Conservation Biology

Contributed Paper

Predicting the conservation status of data-deficient species

Lucie M. Bland,*† Ben Collen,‡ C. David L. Orme,† and Jon Bielby*

*Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, United Kingdom, email lucie.bland@ioz.ac.uk
†Division of Biology, Imperial College London, Silwood Park, Ascot, SL5 7PY, United Kingdom
‡Centre for Biodiversity and Environment Research, University College London, Gower Street, London, WC1 E6BT, United Kingdom

Abstract: *There is little appreciation of the level of extinction risk faced by one-sixth of the over 65,000 species assessed by the International Union for Conservation of Nature. 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 knowledge gap, we used predictive models incorporating species' life history, geography, and threat information to predict the conservation status of DD terrestrial mammals. We constructed the models with 7 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 centers of threatened species richness. Applying the best model to DD species, we predicted 313 of 493 DD species (64%) to be at risk of extinction, which increases 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 species in these areas show considerably higher levels of risk than previously recognized. We conclude that unless directly targeted for monitoring, species classified as DD are likely to go extinct without notice. Taking into account information on DD species may therefore help alleviate data gaps in biodiversity indicators and conserve poorly known biodiversity.*

Keywords: indicators, mammals, predictive modeling, red lists, threatened species

Predección del Estado de Conservación de Especies con Deficiencia de Datos

Resumen: *Existe poca apreciación del nivel de riesgo de extinción que enfrenta un sexto de las más de 65, 000 especies evaluadas por la Unión Internacional para la Conservación de la Naturaleza. La determinación el estado de estas especies con deficiencia de datos (DD) es esencial para desarrollar una imagen precisa de la biodiversidad global e identificar a las especies con DD potencialmente amenazadas. Para enfocarnos en esta interrupción en el conocimiento, usamos modelos predictivos incorporando la historia de vida de las especies, geografía e información sobre las amenazas para predecir el estado de conservación de mamíferos terrestres con DD. Construimos los modelos con siete herramientas que aprenden de máquinas (ML, en inglés) entrenadas con especies de estado conocido. Los modelos resultantes mostraron una precisión muy alta en la clasificación de especies (hasta 92%) y una habilidad muy alta para identificar correctamente centros de riqueza de especies amenazadas. Al aplicar el mejor modelo a las especies con DD, pronosticamos que 313 de las 493 especies con DD (64%) se encuentran en riesgo de extinción, lo que incrementa la proporción estimada de mamíferos terrestres amenazadas de 22% a 27%. Las regiones que se predijo tendrían un gran número de especies con DD amenazadas ya son prioridades de conservación, pero las especies en estas áreas muestran un nivel de riesgo considerablemente más alto que el que se reconocía previamente. Concluimos que a menos que sean objetivo directo de monitoreo, las especies clasificadas como especies con DD probablemente se extingan sin que nos enteremos. Tomar en cuenta la información sobre las especies con DD por lo tanto puede ayudar a aliviar la interrupción de datos en los indicadores de la biodiversidad y a conservar a la biodiversidad de la que se conoce poco.*

Palabras Clave: especies amenazadas, indicadores, listas rojas, mamíferos, modelado predictivo

Paper submitted July 5, 2013; revised manuscript accepted May 12, 2014.

DOI: 10.1111/cobi.12372

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 poorly known 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 toward 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. 2005).

There has been much improvement in the taxonomic coverage of the IUCN Red List over recent years that has resulted in a more comprehensive understanding of species' extinction risk (Collen & Bailie 2010; Böhm et al. 2013). However, one-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 understanding of patterns of threat and risk (Butchart & Bird 2010; Bland et al. 2012) because 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 not threatened to 36% if all DD species were threatened (Hilton-Taylor et al. 2009). Genuinely threatened DD species may be neglected by conservation programs due to their uncertain conservation status.

Determining the true conservation status of DD species is essential to developing an accurate picture of global biodiversity and enabling the protection of threatened species. Recategorization 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, which allows estimation of species' geographical range size, environmental niche, and exposure to anthropogenic threats. These data alone are insufficient for making a decision on formal IUCN Red List status, but they 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 reassessment of DD species.

Comparative data sets frequently contain many variables with nonlinear relationships, complex interactions, and missing values (Cutler et al. 2007), as such traditional statistical methods may lack predictive ability. Machine learning (ML) methods, derived from the artificial intelligence research, are flexible and powerful tools for finding patterns in data sets (Webb 2002; Olden et al. 2008; Hastie et al. 2009). They rely on few assumptions and can accommodate large amounts of data, which has made them increasingly popular with ecologists (Ozesmi et al. 2006; Prasad et al. 2006; Cutler et al. 2007). A wide range of ML algorithms are available, and their relative predictive performance depends on the study objectives and available data (Webb 2002; Hastie et al. 2009). 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 to deriving rules of thumb to predict the level of risk faced by DD species.

We investigated the performance of ML algorithms in predicting extinction risk and in estimating the prevalence of risk in DD terrestrial mammals. For the purposes of our study, terrestrial mammals are a well-suited model taxon because they contain a high proportion of species of known conservation status (85%) and a previous study (Davidson et al. 2009) provides a benchmark against which to measure improvements in predictive accuracy. In addition, large amounts of species-level data are available for the clade, even for DD species. We predicted 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 addressed the following questions: What are the relative abilities of 7 different ML methods (classification trees, random forests, boosted trees, k nearest neighbors, support vector machines, neural networks, and decision stumps) to predict extinction risk in terrestrial mammals? How accurately can those methods predict current geographical patterns of extinction risk? Using the models obtained, what is the predicted level of extinction risk faced by DD species? How

do our findings change current geographical patterns of extinction risk for terrestrial mammals?

Methods

Data Set

We collated a database for 4461 terrestrial mammal species with threat status classified as nonthreatened (LC, least concern; NT, near threatened), threatened (VU, vulnerable; EN, endangered; CR, critically endangered), and DD (IUCN 2008) (Table 1). For each species, we collated the following life-history traits (IUCN 2008; Jones et al. 2009) (available for at least 40% of species): body mass, litter size, habitat breadth, trophic level, and number of IUCN-listed habitats. Because some ML methods require complete data, missing data were either phylogenetically imputed (Bruggeman et al. 2009; Fritz 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), latitude of range centroid (IUCN 2010), and extract summary statistics within ranges for the following 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 2005*a*); average primary productivity (NPP) (Imhoff et al. 2004), human footprint (CIESIN 2005*b*), and gross domestic product for the year 1990 (CIESIN 2002); and human appropriation of net primary productivity (1976–2000) (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. All geographical variables were 100% complete for each species. See Supporting Information for details on explanatory variables.

We did not undertake variable selection and focused on using all available traits to make the best predictions because previous researchers reached inconsistent conclusions about the traits explaining variation in extinction risk among 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).

Training of ML Tools

Six ML tools were used to model risk status across all variables: classification trees, random forests, boosted trees, k nearest neighbors, support vector machines, and neural networks (Table 2). We also computed decision stumps with geographical range size alone to assess the predictive power of that variable and indicate to what extent range size (IUCN criterion B) approximates IUCN risk classifications. We developed models for all mammals and separate models 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 data set, 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 centered to a mean of zero and scaled to a standard deviation of one 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 to independently assess the performance of different ML methods. For each ML method, we used 10-fold cross-validation on the 75% training set to optimize model tuning parameters by maximizing the area under the receiver operating characteristic curve (AUC), which is insensitive to class imbalance and does not require the specification of misclassification costs (Fawcett 2006). The best ML tool for each data set for predicting threatened and nonthreatened status was then found by comparing AUC values of various tuned models on the 25% validation set. In all models, we identified a probability threshold above which species were identified as threatened by maximizing the Youden index (Y = sensitivity + specificity − 1) (Youden 1950; Perkins & Schisterman 2006). The Youden index effectively lends equal weight to detecting threatened and nonthreatened species whilst accounting for differences in the number of threatened and nonthreatened species. To investigate the role of performance measure on our results, we repeated all analyses by maximizing the H measure, a recently developed alternative to AUC which allows the specification of the distribution of misclassification costs (Hand 2009 but see Flach et al. 2011). Assessing model performance with the H measure (Supporting Information) did not qualitatively affect our results. All analyses were conducted in R version 2.14.1 with the *caret* package (Kuhn 2008) to optimize model parameters. For further details see Supporting Information.

Spatial Analysis of Predictions

We assessed the ability of the best global ML model to predict known patterns of extinction risk. Using species' range maps (IUCN 2010), we computed the observed and predicted proportion of threatened species from the 991 species in the 25% validation set across a global grid of 4505 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

Table 1. Number of data-sufficient species, proportion of threatened species, and number of explanatory variables in our models of extinction risk across data sets.

Data set	*Number of data-sufficient species*	*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. Relative performance[a] of different machine learning methods in predicting conservation status, adapted from Hastie et al. (2009) and Kampichler et al. (2010).

Characteristic	*Trees[b]*	*Neural networks*	*Support vector machines*	*K nearest neighbors*
Handling of multilevel categorical variables	+	−	−	−
Handling of missing values	+	−	−	−
Robustness to outliers in explanatory variables	+	−	−	+
Insensitive to monotone transformations of explanatory variables	+	−	−	−
Ability to extract linear combinations of features	−	+	+	=
Interpretability	+	−	−	−

[a] *Key: +, good; = , fair; −, poor.*
[b] *Trees include decision stumps, classification trees, random forests, and boosted trees.*

10 species (Lee & Jetz 2011). We also fitted simultaneous autoregressive models to account for spatial autocorrelation (Supporting Information). We produced maps in ArcGIS 9.3 and conducted all analyses in R version 2.14.1.

Predictions for DD Species

We predicted the status of 493 DD species from the best-performing global model by using the same threshold as for the validation data set (Supporting Information). We tabulated the number of DD species predicted to be threatened and nonthreatened in 6593 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 ML Tools and Taxonomic Levels

Area under receiver operator characteristic curve (AUC) for best models ranged between 0.873 and 0.961 (Table 3), indicating that ML tools calibrated on species-specific information can accurately predict species threat. The best model for the global data set identified correctly 93.5% of threatened species and 88.7% of nonthreatened species (Supporting Information). There were significant differences in performance across tools (Friedman test, $\chi^2 = 18.3$, $P = 0.005$, 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 relative 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 AUC than predictions from the order-specific models (Supporting Information). Near threatened species had the lowest classification accuracy; 66% of near threatened species were correctly classified as nonthreatened (Supporting Information). Classification accuracy was homogeneous among other red-list categories (87–98%) and among threat types (94–100%). Threatened species with ranges >1 million km^2 were less likely to be correctly classified (87%); conversely, nonthreatened species with very small ranges (<20,000 km^2) were less likely to be correctly classified (74%).

Spatial Predictions

Observed and predicted proportions of threatened species in assemblages of the validation set were broadly consistent (Fig. 2), indicating that ML tools can correctly predict macroecological patterns of extinction risk. In

Table 3. Area under the receiver operator characteristic curve for predictions of extinction risk on the validation sets, shown for each combination of machine learning tool and data set.

	Machine learning tool						
Data set	*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

Abbreviations: CT, classification tree; RF, random forests; BT, boosted trees; KNN, K-nearest neighbors; SVM, support vector machine; NNET, neural networks; DS, decision stump.

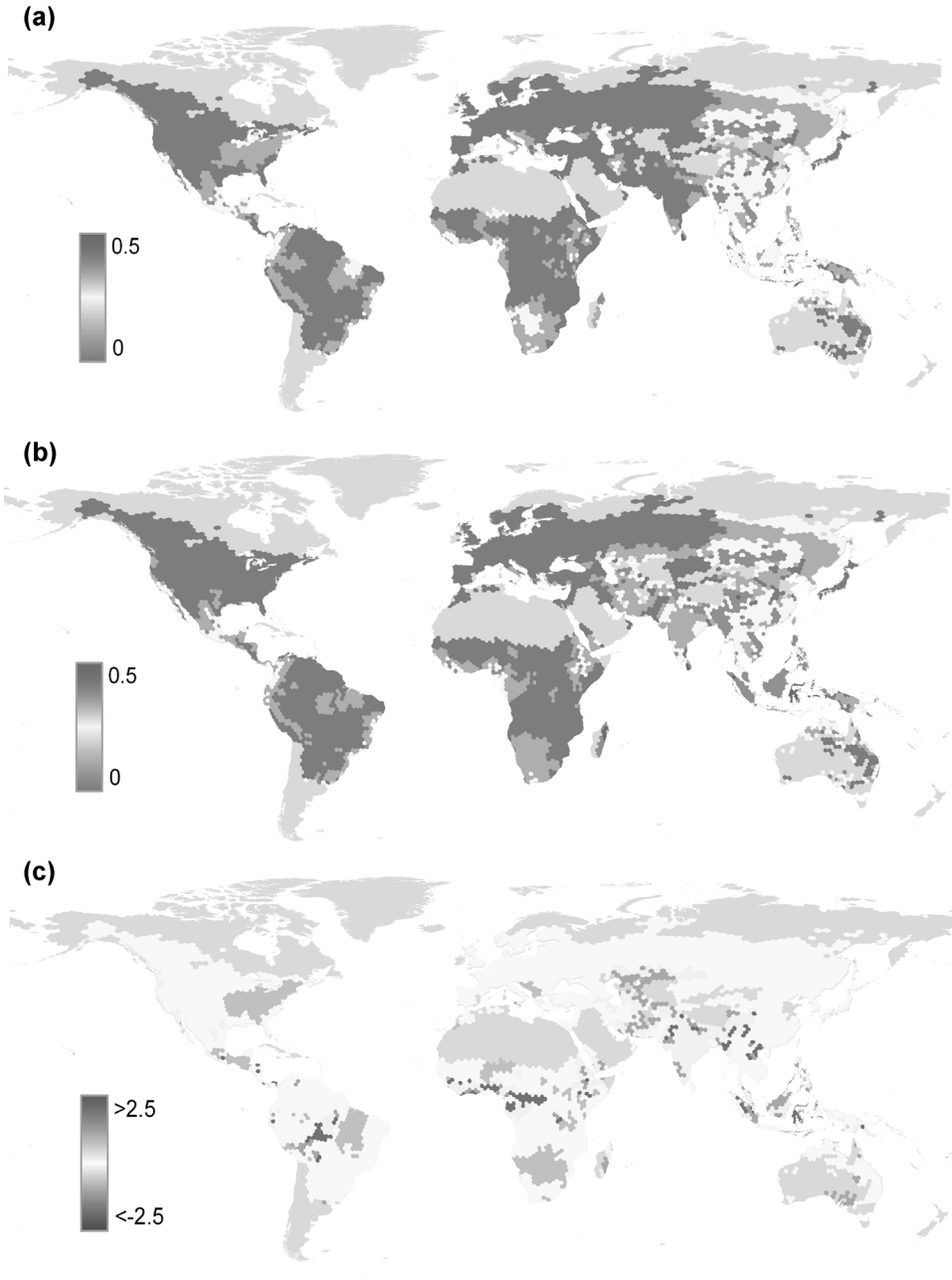

Figure 1. Global distribution of the proportion of threatened terrestrial mammals in the validation set. Proportion (a) observed and (b) predicted from the best machine learning model. (c) Standardized residuals are the difference between the observed and predicted risk levels scaled to SD 1. The validation set contains 991 species.

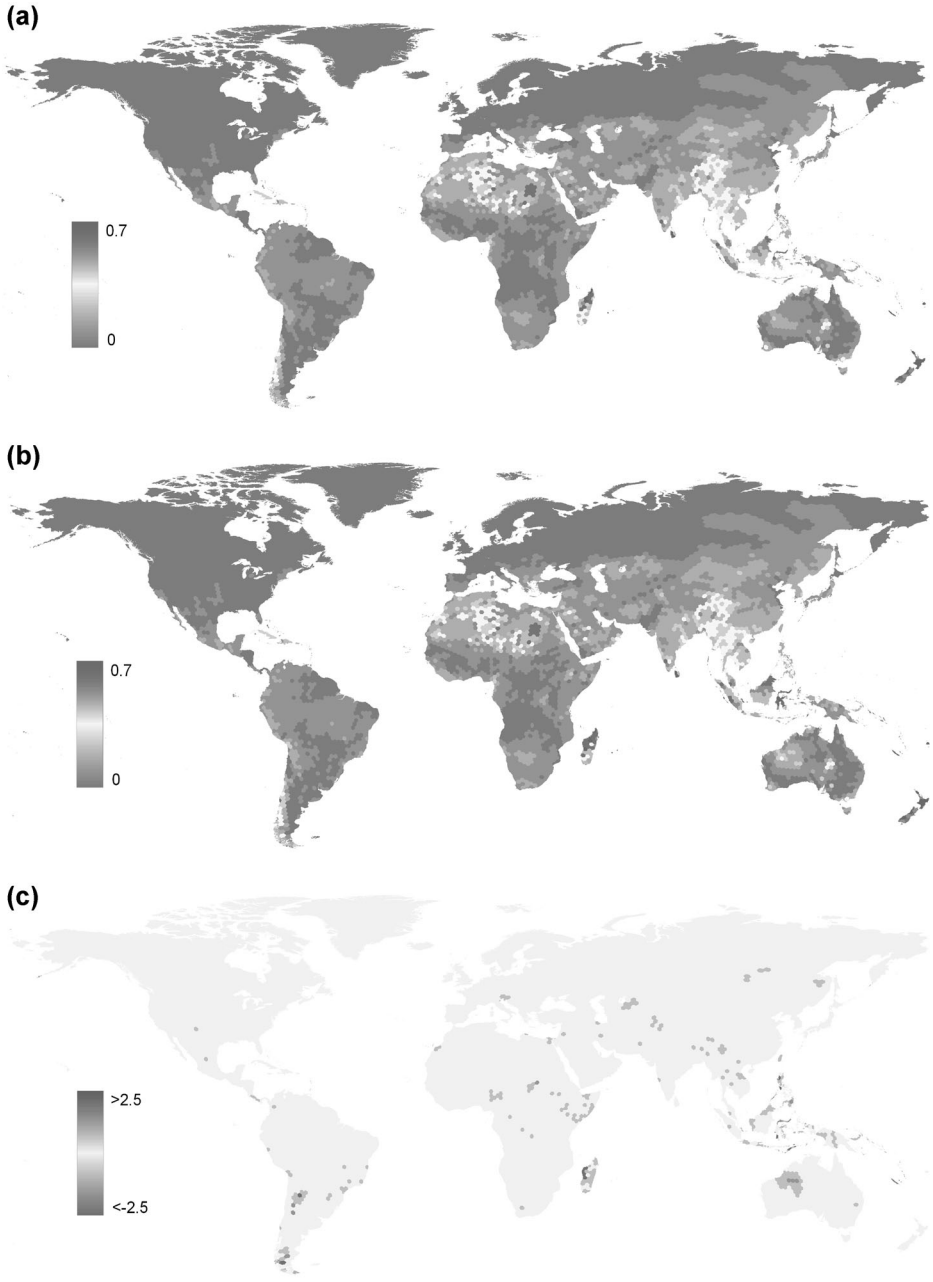

Figure 2. Global distribution of the proportion of all mammal species threatened with extinction. Proportion of threatened species when (a) data-deficient species are excluded from calculations (assumed as equally threatened as data-sufficient species) and (b) data-deficient species model predictions are included. (c) Standardized residuals are the difference between observed and predicted risk levels scaled to SD 1. Distribution based on 4461 species.

both ordinary least squares (OLS) and spatial regression (SAR) models, we found a strong positive relationship 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, $P < 0.001$, $t_{1,4501} = 3.865$ & SAR: slope = 0.096, $P < 0.0001$, = 5.448); model fit improved as assemblage size increased (Supporting Information). Mean assemblage risk was globally overpredicted (observed: 36.8%, predicted: 46.7%), mirroring overpredictions at the species level (observed: 22.1%, predicted: 26.7%).

Predictions for DD Species

Our model outputs showed 313 of 493 DD species were threatened with extinction, implying that underlying risk levels are much greater in DD species (63.5%) than data-sufficient species (22.1%) (Supporting Information). The spatial congruence between threat hotspots identified with only data-sufficient species and hotspots incorporating our DD species predictions was very high (spatial rank correlation = 0.987, $P < 0.001$,

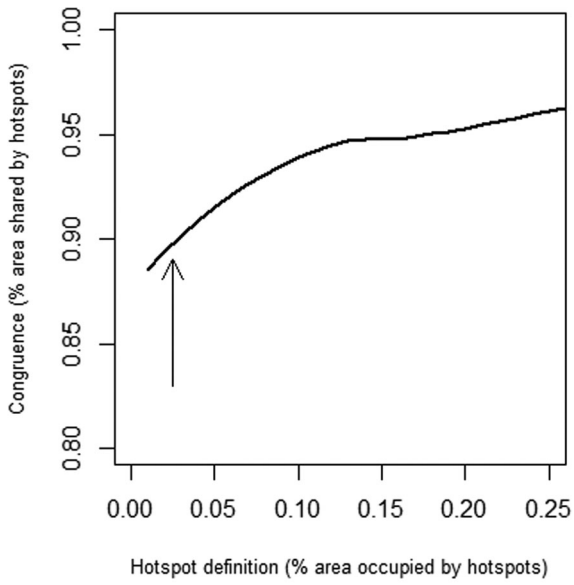

Figure 3. Extent of congruence (y-axis) in hotspots of proportion of threatened species determined with 2 methods relative to area occupied by hotspots. The 2 methods are exclusion of data-deficient species (assuming data-deficient species as equally threatened as data-sufficient species) and inclusion of data-deficient species model predictions. A value of one indicates perfect congruence; the vertical arrow shows the 2.5% threshold for area occupied by hotspots.

Figs. 1 and 3). Additionally, the levels of threat in centers of threatened species richness may previously have been underestimated according to our regression model of observed versus predicted threat (testing for slope≠1: OLS: slope = 1.036, $P < 0.0001$, $F_{1,6591} = 242.96$ & SAR: slope = 1.043, $P < 0.0001$, $x^2_{1,6589} = 214.15$).

Discussion

Predictions for DD Species

There is little appreciation of the true level of extinction risk faced by 1 in 6 species on the IUCN Red List. These DD species are of great conservation concern because they contribute to considerable uncertainty in estimates of risk (Butchart & Bird 2010; Bland et al. 2012) and are neglected by conservation programs because of their uncertain status. Based on our methods, we found that 313 of 493 (63.5%) DD mammal species were threatened with extinction (Supporting Information). A recently published prediction of species extinction risk based on 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). In a previous random forests model, Davidson et al. (2009) predicted 28 of 341 (8.2%) DD terrestrial mammals are at risk, perhaps reflecting the low sensitivity of the model to detect threatened species (sensitivity of 47.7% compared with 93.5% in our best model [Supporting Information]). Our predictions increase the estimated global proportion of threatened terrestrial mammals from 22% (Schipper et al. 2008) to 27%.

Despite this apparent difference in risk, the spatial distribution of predicted risk suggests that global spatial prioritization of mammals based on current knowledge is robust to uncertainty. Our findings echo those of Joppa et al. (2011), who found that regions predicted to contain large numbers of undiscovered plant species are already conservation priorities, but they have 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 and for 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 = 9891 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 show underestimated ranges 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 used the most up-to-date information available for each species, but risk predictions for individual DD species should be interpreted in the context of their IUCN Red List documentation. Since 2008, 2 DD mammal species (pale fox [*Vulpes pallida*] and long-nosed mosaic-tailed rat [*Paramelomys levipes*]) have been reassigned as LC; 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 (Supporting Information), indicate that DD species assigned a high probability of threat are likely to be at imminent risk of extinction.

Nearly 40% of DD species are only known from a few specimens, old records, or records of unknown provenance (Supporting Information), indicating a severe lack of knowledge of mammalian diversity. Predicted threat levels for these very poorly known species are particularly high (79.6%) relative to species classified as DD due to unknown population trends and threats (51.2%) or uncertain taxonomic status and new discoveries (61.7%).

High rates of species rediscoveries indicate that many species missing for long periods remain extant (particularly those that are only known from type specimens), but these rediscovered species 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 were assigned to a data-sufficient category in 2008 (Collen et al. 2011); 31 (34%) were listed as threatened. Our results showed that 53 out of 90 species (59%) listed as DD in both the 1996 and 2008 IUCN Red Lists are at risk of extinction. This suggests that species already reassigned to a data-sufficient category are more abundant and widespread than species still listed as DD on the 2008 IUCN 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 IUCN Red List. This finding highlights the importance of prioritizing potentially threatened DD species for monitoring and reassessment. Collection of life history and distribution information is especially urgent for the 184 DD species excluded from our analysis due to insufficient data (Supporting Information).

Comparison of ML Tools and Taxonomic Levels

For all mammals and within the orders analyzed, ML tools achieved very clear discrimination between threatened and nonthreatened species in the independent validation sets. Classification trees and k nearest neighbors 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. (Elith & Graham [2009] for an example in species distribution modeling.) 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 use in categorizing species under IUCN criterion B (Purvis et al. 2000; Cardillo et al. 2005). Geographical range size alone can however provide misleading information on conservation status: our model was less likely to assign narrow-ranging nonthreatened species and wide-ranging threatened species to their correct status (Supporting Information). The high AUC 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. 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 nonthreatened 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 4 order-level models. However, order-level models achieved lower predictive performance than order-level predictions from the global model (Supporting Information), indicating a modest role of order-specific processes in determining extinction risk. Future studies could focus on efficiently incorporating phylogenies into ML and quantifying the importance of phylogenetic information in predicting extinction risk.

Missing and inexact explanatory variables 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 we were unable to use this variable due to its poor coverage across terrestrial mammals. Analyses based on species' geographic range maps have been criticized because species are not evenly distributed across their range and because some areas 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 modeling (Rondinini et al. 2011), may shed light on how higher resolution range data inform extinction risk prediction.

Finally, model misclassifications may highlight species likely to have been erroneously assessed by the IUCN and

may inform future assessments. Three of the 15 species incorrectly classified as nonthreatened by our models (*Proechimys roberti*, *Reithrodontomys microdon*, and *Scotonycteris ophiodon*) were downlisted to a nonthreatened category in 2010.

Conclusions and Future Prospects

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 7 ML tools we used across 2 taxonomic levels of terrestrial mammals accurately predicted species extinction risk and centers of threatened species richness. DD mammal species are likely to be disproportionately at risk, and unless directly targeted for conservation action may slide toward extinction unnoticed. Although our results leave global mammalian conservation priorities generally unaffected, they suggest risk levels in terrestrial mammals 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 may 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 fill data gaps in biodiversity indicators, as well as conserve Earth's poorly known biodiversity.

Acknowledgments

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 and results (Appendix S1); predicted conservation status of DD terrestrial mammals (Appendix S2); and DD species excluded from our study (Appendix S3) 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.

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

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.

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.

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, 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.

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** DOI: 10.1371/journal.pbio.0020197.

CIESIN. 2002. Country-level population and downscaled projections based on the B2 scenario (1990). Palisades, New York. Available from http://www.ciesin.columbia.edu/datasets/downscaled.

CIESIN. 2005*a*. Gridded population of the world (2000). Version 3 (GPWv3). Socioeconomic Data and Applications Center (SEDAC), Columbia University, Palisades, New York. Available from http://sedac.ciesin.columbia.edu/gpw.

CIESIN. 2005*b*. Last of the Wild Data Version 2 (LWP-2): global human footprint dataset (HF). Available from, from http://sedac.ciesin.columbia.edu/data/collection/wildareas-v2. (accessed October 1, 2011).

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. COP10 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:**2783–2792.

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

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.

Flach, P., J. Hernandez-Orallo, and C. Ferri. 2011. A coherent interpretation of AUC as a measure of aggregated classification performance. Proceedings of the 28th International Conference of Machine Learning. International Machine Learning Society.

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–E17.

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:**538–549.

Hand, D. J. 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning **77:**103–123.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning. 746 p. Springer, NY.

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 in 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.

IUCN. 2008. 2008 IUCN Red List of threatened species. IUCN, Gland, Switzerland. Available from www.iucnredlist.org (accessed October 2011).

IUCN. 2010. 2010 IUCN Red List of threatened species. Version 2010.3. IUCN, Gland, Switzerland. Available from www.iucnredlist.org (accessed October 2011).

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.

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.

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:**13171–13176.

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.

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

Lee, T. M., and W. Jetz. 2011. Unravelling the structure of species extinction risk for predictive conservation science. Proceedings of the Royal Society B: Biological Sciences **278:**1329–1338.

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:**171–193.

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.

Perkins, N. J., and E. F. Schisterman. 2006. The inconsistency of "optimal" cutpoints obtained using two criteria based on the Receiver Operating Characteristic Curuve. American Journal of Epidemiology **163:**670–675.

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 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:**2633–2641.

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:**1136–1145.

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. Evolution lost: status and trends of the world's vertebrates. Zoological Society of London, London.

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**. DOI: 10.1371/journal.pone.0022531

Schipper, J., et al. 2008. The status of the world's land and marine mammals: diversity, threat, and knowledge. Science **322:**225–230.

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.

Webb, A. 2002. Statistical pattern recognition. Wiley, Chichester, United Kingdom.

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