

We predict 313 of 493 (63.5%) DD species are threatened with extinction (Appendix S2). A previous random forests model (Davidson et al. 2009) predicted only 28 of 341 (8.2%) DD terrestrial mammals to be at risk, perhaps reflecting the low sensitivity of the model to detection of threatened species (sensitivity of 47.7% compared to 93.5% in our best model).

A recently published prediction of species extinction risk using eigenvector methods predicted 35% of 481 DD species to be at risk (Jones & Safi 2011), but the ability of the method to integrate phylogenetic signal has been questioned (Freckleton et al. 2011). Our estimates are considerably larger, increasing the estimated proportion of threatened terrestrial mammals from 22% to 27% globally.

Despite this apparent increase in risk, spatial distribution of predicted risk suggests that global spatial prioritization based on current knowledge is robust to uncertainty. Large model residuals (Fig. 2) were caused by the predicted threatened status of a few wide-ranging DD species, such as the northwestern Australian marsupial mole *Notoryctes caurinus*. Our findings echo those of Joppa *et al.* (Joppa et al. 2011), who found that regions predicted to contain large numbers of undiscovered plant species are already conservation priorities, but show considerably higher levels of species risk than previously acknowledged. Additionally, areas containing DD species have been shown to contain more recently described amphibian species than expected by chance (Brito 2010), suggesting that these sites might hold many undescribed species (Bini et al. 2006). A better understanding of the likely status of DD species may therefore provide an efficient method for targeting surveys, as well as incorporating the world's poorly-known and undescribed species in conservation planning.

Our results suggest that DD species are of great conservation concern. DD species have smaller ranges (median=9,891 km^2) than their data-sufficient counterparts (median= 1,666,107 km^2), which contributes to their high extinction risk. Maps of DD species ranges may be uncertain and underestimated when collection effort is low. Nonetheless, the data suggest that many DD species are likely to be range-restricted and that geographical measures derived from the species' range maps are broadly representative of the species' environment. We make the best use of the information available for each species, and note that risk predictions for individual DD species should be interpreted in the context of their IUCN Red List documentation. Since 2008, two DD mammal species (pale fox *Vulpes pallida* and long-nosed mosaic-tailed rat *Paramelomys levipes*) have been re-assigned as least concern; both were predicted not to be at risk by our model. These cases, along with the high consistency between predicted probability of threat and Red List category in our validation set (Appendix S1), indicate that DD species that are assigned a high probability of threat are likely to be at imminent risk of extinction.

Many Data Deficient mammals are nocturnal, and most are bats and rodents (75%), which are difficult to observe and identify in the field without expert knowledge. Worryingly, nearly 40% of DD species are only known from few specimens, old records or from unknown provenance (Appendix S1), indicating a severe lack of knowledge of mammalian diversity. Predicted threat levels in those very-poorly known species are particularly high (79.6%), compared to species classified as DD due to unknown population trends and threats (51.2%) or uncertain taxonomic status and new discoveries (61.7%). High rate of species rediscoveries indicate that many species missing for long periods of time remain extant (particularly those that are only known from type specimens (Scheffers et al. 2011)), but show considerably higher levels of threat than other species (Scheffers et al. 2011). We may therefore expect very poorly-known DD species to be extant, but on the brink of extinction.

Ninety-one species listed as DD in the 1996 IUCN Red List assessment were assigned to a data-sufficient category in 2008 (Collen et al. 2011), including 31 (34%) as threatened. We predict 53 out of 90 species (59%) listed as DD in both the 1996 and 2008 IUCN Red Lists to be at risk of extinction. This suggests that species already re-assigned to a data-sufficient category are more abundant and widespread than species still listed as DD on the 2008 Red List. Hence, we expect threatened DD species to be the last species to be assigned their true conservation status in future iterations of the Red List. This finding highlights the importance of prioritizing potentially threatened DD species for monitoring and re-assessment. Collection of life-history and distribution information is especially urgent for the 174 DD species excluded from our analysis due to insufficient data.

Comparison of Machine Learning tools and taxonomic levels

For all mammals and within the orders analysed, ML tools achieved very clear discrimination between threatened and non-threatened species in the independent validation sets. Classification trees and k-nearest neighbours are conceptually simpler and computationally less intensive than other tools, and never achieved highest classification performance. Random forests, boosted trees, support vector machines and neural networks performed particularly well, and we recommend them as powerful methods for predicting species extinction risk. Why tools differ in predictive performance depends on the link between the algorithm, fitted functions and data distribution, which can be investigated by simulating data (see Elith & Graham (2009) for an example in species distribution modelling). In addition, studies focusing on explaining the role of underlying risk drivers rather than risk prediction could undertake variable selection and model simplification.

Whether one or all of the recommended methods should be applied to a given situation of extinction risk prediction depends on available computational resources. We believe that even small increases in performance achieved by using multiple techniques justify their combined use, given the importance of accurately predicting species conservation status. Geographical range size alone provided reasonable discriminatory power in decision stumps, as expected from its role in categorising species under IUCN criterion B (Purvis et al. 2000). However, the high AUROC observed in models with all explanatory variables included indicates that these extra data are necessary to identify species not listed under criterion B, and to achieve suitable performance for use in conservation decision-making.

Although comparative studies of extinction risk have been criticized for not providing findings that are applicable across taxa (Cardillo & Meijaard 2012), our results suggest that, at least in mammals, information obtained from a wider range of species improves extinction risk prediction. The additional power provided by including all terrestrial mammal species has important implications for the development of predictive systems for conservation. Transferability of predictive power across taxa, and the trade-off between amount of contextual information and predictive ability should be the focus of future research.

Limitations

Although our models achieved high discrimination between threatened and non-threatened species, a number of factors may have negatively affected predictive performance. Discarding species due to the absence of a range map and setting aside 25% of the data as validation reduced the sample size. Our study also lacked a phylogenetic framework, though we took into account taxonomy in our models by including taxonomic levels (order, family and genus) and building four order-level models. However, order-level models achieved lower predictive performance than order-level predictions from the global model (Appendix S1), indicating a modest role of order-specific processes in determining extinction risk.

Missing and inexact explanatory variables and incomplete characterization of the threatening processes may also have caused misclassifications. For example, Purvis et al. (2000) identified population density as a significant predictor of elevated extinction risk in primates, but were unable to use this variable due to its poor coverage across terrestrial mammals. Analyses based on species' geographic range maps have been criticized as species are not evenly distributed across their range, and because some habitats may be unsuitable or inaccessible for species (Rondinini et al. 2006). Making use of more refined maps of species range, such as those derived from habitat suitability modelling (Rondinini et al. 2011), may shed light on how higher resolution range data inform extinction risk prediction. Anthropogenic threat impacts included in the model were mainly based on properties of the human population in the area, e.g. human population density and gross domestic product. Due to the limited characterization of threatening processes, our models are less likely to identify species threatened by over-exploitation and invasive species than those affected by habitat loss.

Finally, model misclassifications may indicate latent potential for recovery or threat and may be used to inform future species assessments. Three of the 15 species incorrectly classified as non-threatened by our models (*Proechimys roberti*, *Reithrodontomys microdon* and *Scotonycteris ophiodon*) were down-listed to a non-threatened category in 2010.

Conclusions

Data Deficient species should be of high conservation interest: they bias our understanding of patterns of extinction risk (Butchart & Bird 2010; Bland et al. 2012) and are neglected by conservation programmes due to their uncertain status. Resolution of taxonomic uncertainty and extensive field surveys are unlikely prospects for all 10,673 species currently listed as DD on the IUCN Red List, given monetary and time costs of surveys (Balmford & Gaston 1999) and risk assessments (Stuart et al. 2010). Predicting species extinction risk from contextual information could be a rapid and inexpensive approach for prioritizing taxa and geographical regions under limited knowledge. ML methods are extremely powerful tools for statistical pattern recognition, which can readily incorporate decision-makers' risk attitudes and quantify prediction uncertainty. As such, they show great potential for predictive conservation science under increasing availability of biodiversity data. The seven ML tools used across two taxonomic levels of terrestrial mammals accurately predicted species extinction risk and centres of threatened species richness. Data Deficient mammal species are likely to be disproportionately at risk, and unless directly targeted for conservation action may slide towards extinction unnoticed. Although our study leaves global mammalian conservation priorities generally unaffected, we conclude risk levels in terrestrial mammals are likely to have been considerably underestimated. Predicting the conservation status of DD species can reduce uncertainty in global patterns of threat, and enable the transparent prioritization for field surveys of potentially threatened DD species. Such an approach could be particularly cost-effective for taxa containing large numbers of DD species, such as invertebrates (Samways & Böhm 2010). Finally, DD species may be indicative of spatial knowledge deficiency and could inform species inventories. Taking into account information on DD species may therefore help tackle data gaps in biodiversity indicators, as well as conserve the earth's poorly-known biodiversity.

Acknowledgements

We thank J. Bruggeman for his assistance with the PhyloPARS programme. We thank E.J. Milner-Gulland, R. Ewers, G. Mace, E. Nicholson, M. McCarthy, and S. Canessa for thoughtful comments on the manuscript. B.C. is partly supported by a grant from the Rufford Foundation.

Supporting Information

Supplementary methods, tables and figures (Appendix S1) and predicted conservation status of Data Deficient terrestrial mammals (Appendix S2) are available online. The authors are solely responsible for the content and functionality of these materials. Queries (other than absence of the material) should be directed to the corresponding author.

Literature cited

Balmford, A., and K. J. Gaston. 1999. Why biodiversity surveys are good value. Nature **398**:204–205. Macmillan Magazines Ltd.

Bielby, J., A. A. Cunningham, and A. Purvis. 2006. Taxonomic selectivity in amphibians: ignorance, geography or biology? Animal Conservation **9**:135–143. Blackwell Publishing Ltd.

Bini, L. M., J. A. F. Diniz-Filho, T. F. L. V. B. Rangel, R. P. Bastos, and M. P. Pinto. 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. Diversity and Distributions **12**:475–482.

Bland, L. M., B. Collen, C. D. L. Orme, and J. Bielby. 2012. Data uncertainty and the selectivity of extinction risk in freshwater invertebrates. Diversity and Distributions **18**:1211–1220.

Böhm, M. et al. 2013. The conservation status of the world's reptiles. Biological Conservation **157**:372–385.

Boyer, A. G. 2008. Extinction patterns in the avifauna of the Hawaiian islands. Diversity and Distributions **14**:509–517.

Brito, D. 2010. Overcoming the Linnean shortfall: Data deficiency and biological survey priorities. Basic and Applied Ecology **11**:709–713.

Bruggeman, J., J. Heringa, and B. W. Brandt. 2009. PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Research **37**:179–184.

Brummitt, N., S. P. Bachman, and J. Moat. 2008. Applications of the IUCN Red List: towards a global barometer for plant diversity. Endangered Species Research **6**:127–135.

Butchart, S. H. M., and J. P. Bird. 2010. Data Deficient birds on the IUCN Red List: What don't we know and why does it matter? Biological Conservation **143**:239–247.

Butchart, S. H. M., A. J. Stattersfield, J. Baillie, L. A. Bennun, S. N. Stuart, H. R. Akçakaya, C. Hilton-Taylor, and G. M. Mace. 2005. Using Red List Indices to measure progress towards the 2010 target and beyond. Philosophical Transactions of the Royal Society B: Biological Sciences **360**:255–268.

Cardillo, M., G. M. Mace, J. L. Gittleman, K. E. Jones, J. Bielby, and A. Purvis. 2008. The predictability of extinction: biological and external correlates of decline in mammals. Proceedings of the Royal Society B: Biological Sciences **275**:1441–8.

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

Cardillo, M., and E. Meijaard. 2012. Are comparative studies of extinction risk useful for conservation? Trends in ecology & evolution **27**:167–171. Elsevier Ltd.

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

CIESIN. 2002. Country-level Population and Downscaled Projections based on the B2 Scenario (1990). Palisades, NY. Retrieved from http://www.ciesin.columbia.edu/datasets/downscaled.

CIESIN. 2005a. Gridded Population of the World (2000), Version 3 (GPWv3). Socioeconomic Data and Applications Center (SEDAC), Columbia University, Palisades, NY. Retrieved from http://sedac.ciesin.columbia.edu/gpw.

CIESIN. 2005b. Last of the Wild Data Version 2 (LWP-2): Global Human Footprint dataset (HF).

Collen, B., and J. M. Bailie. 2010. The barometer of life: sampling. Science **329**:140.

Collen, B., S. T. Turvey, C. Waterman, H. M. R. Meredith, T. S. Kuhn, J. E. M. Baillie, and N. J. B. Insaac. 2011. Investing in evolutionary history: implementing a phylogenetic approach for mammal conservation. Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2611–2622.

Convention on Biological Diversity. 2010. TARGET 12 - Technical Rationale. COP 10 Decisions Tenth meeting of the Conference of the Parties to the Convention on Biological Diversity. CBD, Nagoya, Japan.

Cumberlidge, N., P. K. L. Ng, D. C. J. Yeo, C. Magalhães, M. R. Campos, F. Alvarez, T. Naruse, S. R. Daniels, L. J. Esser, and F. Y. K. Attipoe. 2009. Freshwater crabs and the biodiversity crisis: importance, threats, status, and conservation challenges. Biological Conservation **142**:1665–1673.

Cutler, R. D., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology **88**:2783–92.

Davidson, A. D., A. G. Boyer, H. Kim, S. Pompa-Mansilla, M. J. Hamilton, D. P. Costa, G. Ceballos, and J. H. Brown. 2012. Drivers and hotspots of extinction risk in marine mammals. Proceedings of the National Academy of Sciences **109**:3395–400.

Davidson, A. D., M. J. Hamilton, A. G. Boyer, J. H. Borwn, and G. Ceballos. 2009. Multiple ecological pathways to extinction in mammals. Proceedings of the National Academy of Sciences **106**:10702–10705.

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

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. Pattern Classification. Page 654 p. Wiley, USA.

Elith, J., and C. H. Graham. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography **32**:66–77.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters **27**:861–874.

Freckleton, R. P., N. Cooper, and W. Jetz. 2011. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. The American Naturalist **178**:E10–7.

Fritz, S. A., O. R. P. Bininda-Emonds, and A. Purvis. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. Ecology letters **12**:538–49.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning. Page 746 p. Springer, NY, USA.

Hijmans, S. E., J. L. Cameron, P. G. Parra, A. Jones, and R. J. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology **25**:1965–1978.

Hilton-Taylor, C., C. M. Pollock, J. S. Chanson, S. H. M. Butchart, T. E. E. Oldfield, and V. Katariya. 2009. State of the world's species. Pages 15–41 Wildlife in a changing world. An analysis of the 2008 IUCN Red List of Threatened Species. IUCN, Gland, Switzerland.

Imhoff, M. L., L. Bounoua, T. Ricketts, C. Loucks, R. Harriss, and W. T. Lawrence. 2004. Global patterns in human consumption of net primary production. Nature **429**:870–873.

IUCN. 2001. IUCN Red List Categories and Criteria: version 3.1. Gland, Switzerland and Cambridge, UK.

IUCN. 2008. 2008 IUCN Red List of Threatened Species. Gland, Switzerland. Retrieved October 10, 2011, from www.iucnredlist.org.

IUCN. 2010. 2010 IUCN Red List of threatened species. Version 2010.3. Retrieved October 10, 2011, from www.iucnredlist.org.

Jones, J. P. G. et al. 2011. The Why, What, and How of Global Biodiversity Indicators Beyond the 2010 Target. Conservation Biology **25**:450–457. Blackwell Publishing Inc.

Jones, K. E. et al. 2009. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. Ecology **90**:2648–2648.

Jones, K. E., and K. Safi. 2011. Ecology and evolution of mammalian biodiversity. Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2451–2461.

Jones, M. J., A. Fielding, and M. Sullivan. 2006. Analysing extinction risk in parrots using decision trees. Biodiversity and Conservation **15**:1993–2007.

Joppa, L. N., D. L. Roberts, N. Myers, and S. L. Pimm. 2011. Biodiversity hotspots house most undiscovered plant species. Proceedings of the National Academy of Sciences **108**:13171–6.

Kampichler, C., R. Wieland, S. Calmé, H. Weissenberger, and S. Arriaga-Weiss. 2010. Classification in conservation biology: A comparison of five machine-learning methods. Ecological Informatics **5**:441–450. Elsevier B.V.

Keller, R. P., D. Kocev, and S. Džeroski. 2011. Trait-based risk assessment for invasive species: high performance across diverse taxonomic groups, geographic ranges and machine learning/statistical tools. Diversity and Distributions **17**:451–461.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of Statistical Software **28**:1–26.

Olden, J. D., J. J. Lawler, and N. L. Poff. 2008. Machine learning methods without tears: a primer for ecologists. The Quarterly review of biology **83**:171–93.

Ozesmi, S., C. Tan, and U. Ozesmi. 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling **195**:83–93.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems **9**:181–199.

Purvis, A. 2008. Phylogenetic approaches to the study of extinction. Annual Review of Ecology, Evolution and Systematics **39**:301–319.

Purvis, A., J. L. Gittleman, G. Cowlishaw, and G. M. Mace. 2000. Predicting extinction risk in declining species. Proceedings of the Royal Society of London Series B: Biological Sciences **267**:1947–1952.

Rondinini, C. et al. 2011. Global habitat suitability models of terrestrial mammals. Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2633–41.

Rondinini, C., K. A. Wilson, L. Boitani, H. Grantham, and H. P. Possingham. 2006. Trade offs of different types of species occurrence data for use in systematic conservation planning. Ecology letters **9**:1136–45.

Samways, M., and M. Böhm. 2010. Invertebrata. Are vertebrates representative of animal biodiversity as a whole? Pages 55–61 in J. E. M. Bailie, J. Griffiths, S. T. Turvey, J. Loh, and B. Collen, editors. Zoological Society of London, London, UK.

Scheffers, B. R., D. L. Yong, J. B. C. Harris, X. Giam, and N. S. Sodhi. 2011. The world's rediscovered species: back from the brink? PloS One **6**:e22531.

Schipper, J. et al. 2008. The status of the world's land and marine mammals: diversity, threat, and knowledge. Science **322**:225–30.

Stuart, S. N., J. S. Chanson, N. A. Cox, B. E. Young, A. S. L. Rodrigues, D. L. Fischman, and R. W. Waller. 2004. Status and trends of amphibian declines and extinctions worldwide. Science **306**:1783–1786.

Stuart, S. N., E. O. Wilson, J. A. McNeely, R. A. Mittermeier, and J. P. Rodríguez. 2010. The barometer of life. Science **328**:177. American Association for the Advancement of Science.

Vié, J.-C., C. Hilton-Taylor, C. M. Pollock, J. S. Ragle, J. Smart, S. S. Stuart, and R. Tong. 2009. The IUCN Red List: a key conservation tool. Pages 1–13 in J.-C. Vié, C. Hilton-Taylor, and S. N. Stuart, editors. Wildlife in a changing world. An analysis of the 2008 IUCN Red List of Threatened Species. IUCN, Gland, Switzerland.

Webb, A. 2002. Statistical Pattern Recognition. Page 496 p. Wiley, Chichester, UK.

Youden, W. J. 1950. An index for rating diagnostic tests. Cancer **3**:32–35.

Tables

Table 1. Number of data-sufficient species, proportion of threatened species, number of Data Deficient species and number of explanatory variables used in the models across datasets.

Dataset	Number of data-sufficient species	Proportion of threatened species	Number of Data Deficient species	Number of explanatory variables
Global	3967	22.1%	493	35
Bats	828	17%	108	36
Carnivores	188	23.2%	14	36
Primates	304	56.7%	12	32
Rodents	1666	17%	263	29

Table 2. Area Under the Receiver Operator Characteristic Curve (AUROC) for each combination of tool and dataset on the validation sets.

	CT	RF	BT	KNN	SVM	NNET	DS
Global	0.895	0.944	0.935	0.906	0.932	0.922	0.75
Bats	0.872	0.894	0.897	0.858	0.871	0.891	0.727
Carnivores	0.896	0.901	0.919	0.849	0.922	0.961	0.736
Primates	0.803	0.854	0.866	0.788	0.873	0.857	0.738
Rodents	0.871	0.951	0.933	0.925	0.949	0.935	0.792

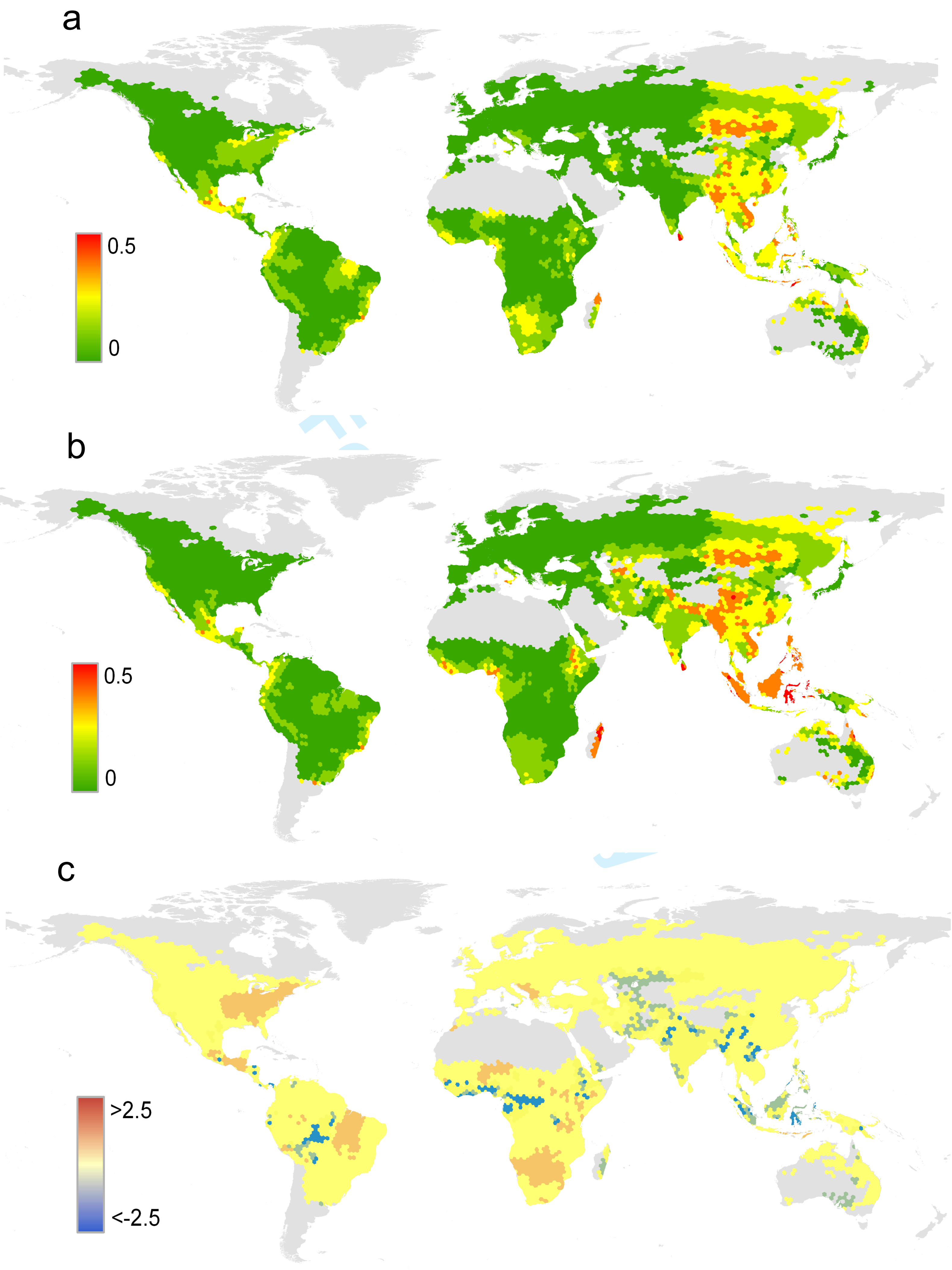
* CT: Classification Tree, RF: Random Forests, BT: Boosted Trees, KNN: K-Nearest Neighbours, SVM: Support Vector Machine, NNET: Neural Networks, DS: Decision Stump.

Figure Legends

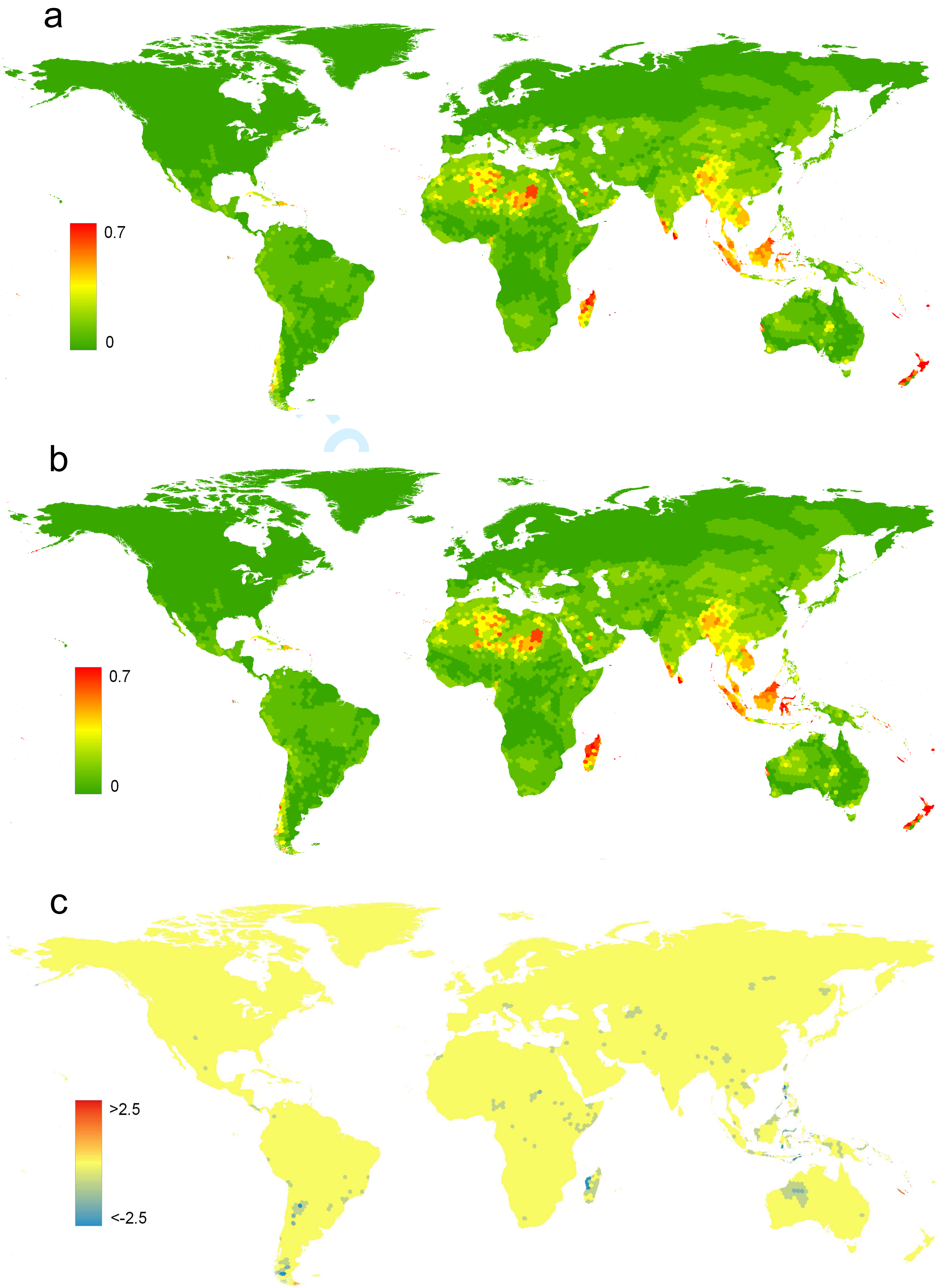
Figure 1. Global geographic distribution of terrestrial mammal extinction risk in the validation set. Observed (a) and predicted (b) proportion of threatened species and standardized model residuals (c).

Figure 2. Global geographic distribution of terrestrial mammal extinction risk. Proportion of threatened species when Data Deficient species are excluded (a), when Data Deficient species model predictions are included (b) and standardized model residuals (c).

Figure 3. Extent of congruence between hotspots of proportion of threatened species under two scenarios, shown across a range of hotspot definitions. The two scenarios are: 1) exclusion of Data Deficient species and 2) inclusion of Data Deficient species model predictions. Horizontal line shows expectation under full congruence; vertical arrow shows 2.5% hotspot definition.



a
0.5
0
b
0.5
0
c
>2.5
<-2.5


a
0.7
0
b
0.7
0
c
>2.5
<-2.5

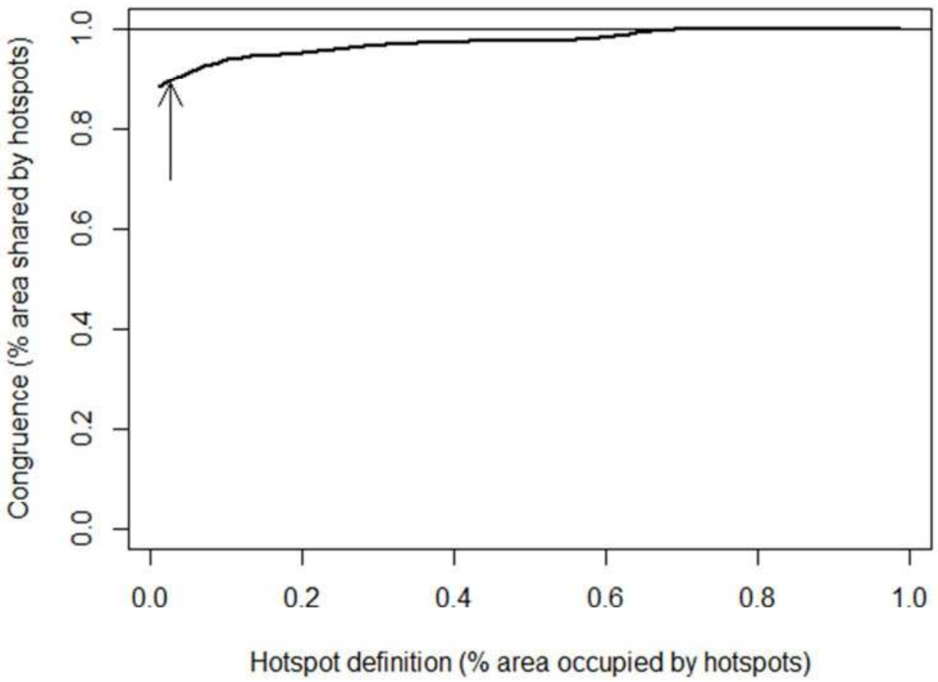


Figure 3. Extent of congruence between hotspots of proportion of threatened species under two scenarios, shown across a range of hotspot definitions. The two scenarios are: 1) exclusion of Data Deficient species and 2) inclusion of Data Deficient species model predictions. Horizontal line shows expectation under full congruence; vertical arrow shows 2.5% hotspot definition.
158x117mm (96 x 96 DPI)