

Resolving the effects of Data Deficient species on the estimation of extinction risk

Lucie Morgane Bland

A thesis submitted for the degree of Doctor of Philosophy from the Division of Ecology and Evolution, Department of Life Sciences, Imperial College London.

June 2014

Abstract

Cost-effective reduction in the uncertainty surrounding global indicators of biodiversity change is a central goal of conservation. In this thesis, I identify and resolve the effects of IUCN Data Deficient species on the estimation of global patterns and levels of extinction risk. I show that gaps in our knowledge of species' conservation status are primarily driven by spatial patterns of ecological research (Chapter 2). Large numbers of species are extremely poorly known, highlighting the importance of basic taxonomic and natural history information in conservation assessments. Using sensitivity analyses (Chapter 3), I show that Data Deficient species contribute to considerable uncertainty in patterns of extinction risk in freshwater invertebrates, limiting our understanding of the factors influencing extinction risks and our capacity to design reliable conservation schemes. To determine the likely conservation status of Data Deficient species, I develop seven machine learning models based on species' life-history traits, niche and threat exposure (Chapter 4). I find that machine learning models accurately predict species conservation status and geographical patterns of threatened species richness. I predict 64% of Data Deficient mammals to be at risk of extinction, increasing the estimated proportion of threatened mammals from 22% to 27% globally. Finally, I use sampling theory to compare the cost-effectiveness of predictive models and IUCN Red List assessments in mammals, amphibians, reptiles and crayfish (Chapter 5). Double sampling with predictive models reduces the cost of determining the proportion of Data Deficient species at risk of extinction by up to 69%, and can be used to reduce the impact of uncertainty in the Red List and Red List Index. My thesis demonstrates how predictive models and decision theory can strengthen indicators of biodiversity change to monitor progress towards international biodiversity targets.

Declaration

All the work presented in this thesis is my own, with the following acknowledgements. The IUCN Red List data used were freely available from the IUCN website (www.iucnredlist.org) or were collected by the Indicators and Assessment Unit at the Institute of Zoology. Species-level life-history data were retrieved or collected from the following sources: mammals: panTHERIA database; amphibians: published databases; reptiles: collaborative efforts among the National Autonomous University of Mexico, Stony Brook University, Nature Serve and the Institute of Zoology coordinated by Andres Garcia, Ana Davidson and Monika Böhm; crayfish: data collection my own. Environmental and threat information were obtained from freely downloadable GIS datasets. All sources are fully acknowledged in the text. The ideas presented in this thesis were formulated in consultation with my supervisors: Ben Collen, Jon Bielby and David Orme. All the statistical analyses and writing are my own, with valuable inputs from my supervisors. Specifically, I acknowledge the following inputs:

Chapter 2: Monika Böhm provided helpful advice on creating species richness maps and accessing IUCN assessments. Chris Carbone and Robin Freeman provided valuable comments on the manuscript. A version of this chapter is submitted to *Global Ecology and Biogeography*.

Chapter 3: I thank Nadia Richman and Helen Meredith for valuable discussions on the ideas presented in the chapter. A version of the chapter is published as: Bland, L. M., Collen, B., Orme, C. D. L. & Bielby, J. 2012 Data uncertainty and the selectivity of extinction risk in freshwater invertebrates. *Diversity & Distributions* **18**, 1211–1220.

Chapter 4: Jörn Bruggeman provided assistance with the PhyloPARS programme. E.J. Milner-Gulland, Rob Ewers, Georgina Mace, Emily Nicholson, Michael McCarthy, and Stefano Canessa provided thoughtful comments on the manuscript. I thank staff and students from the Global Mammal Assessment Unit at Sapienza University in Rome, the University of Melbourne, and the Australian National University for helpful discussions. A version of this chapter is in revision in *Conservation Biology*.

Chapter 5: Ideas for this chapter originated from discussions with Michael McCarthy and Emily Nicholson at the University of Melbourne. Monika Böhm helped me access reptile life-history data. I thank curation staff the Natural History Museum London, Musee d'Histoire Naturelle Paris, Museum Victoria and the Australian Museum for assistance with crayfish data collection. I also thank members from the International Association of Astacology for

discussions on crayfish macroecology, including Keith Crandall, Jesse Breinholt, Jim Fetzner and Jason Coughran. IUCN assessment costs were obtained from personal communications with IUCN Specialist Group chairs and IUCN assessments coordinators. A version of this chapter is submitted to *Ecology Letters*.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acknowledgements

I thank my three supervisors for their continued support, and for encouraging me to develop research skills beyond the requirements of a PhD. I thank Ben Collen for his very regular availability and helping me plan the progress of my PhD. I thank Jon Bielby for helping me clarify my writing and always being available for cups of tea. I thank David Orme for carefully checking through my analyses and providing much-needed help on designing figures. I feel lucky to have worked with three supervisors with whom I got along very well.

I also thank members from my host organizations for advice over the last three years. I thank Monika Böhm for assistance with ArcGIS and IUCN databases; Nadia Richman for providing valuable information and contacts on crayfish; Helen Meredith for shedding light on the wider conservation picture. I thank E.J. Milner-Gulland and Rob Ewers at Imperial College for feedback on my manuscripts during progress review meetings.

I am grateful for the support of friends at the Institute of Zoology during lunch breaks and late nights in the Social Club. Special mentions to: Nadia Richman, Helen Meredith, Monika Böhm, Clare Duncan, Jennifer Crees, Claire Asher, Amy Collins, Kirsten McMillan, Henry Ferguson-Gow and the SRLI interns. I am also grateful to the administrative staff at the Institute (Jo, Amrit and David). I also want to thank Martina DiFonzo for valuable discussions and great experiences in Australia and the USA.

I also acknowledge the support of the Centre for Excellence in Environmental Decisions Australia for funding a five-month visit in 2013, which considerably influenced my perspective on conservation biology. I thank Michael McCarthy and Emily Nicholson for discussing ideas expressed in Chapter 5, and Ayesha Tulloch and Hugh Possingham for discussing ideas for future work on Data Deficient species. I thank staff and students at the University of Melbourne, the Australian National University and the University of Queensland for their friendship. Special mentions to: Kylie Soanes, Emily Hannah, Darren Southwell, Michaela Plein, Stefano Canessa, Skip Woolley, and Payal Bal.

Finally I want to thank my family and friends for putting up with me through the (sometimes) painful moments of my PhD. I am grateful to my parents, for encouraging me to thrive within the educational system over the last 20 years; and my sister, for helpful moaning on the realities of the PhD existence. I thank Martha for providing homely comfort for seven years of house-sharing and biology. I thank all of my great friends for their optimistic attitude and strong livers: Tom Mason, Alex White, Ashley Atkins, Ali Bigos, Sella Mak, and Yallene Thirukkumar.

Table of Contents

Abstract	2
Declaration.....	3
Acknowledgements.....	5
Table of Contents	6
List of Figures.....	8
List of Tables	9
List of Equations.....	10
Chapter 1. Introduction.....	11
Data gaps in conservation knowledge.....	15
Addressing data deficiency in the IUCN Red List	18
Cost-effective assessment of extinction risk with limited information	19
Recommendations for the use of the Data Deficient category by IUCN.....	20
Chapter 2. Known unknowns: global patterns of conservation data deficiency.....	21
Introduction.....	21
Methods.....	24
Data.....	24
Cross-taxa congruence in centres of Data Deficient species richness	25
Geographical correlates of the spatial distribution of data deficiency.....	26
Biological and geographical correlates of Data Deficient species status.....	27
Justification for listing as Data Deficient.....	27
Results.....	28
Cross-taxa congruence in centres of Data Deficient species richness	28
Geographical correlates of the spatial distribution of data deficiency.....	29
Biological and geographical correlates of Data Deficient species status.....	33
Justification for listing as Data Deficient.....	35
Discussion.....	37
Cross-taxa congruence in centres of Data Deficient species richness	37
Geographical correlates of the spatial distribution of data deficiency.....	38
Correlates and justifications of Data Deficient species status.....	39
Limitations and prospects	40
Conclusions.....	41
Chapter 3. Data deficiency and the selectivity of extinction risk in freshwater invertebrates	42
Introduction.....	42
Methods.....	44
Data.....	44
Analyses.....	45
Results.....	46
Taxonomic and geographical selectivity of data deficiency.....	46
Taxonomic and geographical selectivity of extinction risk.....	46
Effect of geographical scale on taxonomic selectivity.....	50
Discussion.....	50
Taxonomic and geographical selectivity of data deficiency.....	51
Taxonomic and geographical selectivity of extinction risk.....	51

Effect of geographical scale on taxonomic selectivity.....	53
Limitations	53
Conclusions.....	54
Chapter 4. Predicting the conservation status of Data Deficient species	56
Introduction.....	56
Methods.....	58
Data.....	58
Training of Machine Learning tools	59
Spatial analysis of predictions	61
Predictions for Data Deficient species.....	61
Results.....	61
Comparison of Machine Learning models and taxonomic levels.....	61
Spatial analysis of predictions	63
Predictions for Data Deficient species.....	65
Discussion.....	68
Predictions for Data Deficient species.....	68
Comparison of Machine Learning models and taxonomic levels	69
Limitations	70
Conclusions.....	71
Chapter 5. Cost-effective assessment of extinction risk with limited information.....	73
Introduction.....	73
Methods.....	75
Double sampling.....	75
Estimating the coefficient of reliability K	77
Estimating the cost ratio R	79
Results.....	81
Estimating the coefficient of reliability K	81
Estimating the cost ratio R	83
Double sampling.....	83
Discussion.....	85
Comparison of Machine Learning models and datasets	87
Limitations	88
Conclusions.....	90
Chapter 6. Conclusions.....	91
Summary of research findings	91
Limitations of the research and future prospects.....	93
Patterns in biodiversity data collection.....	93
Predictive modelling of biodiversity patterns.....	94
Accurately specifying conservation problems and objectives	95
Developing a framework to resolve the effects of data gaps in biodiversity patterns	97
Recommendations to IUCN for the application of the Data Deficient category.....	99
Concluding remarks.....	102
References	103
Appendix I.....	118
Appendix II.....	123
Appendix III	126
Appendix IV	138

List of Figures

Fig 1.1 Structure of the IUCN Red List categories, redrawn from IUCN (2001)	12
Fig 2.1 Uncertainty in estimates of the proportion of threatened species among taxonomic groups	22
Fig 2.2 Richness of terrestrial and freshwater data-sufficient and Data Deficient species in mammals, amphibians, reptiles, freshwater crabs, and crayfish	30
Fig 2.3 Relationship between the prevalence of data deficiency and species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish	32
Fig 2.4 Justifications for listing as Data Deficient in mammals, amphibians, reptiles, freshwater crabs, crayfish, and odonates	36
Fig 4.1 Global distribution of the proportion of threatened terrestrial mammals in the validation set	64
Fig 4.2 Global distribution of the proportion of threatened species for all terrestrial mammals.....	66
Fig 4.3 Extent of congruence in hotspots of proportion of threatened species between two scenarios, show against a range of hotspot definitions	67
Fig 5.1 Proportional reductions in cost and optimal sampling proportion for double sampling assessments of extinction risk.....	84
Fig 6.1 Framework to resolve the effects of data gaps in biodiversity patterns	99

List of Tables

Table 1.1 Case studies of Data Deficient species among taxonomic groups	13
Table 1.2 Correlates of knowledge availability and extinction risk among species	17
Table 2.1 IUCN Red List assessments and available data for mammals, amphibians, reptiles, freshwater crabs, crayfish, and odonates	25
Table 2.2 Matrix of spatial congruence in Data Deficient species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish	28
Table 2.3 Spatial correlates of the prevalence of data deficiency in mammals, amphibians, reptiles, freshwater crabs, and crayfish	31
Table 2.4 Single predictor phylogenetic logistic regression of Data Deficient status in mammals	33
Table 2.5 Multiple predictor phylogenetic logistic regression of Data Deficient status in mammals	34
Table 2.6 Single predictor generalized mixed models of Data Deficient status with nested taxonomic levels in reptiles and crayfish	34
Table 2.7 Multiple predictor generalized mixed model of Data Deficient status in reptiles with nested taxonomic levels	35
Table 3.1 Number of threatened species, Data Deficient species and non-threatened species in crayfish, freshwater crabs and odonates	45
Table 3.2 Global taxonomic selectivity of data deficiency and extinction risk in crayfish, freshwater crabs and odonates	47
Table 3.3 Geographical selectivity of data deficiency and extinction risk in crayfish, freshwater crabs and odonates	48
Table 3.4 Taxonomic selectivity of data deficiency and extinction risk of freshwater crabs within biogeographical realms	49
Table 3.5 Taxonomic selectivity of data deficiency and extinction risk of odonates within biogeographical realms	49
Table 4.1 Characteristics of the datasets used to model extinction risk in mammals	59
Table 4.2 Characteristics of different machine learning methods, adapted from Hastie et al. (2009) and Kampichler et al. (2010)	60
Table 4.3 Area under the receiver operator characteristic curve (AUC) for each combination of tool and dataset on the validation sets	62
Table 4.4 Proportion of species in the validation set correctly identified as threatened or non-threatened by the best machine learning model	63
Table 5.1 Description of IUCN Red List assessments and predictive models of extinction risk for terrestrial mammals, amphibians, reptiles, and crayfish	78
Table 5.2 Model performances among predictive models and taxonomic groups	81
Table 6.1 IUCN classification of Research Needed actions from IUCN (2013b)	101

List of Equations

Equation 1 Variance under single sampling	75
Equation 2 Variance under double sampling	75
Equation 3 Coefficient of reliability	76
Equation 4 Optimal sampling proportion	77
Equation 5 Proportional reduction in cost or variance	77

Chapter 1. Introduction

In 2010 Conservation International launched its *Search for lost frogs*, in an attempt to find a hundred amphibian species not seen in over a decade. Only four of those one hundred species were re-discovered, highlighting both the increasing risk of extinction to amphibian species, and the limited knowledge of their survival status. Limited knowledge of the biological world is a considerable obstacle to the development of reliable and effective conservation measures (The Royal Society 2003; Whittaker *et al.* 2005). Only 1.2 million eukaryotic species have been described out of a putative 8.7 million (Mora *et al.* 2011); we lack geographical distribution data for many species (Lomolino 2004), as well as ecological, behavioural and life-history information (Trimble & Van Aarde 2010; González-Suárez *et al.* 2012).

Documenting species' distributions, population status and ecology is fundamental to evaluating risks to biodiversity (The Royal Society 2003; Whittaker *et al.* 2005; Sousa-Baena *et al.* 2013). As a consequence, limitations in natural history information cause significant data gaps in indicators of biodiversity change adopted by the Convention on Biological Diversity (Balmford *et al.* 2005; Butchart & Bird 2010). To date the conservation status of only 5.8% of the world's described species has been assessed on the IUCN Red List of Threatened Species, and within those, one in six is too poorly-known to assign to an extinction risk category (IUCN 2013a).

The IUCN assigns a species to the Data Deficient category “when there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status” (IUCN 2001). The Data Deficient category therefore does not correspond to a level of extinction risk, but is an assessment of the lack of information on the taxonomy, population status, ecology or threats to a species (Table 1.1). For example, the frog *Hyperolius thoracotuberculatus* was described from an unknown location in Africa in 1931, and cannot be matched to any known species in the wild (IUCN 2013a); the red brocket deer *Mazama americana* is known from more than 1,000 records (Global Biodiversity Information Facility 2013), but its karyotypic pattern and taxonomy remain uncertain (IUCN 2013a); and the Madagascar crayfish *Astacoides madagascariensis* may be threatened by the spread of invasive species, but no estimates of population decline are available (IUCN 2013a). These three species are assessed as Data Deficient, alongside 10,670 other species on the 2013 update of the Red List (IUCN 2013a).

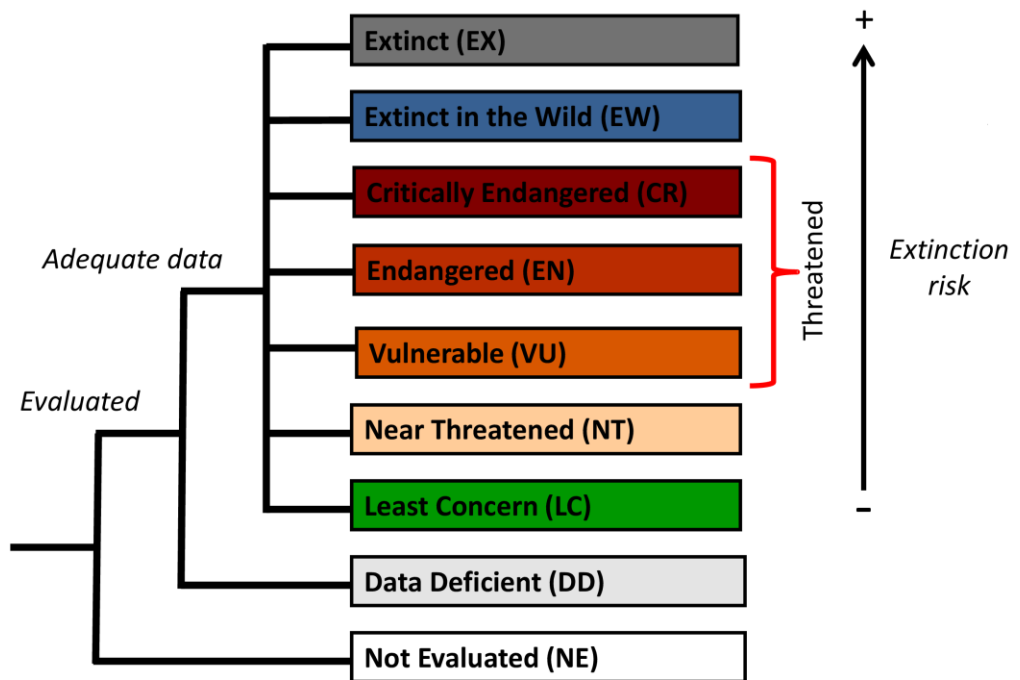




Figure 1.1 Structure of the IUCN Red List categories, redrawn from IUCN (2001).

Table 1.1 Case studies of Data Deficient species among taxonomic groups.

Species	Countries	Assessment justification	Conservation actions	References
<i>Geophis dunni</i> (Dunn's earth snake)	Nicaragua	This species is only known from a single specimen found in the stomach of the coral snake <i>Micrurus nigrocinctus</i> in Nicaragua in 1932.	More research is needed into the distribution, ecology, threats, and habitat status of this species before a full assessment can be made.	IUCN (2013); Schmidt (1932)
<i>Melogale everetti</i> (Bornean ferret badger)	 Borneo	This species is listed as Data Deficient as the impact of potential threats is unknown. Other <i>Melogale</i> species are very adaptable to forest fragmentation and degradation, but the same cannot be assumed for this species. <i>Melogale everetti</i> may be the target of non-specific hunting. Nothing is known about the species' population status or size.	Survey work and research on this species is needed to evaluate its true status and threats. The species was only recorded from Mount Kinabalu National Park in 2002, and surveys are needed to identify additional populations.	IUCN (2013) Photo credit: BBC Natural History Unit via Arkive (2013)
<i>Lepilemur dorsalis</i> (Gray's sportive lemur)	 Madagascar	Listed as Data Deficient as, in light of recent taxonomic upheavals in the genus <i>Lepilemur</i> , the taxonomy, type locality and precise distribution range of this species have become unclear.	A major reassessment of the <i>Lepilemur</i> species currently described in northwestern Madagascar is needed before a full assessment can be made..	IUCN (2013) Photo credit: Pierre Huguet/ Biosphoto via Arkive (2013)
<i>Austrochaperina parkeri</i>	 Papua New Guinea	Listed as Data Deficient since it has only recently been described in 2001. Only one specimen was obtained from the type locality, and the population size is unknown.	Research on its extent of occurrence, status and ecological requirements is needed before a full assessment can be made.	IUCN (2013) Photo credit: Zweifel (2000)
<i>Isomma elouardi</i>	Madagascar	This dragonfly is only known from the type specimens with locality "Madagascar".	Unknown.	IUCN (2013)
<i>Austropotamobius torrentium</i> (Stone crayfish)	 Albania; Austria; Bosnia and Herzegovina; Bulgaria; Croatia; France; Germany; Greece; Hungary; Italy; Montenegro; Romania; Serbia; Slovakia; Slovenia; Switzerland; Turkey	<i>Austropotamobius torrentium</i> is widespread across Europe, but is undergoing significant declines throughout much of its range. However, rates of decline have not been quantified and therefore this species cannot be assessed under criterion A.	Long-term monitoring to assess rate of decline is needed before a full assessment can be made.	IUCN (2013) Photo credit: Daniel Kane via Arkive (2013)

Data Deficient species have typically been excluded from analyses on global patterns of extinction risk (Purvis *et al.* 2000a; Grenyer *et al.* 2006) and conservation prioritization (Carwardine *et al.* 2008; Wilson *et al.* 2011) due to their uncertain conservation status. Moreover, the recommendation to afford Data Deficient species the same level of protection as threatened species (Mace *et al.* 2008) appears to have rarely been followed, given the very large number of Data Deficient present globally. For example, less than 1% of the awards from the People's Trust for Endangered Species (People's Trust for Endangered Species 2013), 3% of the awards from the Mohamed Bin Zayed Species Conservation Fund (Mohamed bin Zayed Species Conservation Fund 2013), and only one project of the World Association of Zoos and Aquaria (World Association of Zoos and Aquaria 2013) exclusively focus on Data Deficient species. Data Deficient species are neglected by global conservation funding programmes, mirroring the plight of other poorly-known species such as species missing for decades (Fisher & Blomberg 2011; Ladle *et al.* 2011; Scheffers *et al.* 2011) and undiscovered species (Bini *et al.* 2006; Giam *et al.* 2012).

Nonetheless, Data Deficient species have received increased interest from the conservation literature in recent years, with studies investigating the rationale for the use of the category (Butchart & Bird 2010; Sousa-Baena *et al.* 2013), the effect of Data Deficient species on conservation priorities (Trindade-Filho *et al.* 2012), their potential for informing future biodiversity inventories (Brito 2010), and potential methods to assess their true conservation status (Good *et al.* 2006; Davidson *et al.* 2009; Jones & Safi 2011; Morais *et al.* 2013). Data Deficient species are of great scientific and conservation interest, as they represent an identifiable gap in knowledge (i.e. a species is either Data Deficient or not), as opposed to study systems relying on relative metrics of inventory completeness (Peterson *et al.* 1998; Reddy & Davalos 2003; Lobo *et al.* 2007). Moreover, the IUCN Red List is considered to be the most comprehensive and authoritative method for assessing extinction risk globally (Hilton-taylor *et al.* 2000; Rodrigues *et al.* 2006; Mace *et al.* 2008), is used to monitor progress towards the Aichi targets of the Convention on Biological Diversity (Convention on Biological Diversity 2010), forms the basis for a range of prioritization schemes (Isaac *et al.* 2007; Carwardine *et al.* 2008), and is embedded in global funding initiatives for conservation (Critical Ecosystems Partnership Fund 2013; Mohamed bin Zayed Species Conservation Fund 2013; People's Trust for Endangered Species 2013). The study of Data Deficient species therefore has substantial implications for our understanding of extinction risk through the IUCN Red List, the design of biodiversity indicators for the Convention on Biological Diversity, and the influence of uncertainty in conservation biology.

In this thesis, I evaluate and address the effects of data gaps in conservation, taking IUCN Data Deficient species as a model. I ask four questions:

- i)* What factors determine the availability of species conservation data?
- ii)* What is the effect of data deficiency on global patterns of extinction risk and conservation prioritization?
- iii)* Can the likely conservation status of Data Deficient species be determined?
- iv)* Can it be determined cost-effectively?

Hereafter, I present the rationales and research methods for investigating each of these questions.

Data gaps in conservation knowledge

A major obstacle to the development of reliable conservation approaches is our limited knowledge of the biological world (The Royal Society 2003; Whittaker *et al.* 2005). Much of the world's species have yet to be formally described (the Linnean shortfall; Brown & Lomolino 1998), and their geographical ranges characterized (the Wallacean shortfall; Lomolino 2004). Our knowledge of biodiversity is not only limited, but biased. The taxonomic (Gaston & May 1992), scientific (Bonnet *et al.* 2002) and conservation (May & Clark 2002; Bajomi *et al.* 2010; Trimble & Van Aarde 2010) literatures show significant biases favouring vertebrates, especially birds and mammals. Biodiversity knowledge is highest in areas that are accessible (Reddy & Davalos 2003; Ficetola *et al.* 2012; Scheffers *et al.* 2012), and close to research infrastructure, such as field stations and universities (Griffiths 2010; Moerman & Estabrook 2006 but see Pautasso & McKinney 2007). Hence, perceived patterns of biodiversity are not only the product of biology, but human observation.

Biased information availability could contribute to considerable uncertainty in patterns of extinction risk. First, bias alters inferences on the level of extinction risk faced by a taxonomic group (González-Suárez *et al.* 2012). If smaller mammals are more likely to be classified as Data Deficient due to low encounter rates, the estimated proportion of mammals faced with extinction is representative of large mammals rather than the taxon as a whole. It therefore follows that our understanding of patterns of extinction risk may also be biased. Second, a strong correlation between factors influencing knowledge availability and those influencing extinction risks may lead to unreliable estimates of risk levels, and may limit our understanding of the factors contributing to high risk (Table 1.2). If small-ranged species are both more likely to be assessed as Data Deficient and more likely to be threatened, levels of extinction risk in a group and the number of species in need of conservation action may have

been severely under-estimated. On the other hand, if small-bodied species are more likely to be assessed as Data Deficient, but are less likely to be affected by anthropogenic processes such as over-exploitation (Owens & Bennett 2000; Fritz *et al.* 2009), levels of extinction risk may have been over-estimated. In addition, species sufficiently known to assign to a threatened category may command more scientific attention, promoting more research and allocation of funds to understanding their conservation problems (Martín-López *et al.* 2011). This positive feedback loop may result in few species being considered conservation priorities (Metrick & Weitzman 1996; Martín-López *et al.* 2011). Considering the range of biological traits and spatial processes contributing to both the scientific study of species (Trimble & Van Aarde 2010), and their endangerment (Purvis *et al.* 2000a, 2000b; Cardillo & Meijaard 2012), disentangling patterns of human observation and true patterns of biodiversity is necessary for reliable understanding of risk and conservation prioritization.

A first step in achieving this goal is characterizing factors that influence conservation knowledge availability on species. Comparative studies of extinction risk have been undertaken widely (Fisher & Owens 2004; Cardillo & Meijaard 2012), and uncovered correlates of risk among taxonomic groups, geographical regions and scales (Purvis *et al.* 2005). Yet, our understanding of the factors influencing knowledge availability at the species level remains poor. In Chapter 2, I characterize global patterns of uncertainty in conservation knowledge among geographical regions and species, focusing on mammals, amphibians, reptiles, freshwater crabs, crayfish and odonates. I also investigate the biological, geographical and anthropogenic factors influencing information availability on species.

Table 1.2 Correlates of knowledge availability and extinction risk among species. Studies selected focused on inter-species variation within a single clade (e.g. mammals, primates), and were of global or regional geographic scale (e.g. Australia). Studies on knowledge availability recorded the description date, number of occurrence records, number of publications or availability of life-history data per species. Studies on extinction risk recorded IUCN Red List status, and did not distinguish among threat types. +: positive effect. - : negative effect. =: no significant effect.

Trait	Knowledge availability	Extinction risk	References on knowledge availability	References on extinction risk
Biology				
Body size	+++++++	+++++++ ==	Blackburn & Gaston 1994; Collen <i>et al.</i> 2004; Diniz-Filho <i>et al.</i> 2005; Gaston <i>et al.</i> 1995; Patterson 1994; Brodie 2009; González-Suárez <i>et al.</i> 2012	+ : Bennett & Owens 1997; Cardillo <i>et al.</i> 2004, 2005; Davidson <i>et al.</i> 2009, 2012; Johnson <i>et al.</i> 2002; Lee & Jetz 2011; Morrow & Pitcher 2003; Purvis <i>et al.</i> 2000a; Sullivan <i>et al.</i> 2000 =: Cooper <i>et al.</i> 2008; Jones <i>et al.</i> 2003
Geographical range size	+++++ ==	-----	+ : Diniz-Filho <i>et al.</i> 2005; Gaston <i>et al.</i> 1995; González-Suárez <i>et al.</i> 2012; Patterson 1994; Trimble & Van Aarde 2010 = : Brodie 2009; Trimble & Van Aarde 2010	Cardillo <i>et al.</i> 2004, 2005, 2008; Cooper <i>et al.</i> 2008; Davidson <i>et al.</i> 2009, 2012; Jones <i>et al.</i> 2003; Lee & Jetz 2011; Purvis <i>et al.</i> 2000a
Habitat specialisation	=	====	González-Suárez <i>et al.</i> 2012	Cooper <i>et al.</i> 2008; Lee & Jetz 2011; Sullivan <i>et al.</i> 2000
Diurnal activity	+	====	Collen <i>et al.</i> 2004	Davidson <i>et al.</i> 2009; Lee & Jetz 2011; Purvis <i>et al.</i> 2000a
Threatened status	++ - =	NA	+ : Trimble & Van Aarde 2010; Martín-López <i>et al.</i> 2011 -: Trimble & Van Aarde 2010 =: Brodie 2009	
Geography				
Tropical distribution/ Latitude	-	==	Collen <i>et al.</i> 2004	Cardillo <i>et al.</i> 2008; Cooper <i>et al.</i> 2008
Human population density	-	+++	Diniz-Filho <i>et al.</i> 2005	Cardillo <i>et al.</i> 2004, 2005, 2008

A second step in assessing the effect of human observation on biodiversity patterns requires quantification of the sensitivity of estimates of extinction risk to the prevalence of data deficiency. Excluding or including Data Deficient species affects the prevalence of threatened species among amphibians families (Bielby *et al.* 2006), and alters the spatial configuration of reserve networks for amphibians in Brazil's Atlantic Forest (Trindade-Filho *et al.* 2012). Data Deficient species are often excluded from calculations of extinction risk levels, disguising considerable uncertainty in estimates of risk. The effect of data deficiency is likely to be particularly strong in taxonomic groups with large numbers of Data Deficient species, such as some recently assessed invertebrate groups (e.g. 49% of freshwater crabs are assessed as Data Deficient; Cumberlidge *et al.* 2009). Moreover, the effect of data deficiency depends on the distribution of both threatened and Data Deficient species among taxonomic levels and geographical regions: if the distribution of Data Deficient species is non-random, treating Data Deficient species as either threatened or non-threatened could dramatically alter observed patterns of extinction risk. In Chapter 3, I assess the sensitivity of taxonomic and geographical patterns of extinction risk to data deficiency, focusing on freshwater crabs, crayfish and odonates.

Addressing data deficiency in the IUCN Red List

To place confidence limits on the proportion of threatened species in a given taxonomic group, the convention is to calculate bounds by treating all Data Deficient species as non-threatened or threatened (e.g. Hoffmann *et al.* 2010). Such an approach may be inadequate if Data Deficient species exhibit traits not consistently associated with a given level of extinction risk (Table 1.2). Integrating life-history and ecological traits with anthropogenic threat information may be necessary to infer the likely extinction risk of Data Deficient species. Although insufficient for formal Red Listing, substantial amounts of life-history and geographical data available for Data Deficient species could inform their extinction risk status. Trait-based comparative studies of extinction risk have typically focused on explaining differences in risk among species, rather than predicting risk in new species (Owens & Bennett 2000; Purvis *et al.* 2000a; Sullivan *et al.* 2000; Adamowicz & Purvis 2006; Jones *et al.* 2006; Cooper *et al.* 2008; Fritz *et al.* 2009; Larson & Olden 2010; Cardillo & Meijaard, 2012). Attempts to predict the extinction risk of Data Deficient species have suffered from methodological flaws, such as use of controversial methods (e.g. eigenvector method in Jones & Safi 2011; Safi & Pettorelli 2010, see criticism by Freckleton *et al.* 2011), arbitrary classification criteria (Morais *et al.* 2013; Sousa-Baena *et al.* 2013), or poor classification performance (Davidson *et al.* 2009). Finally, most studies have failed to account

for the justification provided by IUCN for listing a species as Data Deficient, and the uncertain nature of the information available on these species. Unravelling the conservation status of Data Deficient species therefore requires the design of a strong predictive framework, resilient to missing and uncertain data, and capable of communicating uncertainty in predictions. In addition, such a framework should be transferable among a wide range of taxonomic groups.

In Chapter 4, I use Machine Learning tools to predict the extinction risk of Data Deficient terrestrial mammals. Machine Learning tools are powerful methods for finding patterns in large datasets, and are increasingly applied to ecological and conservation problems (Lek & Gue 1999; De'ath & Fabricius 2000; Drake *et al.* 2006; Ozesmi *et al.* 2006; Prasad *et al.* 2006; Cutler *et al.* 2007; De'ath 2007; Elith *et al.* 2008; Olden *et al.* 2008; Kampichler *et al.* 2010). I assess the ability of seven Machine Learning tools to predict the extinction risk of species of known conservation status, and predict centres of threatened species richness using terrestrial mammals as a study taxon. I then use the best model to predict the likely status of Data Deficient species. In Chapter 5, I extend the method to an existing dataset on amphibians (Bielby *et al.* 2008; Cooper *et al.* 2008), and two new global datasets on reptiles and crayfish.

Cost-effective assessment of extinction risk with limited information

Chapter 5 investigates the potential for predictive models to cost-effectively determine the status of Data Deficient species. Collecting information and updating Red List assessments for the 10,673 species currently listed as Data Deficient will require considerable resources, given the costs of biodiversity surveys (Balmford & Gaston 1999) and Red List assessments (Stuart *et al.* 2010). Biodiversity monitoring should be undertaken in the most cost-efficient manner to inform conservation decisions (Mace & Baillie 2007; McDonald-Madden *et al.* 2010; Jones *et al.* 2011); this is particularly the case for data collection for global biodiversity indicators, which synthesize large amounts of information at high running costs (Jones *et al.* 2011). Indeed, the Red List is already under pressure of expanding the coverage of biodiversity assessments (Collen *et al.* 2009; Stuart *et al.* 2010), whilst keeping those up-to-date (Rondinini *et al.* 2013). Designing a cost-effective strategy for the reduction of data gaps will therefore ensure the Red List meets its conservation objective “to provide information and analyses on the status, trends and threats to species in order to inform and catalyse action for biodiversity conservation” (IUCN 2013c).

Predictive models such as those developed in Chapter 4 could be used to cheaply and accurately estimate risk levels among groups, in order to monitor progress towards the Aichi targets of the Convention on Biological Diversity (Convention on Biological Diversity 2010). In Chapter 5, I determine the proportion of Data Deficient at risk of extinction with double sampling theory. Double sampling theory is frequently used in medicine to compare diagnostic tests differing in cost and reliability (Baker 1991; Zhou *et al.* 2002). I apply the method to compare the cost-effectiveness of predictive models of risk and IUCN Red List assessments in mammals, amphibians, reptiles and crayfish. I take into account multiple scenarios of information quality and availability on Data Deficient species, as well as uncertainty in field survey costs among species.

Recommendations for the use of the Data Deficient category by IUCN

The IUCN “discourages the liberal use of the Data Deficient category”, as “taxa that are poorly-known can often be assigned a threat category on the basis of background information concerning the deterioration of the habitat and/or other causal factors” (IUCN 2001). Yet, in the absence of formal thresholds, risk attitudes of individual assessors may cause discrepancies in the application of the category. An evidentiary attitude results in a higher number of species listed as Data Deficient, and increased risk of neglecting species in need of urgent conservation action, whilst, a precautionary attitude may generate inaccurate classifications of extinction risk. As a consequence, Butchart & Bird (2010) hypothesized the Data Deficient category to be the most misunderstood and controversial category on the Red List, and the most heterogeneous among assessments of different taxonomic groups. In Chapter 6, I provide recommendations for the use and consistent reporting of the category to inform future conservation actions directed towards Data Deficient species.

Chapter 2. Known unknowns: global patterns of conservation data deficiency

A version of this chapter is submitted to *Global Ecology and Biogeography*.

Introduction

Limited knowledge of the biological world is a major obstacle to the development of reliable conservation approaches (The Royal Society 2003; Whittaker *et al.* 2005). Only 1.2 million eukaryotic species have been described out of a putative 8.7 million (Mora *et al.* 2011), and rates of species discoveries show little sign of abatement even in well-known groups (Köhler *et al.* 2005; Ceballos & Ehrlich 2009). Among described species, natural history and geographic information is strongly biased towards terrestrial plants and vertebrates (Bonnet *et al.* 2002; May & Clark 2002; Millenium Ecosystem Assessment 2005) and towards temperate rather than tropical regions (Collen *et al.* 2008). Documenting species' distributions, population status and natural history is fundamental to evaluating risks to biodiversity (The Royal Society 2003; Whittaker *et al.* 2005). As a consequence, limitations and biases in the availability of biological data may hinder our ability to monitor trends in biodiversity loss and develop sound conservation schemes.

The IUCN Red List of Threatened Species is a key conservation tool used to monitor progress towards the Aichi targets of the Convention on Biological Diversity (Convention on Biological Diversity 2010) and develop a range of prioritization schemes globally (Isaac *et al.* 2007; Carwardine *et al.* 2008). However, the IUCN Red List suffers from significant data gaps, as one in six assessed species are classified as Data Deficient due to lack of knowledge on their taxonomy, population status and threats (IUCN 2013a). The proportion of species assessed as Data Deficient varies widely among groups (from 1% in birds to 49% in freshwater crabs; Butchart & Bird 2010; Cumberlidge *et al.* 2009), and is particularly high in recently assessed invertebrate groups (Samways & Böhm 2010). Data Deficient species are excluded from calculations of extinction risk levels since there is no realistic appreciation of their relative level of risk, but this approach disguises considerable uncertainty (Figure 2.1; Bland *et al.* 2012). For example, 25% of data-sufficient (species classified in non-Data Deficient categories) mammals are threatened with extinction, but estimates range from 21%

if all Data Deficient species were non-threatened to 36% if all Data Deficient species were threatened (Schipper *et al.* 2008). If Data Deficient species are non-randomly distributed among taxonomic levels and geographical regions, treating all Data Deficient species as either threatened or non-threatened dramatically alters observed patterns of extinction risk within a group (Bland *et al.* 2012). Extinction risk patterns and prioritization schemes based on Red List data may therefore exhibit substantial uncertainty associated with Data Deficient species.

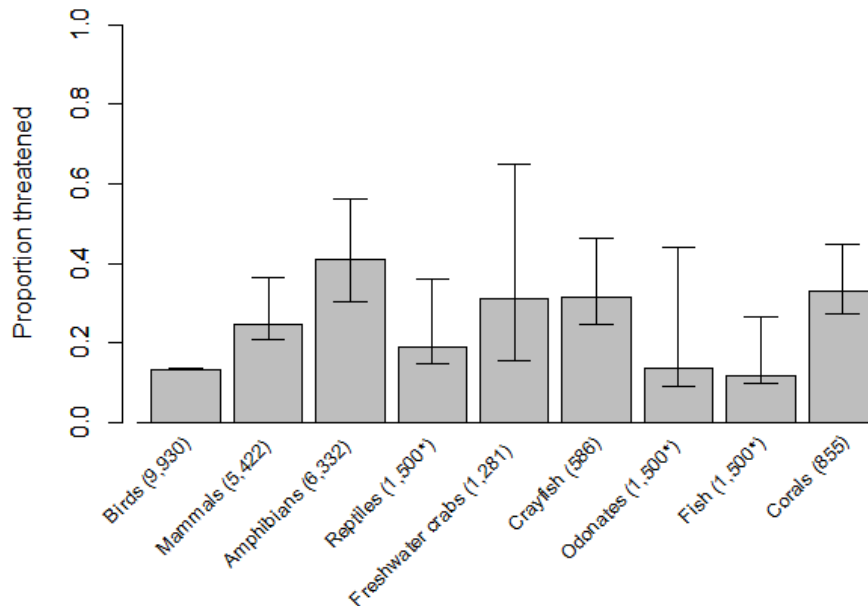


Figure 2.1 Uncertainty in estimates of the proportion of threatened species among taxonomic groups. Histogram bars indicate the proportion of threatened species if Data Deficient species are excluded, or assumed to be threatened in equal proportions to data-sufficient species. Upper error bars indicate the proportion if all Data Deficient species are considered threatened. Lower error bars indicate the proportion if all Data Deficient species are considered non-threatened. Numbers in brackets indicate the number of species in the IUCN Red List group assessment. *: groups assessed with the Sampled Red List approach.

Characterizing the geographic distribution of conservation data deficiency is a first step in assessing the reliability of current biodiversity patterns, and prioritizing areas for conservation research. Most of our understanding of the representativeness of biodiversity data stems from studies of restricted geographic scope (Lobo *et al.* 2007; Hortal 2008; Trimble & Van Aarde 2010; Vale & Jenkins 2012), or from broad comparisons between vertebrates and invertebrates (May & Clark 2002). These approaches are insufficient for informing international biodiversity targets and prioritization schemes, which typically focus on differences within and among taxonomic groups globally (Brooks *et al.* 2006; Wilson *et al.* 2006; Rondinini *et al.* 2011a). Investigating the use of the Data Deficient category therefore provides an opportunity to systematically compare data gaps among taxonomic groups and geographical areas at multiple spatial scales.

Congruent centres of data deficiency among groups may reflect similar driving processes, such as low sampling intensity in inaccessible and sparsely populated areas. Knowledge of biodiversity is highest in areas that are accessible (Reddy & Davalos 2003; Ficetola *et al.* 2013), and close to research infrastructure, such as field stations and universities (Moerman & Estabrook 2006; Griffiths 2010 but see Pautasso & McKinney 2007), whilst knowledge is particularly scarce in equatorial and species-rich regions (Collen *et al.* 2008). Complex interactions among geographical location, sampling intensity and species diversity are therefore likely to shape patterns of biodiversity knowledge. In addition, congruent hotspots of data deficiency would allow research actions directed towards poorly known species to be shared among groups. Re-assessing all 10,673 Data Deficient species currently on the Red List (IUCN 2013a) to data-sufficient categories will require considerable financial resources, given the high costs of field surveys (Gardner *et al.* 2008) and Red List assessments (Stuart *et al.* 2010). Assessing the taxonomic transferability of conservation research actions is therefore imperative to cost-effectively reducing uncertainty in the IUCN Red List.

Biased distribution of data deficiency in respect to species traits could contribute to considerable uncertainty in our understanding of extinction risks. Small species can be less apparent to biologists, and tend to be discovered later (Blackburn & Gaston 1994; Patterson 1994; Diniz-Filho *et al.* 2005), whilst small-ranged species may be encountered less frequently (Patterson 1994; Gaston *et al.* 1995; Diniz-Filho *et al.* 2005). If small-ranged species are both more likely to be assessed as Data Deficient and more likely to be threatened, levels of extinction risk in a group may have been severely under-estimated. Differences in data deficiency may also result from taxon-specific trends in the study of species and the application of the Data Deficient category. Butchart and Bird (2010) considered the Data Deficient category to be the most misunderstood and controversial category on the Red List,

due to the absence of formal thresholds for its application. Investigating assessment rationales for the classification of species into the Data Deficient category is crucial to quantifying knowledge deficiency, ensuring consistent application of the category, and prioritizing Data Deficient species for further research.

Disentangling true patterns of extinction risk from biases in human observation is necessary for reliable conservation prioritization, given the range of biological traits and spatial processes contributing to both the scientific study of species (Trimble & Van Aarde 2010), and their endangerment (Purvis *et al.* 2000a, 2000b; Cardillo & Meijaard 2012). In this chapter, I investigate global patterns of conservation data deficiency in six groups of non-marine species: mammals, amphibians, reptiles, freshwater crabs, crayfish and odonates. First, I assess the congruence of geographical patterns of species classified as Data Deficient among groups, to inform conservation research directed towards poorly known species. I then assess the relative roles of species biology and human sampling effort in driving patterns of data deficiency, both at the geographical assemblage (grid cell) level and at the species level. I use two proxies of global sampling intensity: human population density (CIESIN 2005a), and a recently developed measure of accessibility quantified as the travel time from the nearest city with land or water-based transport (Nelson 2008). Finally, I review IUCN Red List assessment rationales for the classification of species in the Data Deficient category, provide recommendations for the use of the category and outline avenues for research on poorly known species.

Methods

Data

I obtained complete IUCN group assessments for mammals (Schipper *et al.* 2008), amphibians (Stuart *et al.* 2004), freshwater crabs (Cumberlidge *et al.* 2009), and crayfish (IUCN 2010)(Table 2.1). I obtained randomly selected, representative global samples of 1,500 reptiles (Böhm *et al.* 2013) and 1,500 odonates (Clausnitzer *et al.* 2009) following the Sampled Red List approach (Baillie *et al.* 2008). I identified Data Deficient (DD) and data-sufficient (LC, NT, VU, EN and CR) species in each group. I excluded species classified as EX and EW from the analyses. I gathered IUCN geographical range maps for all groups except odonates, since there are no maps available for that group (Clausnitzer *et al.* 2009). The availability of global species-level trait datasets is limited, hence I focussed my analyses on body size, number of IUCN-listed habitats, and geographical range size species data for mammals (Jones *et al.* 2009), reptiles (Appendix IV), and crayfish (Appendix IV). Body size

was measured as median body mass (g) in mammals, maximum snout-vent length (mm) in reptiles, and maximum carapace length (mm) in crayfish. All analyses were conducted in ArcGIS 9.3 and R version 2.12.0 (R Development Core Team 2010).

Table 2.1 IUCN Red List assessments and available data for mammals, amphibians, reptiles, freshwater crabs, crayfish, and odonates. Data-sufficient species are listed as Least Concern, Near Threatened, Vulnerable, Endangered or Critically Endangered on the Red List. Extinct and Extinct in the Wild species are excluded from calculations. *: groups assessed with the Sampled Red List approach.

	Number of assessed species	Percentage of species classified as Data Deficient	Percentage of threatened data-sufficient species	Number of mapped species	Number of species with trait data
Mammals	5,282	12.8	24.5	5,275	4,997
Amphibians	6,260	25.4	41	5,958	NA
Reptiles*	1,500	21.8	18.9	1,467	1,416
Freshwater crabs	1,281	49.3	31.1	1,279	NA
Crayfish	586	21.1	31.3	579	576
Odonates*	1,500	35.1	13.9	NA	NA

Cross-taxa congruence in centres of Data Deficient species richness

I assessed the spatial congruence in patterns of conservation data deficiency among groups by generating spatial overlays of Data Deficient species richness. I overlaid Data Deficient species ranges with an equal-area grid of 21,583 hexagons of 23,529 km². The grain was selected to obtain a reasonable number of Data Deficient species in each cell for congruence and spatial regression analyses (the maximum number of Data Deficient species in a cell ranged from 11 to 35 among groups). Cells not containing any species may inflate covariation measures (the double zero problem: Legendre & Legendre 1998), so I excluded those from my analysis. Following studies of similar species richness patterns (Grenyer *et al.* 2006; Collen *et al.* 2014), I identified the 5% of cells richest in Data Deficient species and calculated the spatial congruence of data deficiency among groups. I examined the sensitivity of this value by repeating the analysis with the richest 2.5% and 10% of cells; this did not qualitatively affect the results (Tables S2.1 and S2.2).

Geographical correlates of the spatial distribution of data deficiency

I assessed the relative roles of species diversity and sampling effort on the spatial distribution of data deficiency in each group, using total species richness and two global proxies of sampling effort: human population density (people/km²; CIESIN 2005a) and remoteness (travel time in hours to the nearest city >50,000 people; Nelson 2008). For each cell in the aforementioned hexagonal grid, I extracted: Data Deficient species richness, total species richness, mean human population density, and mean remoteness. For each group, I removed cells that did not contain any species. I modelled the prevalence of data deficiency in each cell by taking a log transformation of the incidence proportion, to achieve equal variance and normality of residuals in regression models. I computed $Z_i = \log \frac{1000(Y_i+1)}{n_i}$ (Waller & Gotway 2004), where Y_i is the Data Deficient species richness and n_i the total species richness. I log transformed species richness, human population density and remoteness. I included in the models main and quadratic forms of all variables and the interactions of species richness with human population density and remoteness.

I first modelled the prevalence of data deficiency with ordinary least square (OLS) regression. I devised minimum adequate models by stepwise model simplification of the full model. I removed the term with the highest p-value until all terms were significant, using a p-value for significance of 0.01 given the number of hypotheses tested among groups. Moran's I tests showed significant spatial autocorrelation in all models (mammals: Moran's I = 1.29, n = 7,544, p<0.0001, amphibians: Moran's I = 6.9, n = 6,025, p<0.0001; reptiles: Moran's I = 4.98, n = 6,237, p<0.0001; freshwater crabs: Moran's I = 1.84, n = 3,282, p<0.0001; crayfish: Moran's I = 3.16, n = 1,863, p<0.0001). I integrated spatial autocorrelation in the models using simultaneous autoregressions (SAR). I defined neighbourhood size as the distance at which OLS residuals were no longer autocorrelated (Cressie 1993): 350 km in mammals, 200 km in amphibians, 550 km in reptiles, 350 km in freshwater crabs, and 200 km in crayfish. I calculated neighbourhood connections matrices with row-standardised weights. I considered two specifications of the error covariance matrix: spatial lag (spatial autocorrelation in the response), and spatial error (spatial autocorrelation in the error term). I used a Lagrange multiplier test (Anselin 1988) to find the best error specification; in all groups, the spatial error model showed higher support. I then undertook stepwise model selection as described for OLS models. SAR models showed lower residual spatial autocorrelation than OLS models (Figure S2.1). All models were built with the *spdep* package in R (Bivand *et al.* 2014); detailed model-fitting procedures are available in Tables S2.3 – S2.5.

Biological and geographical correlates of Data Deficient species status

I assessed the relative roles of biological traits and geographical proxies of sampling effort in determining whether a species was assessed as Data Deficient or data-sufficient for mammals, reptiles and crayfish. Associations between Data Deficient status and five predictor variables were investigated: species' body size, geographical range size, habitat specialization, and two proxies of sampling effort: mean human population density in the species' range (CIESIN 2005a), and mean remoteness of the species' range (Nelson 2008). All variables were log transformed.

In mammals, I quantified the strength of phylogenetic signal in Data Deficient status with the D statistic for binary data (Fritz & Purvis 2010; Orme *et al.* 2012) using 1,000 permutations for 4,461 species present in a global phylogeny (Fritz *et al.* 2009). In the presence of phylogenetic signal ($D= 0.876$, $p(D>0)<0.001$, $p(D<1)<0.001$), I used Ives & Garland's (2010) phylogenetic logistic regression for binary dependent variables to investigate correlates of data deficiency (Ho & Ane 2013). First, I regressed each variable as a single predictor of data deficiency. I then used multiple regressions to investigate shared information content among variables, including variable main and quadratic forms, and first order interactions with geographical range size. I ran two sets of multiple regressions with and without body size, as inclusion of body size severely reduced sample size which could in turn affect inference. I devised minimum adequate models by stepwise model simplification, removing the term with the highest p-value (deviance and AIC values are not available in phylogenetic logistic regression) until all terms were significant ($p<0.01$).

Global phylogenies for reptiles and crayfish are not available, so I created generalized linear mixed models (GLMM) with binomial error and taxonomic information (order, family, genus) as nested random factors (Pandit *et al.* 2011). First, I regressed each variable as a single predictor, and then regressed variables together in a multiple regression. I computed a maximal model including variable main effects and first order interactions with geographical range size, and conducted model simplification as described for mammals. I also ran GLMMs for mammals based on taxonomic information to investigate differences with phylogenetic logistic regression (Tables S2.6 and S2.7). I calculated marginal and conditional R^2 in GLMMs following Nakagawa & Schielzeth (2013).

Justification for listing as Data Deficient

I used the assessment rationales recorded on the IUCN Red List to assign species to eight justifications of Data Deficient status: new species, taxonomic uncertainty, type series, few

records, old records, unknown record provenance, unknown population status or distribution, and unknown threats. I defined “new species” as species discovered within 10 years of the group assessment (mammals: 2008, amphibians: 2004, reptiles: 2011, freshwater crabs: 2008, crayfish: 2010, odonates: 2009). Species listed under “few records” were known from five records or fewer. I categorised “old records” as those collected prior to 1970, to ensure comparability with other biodiversity indicators (Collen *et al.* 2008a). I characterized as “severe uncertainty” justifications for Data Deficient status based on type series, few records, old records and records of unknown provenance. I assigned to these eight categories all Data Deficient mammals, reptiles, freshwater crabs, crayfish and odonates, and categorized a random sample of 600 (38%) Data Deficient amphibians. Justifications for listing as Data Deficient were not mutually exclusive; hence a single species may be included under more than one justification.

Results

Cross-taxa congruence in centres of Data Deficient species richness

Pairwise analysis of the geographical distribution of the top 5% richest cells in Data Deficient species showed low congruence among taxonomic groups (Table 2.2 and Figure 2.2). I observed the greatest congruence between reptiles and freshwater crabs (34%), and both groups showed lower levels of congruence with mammals and amphibians (10 – 29%). Crayfish showed lowest congruence with other groups (2 – 5%).

Table 2.2 Matrix of spatial congruence in Data Deficient species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish. The comparison is presented for the richest 5% of cells in each group. Numerical values indicate, for each column, the proportion of hotspot cells encompassed by the hotspot cells of the row. A value of 1 indicates perfect coverage of hotspots of the column taxon by the hotspots of the row taxon.

	Mammals	Amphibians	Reptiles	Freshwater crabs	Crayfish
Mammals		0.33	0.07	0.16	0.02
Amphibians	0.28		0.14	0.25	0.02
Reptiles	0.10	0.24		0.34	0.02
Freshwater crabs	0.16	0.29	0.23		0.02
Crayfish	0.04	0.05	0.02	0.04	

Geographical correlates of the spatial distribution of data deficiency

I investigated the effects of species richness, human population density and remoteness on the spatial distribution of the prevalence of data deficiency (modified incidence proportion; Waller & Gotway 2004). I found that spatial autoregressive models explained 80 to 89% of the spatial variation in data deficiency in mammals, amphibians, reptiles and crayfish; explained variation was lower in freshwater crabs (63.2%; Table 2.3). In all groups, spatial models revealed a strong negative effect of species richness on data deficiency, although the effect decreased for high values of species richness (Table 2.3). The effects of human population density and remoteness varied among levels of species richness and taxonomic groups (Figure 2.3). Densely populated, species-rich areas were more poorly known in amphibians and reptiles, but better-known in freshwater crabs and crayfish. Remote, species-rich areas were more poorly known in mammals, amphibians and reptiles, but better-known in freshwater crabs and crayfish.

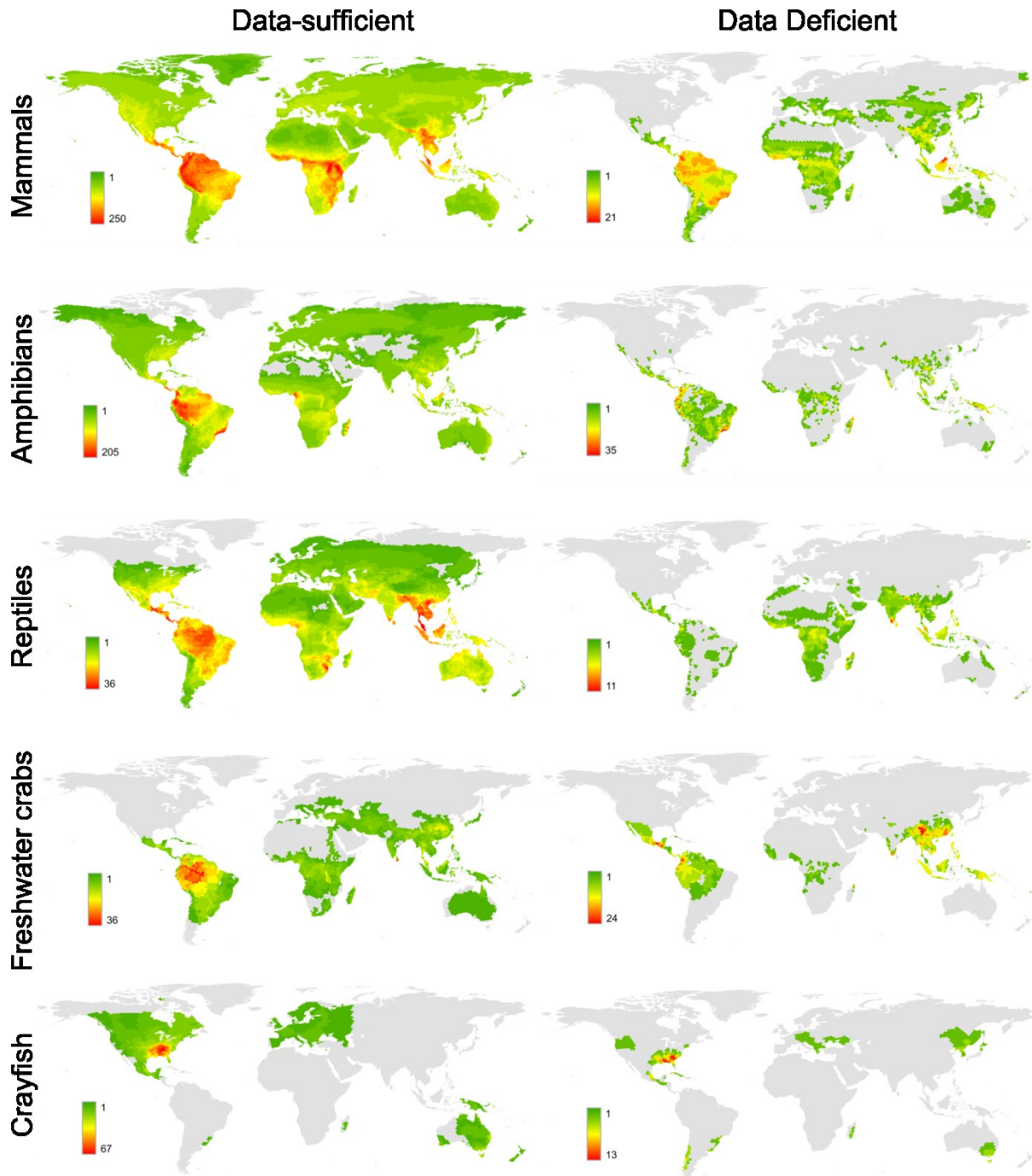


Figure 2.2 Richness of terrestrial and freshwater data-sufficient and Data Deficient species in mammals, amphibians, reptiles, freshwater crabs, and crayfish. Data-sufficient species are listed as Least Concern, Near Threatened, Vulnerable, Endangered or Critically Endangered.

Table 2.3 Spatial correlates of the prevalence of data deficiency in (a) mammals, (b) amphibians, (c) reptiles, (d) freshwater crabs, and (e) crayfish. Parameter estimates are given for spatial autoregressive error models; AIC values are given for both spatial autoregressive (SAR) and ordinary least squares (OLS) models; Nagelkerke R^2 values are given for SAR models, adjusted R^2 for OLS models. S.E.: standard error. HPD: human population density. *: $p < 0.01$, **: $p < 0.001$, *** $p < 0.0001$.

Parameter	Estimate	S.E.	z value
a) Mammals (residual d.f. = 7,536) $AIC_{SAR} = 9,056$ $AIC_{OLS} = 10,324$ $R^2_{SAR} = 0.803$ $R^2_{OLS} = 0.766$			
Intercept	7.45	0.219	33.97***
Species richness	-1.861	0.0257	-72.16***
HPD	0.051	0.004	14.02***
Remoteness	-0.039	0.009	-4.26***
Species richness ²	0.132	0.003	46.65***
Species richness x remoteness	0.064	0.0031	20.04***
b) Amphibians (residual d.f. = 6,014) $AIC_{SAR} = 5,340$ $AIC_{OLS} = 5,786$ $R^2_{SAR} = 0.889$ $R^2_{OLS} = 0.881$			
Intercept	7.32	0.113	64.55***
Species richness	-1.54	0.042	-36.72***
HPD	-0.039	0.008	-4.56***
Remoteness	-0.033	0.014	-2.37
Species richness ²	0.095	0.004	25.9***
Species richness x HPD	0.025	0.004	7.21***
Species richness x remoteness	0.044	0.006	7.59***
c) Reptiles (residual d.f. = 6,228) $AIC_{SAR} = 5,425$ $AIC_{OLS} = 5,766$ $R^2_{SAR} = 0.829$ $R^2_{OLS} = 0.819$			
Intercept	6.99	0.091	76.4***
Species richness	-1.438	0.037	-38.26***
HPD	-0.047	0.005	-8.36***
Remoteness	-0.01	0.008	-1.21
Species richness ²	0.057	0.0053	10.72***
Species richness x HPD	0.061	0.003	21.63***
Species richness x remoteness	0.048	0.005	9.12***
d) Freshwater crabs (residual d.f. = 3,272) $AIC_{SAR} = 4,222$ $AIC_{OLS} = 5,360$ $R^2_{SAR} = 0.632$ $R^2_{OLS} = 0.507$			
Intercept	7.62	0.332	22.91***
Species richness	-0.769	0.081	-9.41***
HPD	-0.149	0.018	-8.17***
Remoteness	0.004	0.017	0.22
Species richness ²	0.162	0.008	18.27***
HPD ²	0.013	0.003	4.69***
Species richness x HPD	0.028	0.006	4.35***
Species richness x remoteness	-0.034	0.011	-2.96*
e) Crayfish (residual d.f. = 1,881) $AIC_{SAR} = 780$ $AIC_{OLS} = 1,688$ $R^2_{SAR} = 0.885$ $R^2_{OLS} = 0.815$			
Intercept	5.53	0.386	14.357***
Species richness	-0.799	0.076	-10.49***
HPD	-0.05	0.016	-3.04*
Remoteness	0.464	0.119	3.87**
Species richness ²	0.12	0.007	16.75***
HPD ²	0.009	0.003	2.82*
Remoteness ²	-0.035	0.009	-3.54**
Species richness x remoteness	-0.04	0.011	-3.41**

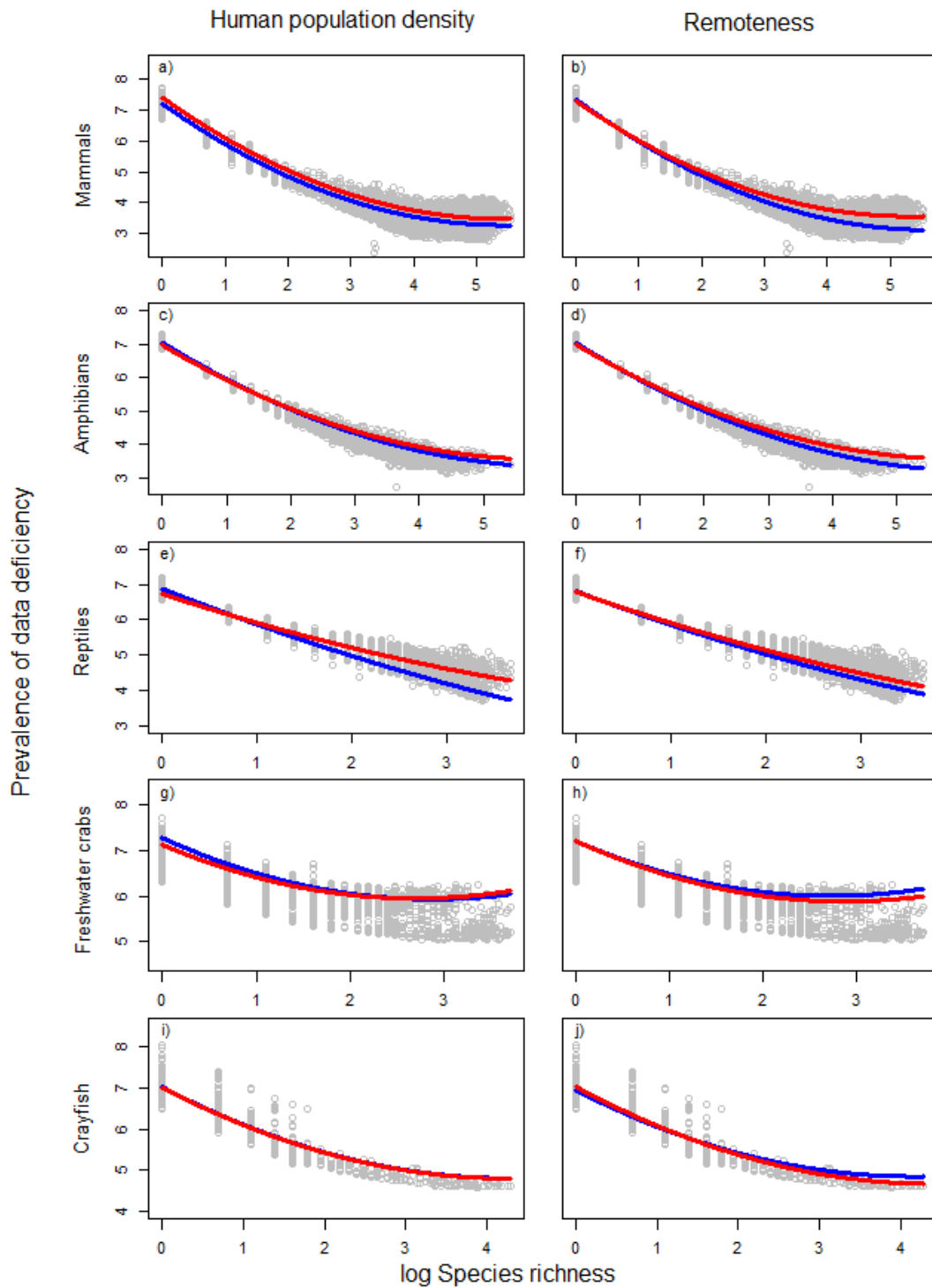


Figure 2.3 Relationship between the prevalence of data deficiency and species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish. The prevalence of data deficiency is measured as the log transformation of the incidence proportion (Waller & Gotway 2004). Grey dots represent the observed prevalence of data deficiency in each cell. Predictions are shown for first quartile (blue) and third quartile (red) values of human population density and remoteness, whilst holding the other variable fixed at its global median. Quartile and median values are available in Table S2.5.

Biological and geographical correlates of Data Deficient species status

Mammals exhibiting small geographical ranges, narrow habitat requirements, and occurring in less remote areas were more likely to be assessed as Data Deficient (single predictor phylogenetic logistic regressions; Table 2.4). Multiple regression on species for which body size was available indicated a negative effect of range size, body size and number of habitats on Data Deficient status. However, this negative effect was weaker for species exhibiting both large range and body size (Table 2.5a). GLMMs for mammals revealed comparable results to the phylogenetic regressions, with fixed effects explaining 21 – 24% of the variation in Data Deficient species status (Table S2.7). Reptiles exhibiting small geographical ranges, narrow habitat requirements and occurring in less remote areas were more likely to be assessed as Data Deficient (single predictor GLMMs; Table 2.6). Multiple regression indicated a similar effect of range size and number of habitats (Table 2.7). Fixed effects explained 36% of the variation in Data Deficient status in reptiles. In crayfish, single and multiple regressions revealed no significant correlates of Data Deficient status (Table 2.6).

Table 2.4 Single predictor phylogenetic logistic regressions of Data Deficient status in mammals. Standard errors of the estimates were obtained with the generalized estimating equations approximation. Residual degrees of freedom in all models equal the total number of species minus three estimated parameters. Among all models phylogenetic signal $\alpha = 0.007$. N_{DS} : number of data-sufficient species. N_{DD} : number of Data Deficient species. S.E.: standard error. HPD: human population density. *: $p < 0.01$, **: $p < 0.001$, *** $p < 0.0001$.

Predictor	N_{DS}	N_{DD}	Estimate	S.E.	t score
Range size	3,655	520	-0.29	0.028	-10.32***
Body size	2,996	220	-0.11	0.068	-1.62
Number of habitats	4,039	559	-1.27	0.176	-7.21***
HPD	3,828	501	-0.04	0.025	-1.60
Remoteness	3,908	534	-0.102	0.039	-2.61*

Table 2.5 Multiple predictor phylogenetic logistic regression of Data Deficient status in mammals. (a) Model including body size: 2,485 data-sufficient and 176 Data Deficient species, 2,858 residual degrees of freedom. (b) Model excluding body size: 3,290 data-sufficient and 410 Data Deficient species, 3,873 residual degrees of freedom. Standard errors of the estimates were obtained with the generalized estimating equations approximation. Across all models phylogenetic signal $\alpha = 0.007$. S.E.: standard error. HPD: human population density. *: $p < 0.01$, **: $p < 0.001$, *** $p < 0.0001$.

Predictor	Estimate	S.E.	t score
a) Including body size			
Intercept	3.74	0.928	4.03***
Range size	-0.42	0.091	-4.59***
Body size	-0.57	0.16	-3.52**
Number of habitats	-1.01	0.19	-5.31***
Range size x body size	0.03	0.013	-2.63*
b) Excluding body size			
Intercept	1.53	0.227	6.74***
Range size	-0.28	0.029	-8.95***
Number of habitats	-0.77	0.103	-7.15***

Table 2.6 Single predictor generalized mixed models of Data Deficient status with nested taxonomic levels in (a) reptiles (b) and crayfish. N_{DS} : number of data-sufficient species. N_{DD} : number of Data Deficient species. S.E.: standard error. HPD: human population density. *: $p < 0.01$, **: $p < 0.001$, *** $p < 0.0001$.

Predictor	N_{DS}	N_{DD}	Estimate	S.E.	z score	Variance due to order; family; genus
a) Reptiles						
Range size	1,124	292	-0.312	0.025	-12.44***	<0.0001;0.595;0.654
Body size	1,019	234	-0.703	0.135	-5.21***	<0.0001;0.324;0.719
Number of habitats	1,124	292	-1.534	0.142	-10.79***	<0.0001;0.148;0.636
HPD	1,108	290	0.05	0.046	1.09	<0.0001;0.215;0.558
Remoteness	1,108	291	-0.297	0.086	-3.43**	0;0.205;0.579
b) Crayfish						
Range size	453	123	-0.04	0.047	-0.906	NA;0;1.55
Body size	450	122	-0.49	0.298	-1.67	NA;<0.0001;1.54
Number of habitats	453	123	-0.46	0.209	-2.18	NA;<0.0001;1.45
HPD	450	123	-0.04	0.106	-0.39	NA;<0.0001;1.56
Remoteness	452	123	0.27	0.215	1.24	NA;<0.0001;1.47

Table 2.7 Multiple predictor generalized mixed model of Data Deficient status in reptiles with nested taxonomic levels. The model is calibrated on 1,108 data-sufficient species and 290 Data Deficient species. AIC = 942.9, marginal $R^2 = 0.364$, conditional $R^2 = 0.499$. Variance due to order: <0.0001 ; family: 0.198; genus: 0.692 S.E.: standard error. HPD: human population density. *: $p<0.01$, **: $p<0.001$, *** $p<0.0001$.

Predictor	Estimate	S.E.	z score
Intercept	1.99	0.33	6.05***
Range size	-0.33	0.033	-10.13***
Number of habitats	-1.09	0.168	-6.52***

Justification for listing as Data Deficient

Severe uncertainty (type series, few records, old records, unknown provenance) was the most frequently used combined justification for listing as Data Deficient in freshwater crabs (92%), dragonflies (83%), amphibians (43%), and mammals (43%) (Figure 2.4). Discovery of new species was the most important single factor in amphibians (24%). Unknown population status and distribution was the main single justification for crayfish (44%), mammals (28%), and reptiles (23%). Large percentages of crayfish (37%) and reptiles (18%) justifications for Data Deficient status invoked unknown threats. Taxonomic uncertainty is an important factor in mammals (16%) and amphibians (13%).

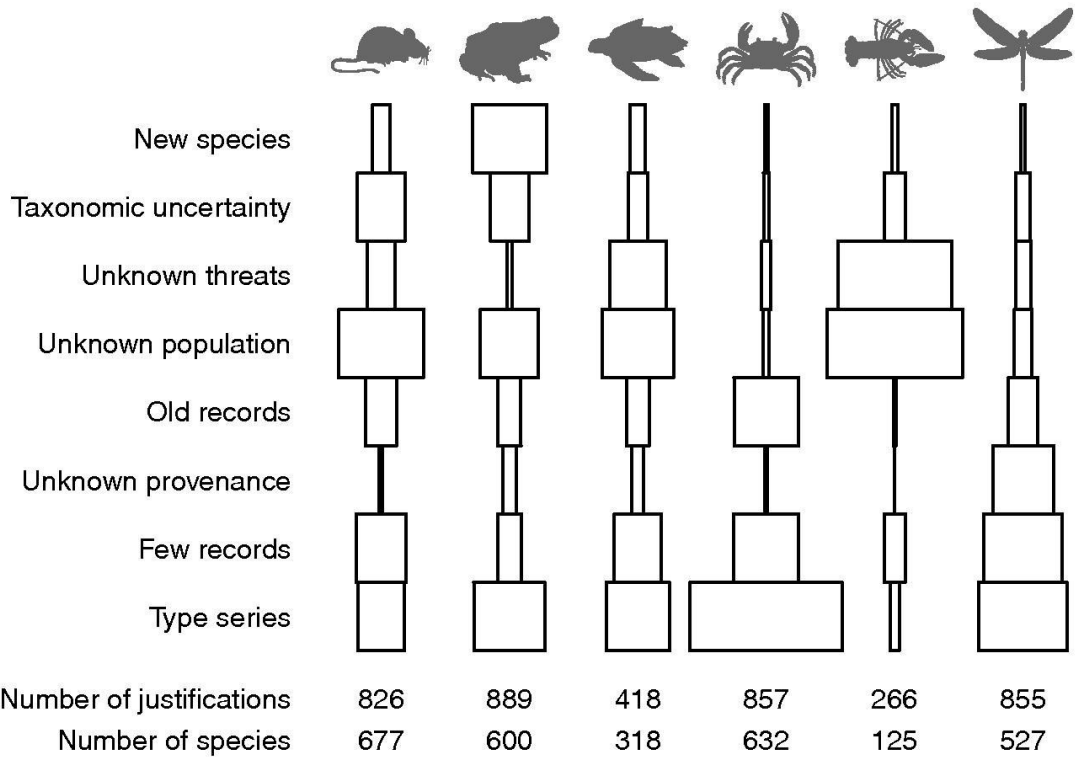


Figure 2.4 Proportional justifications for listing as Data Deficient in mammals, amphibians, reptiles, freshwater crabs, crayfish, and odonates. Species listed under “old records” have only been recorded prior to 1970. Species listed under “few records” are known from five records or less. Species listed under “new species” have been discovered within 10 years of the group assessment. Multiple justifications can apply to one species. Justifications were obtained for all Data Deficient mammals, reptiles, freshwater crabs, crayfish and odonates. Justifications were obtained for 600 Data Deficient amphibians (38% of 1,578 Data Deficient species).

Discussion

In the midst of a sixth extinction crisis, much attention has been given to characterizing global patterns of biodiversity loss (Hilton-Taylor *et al.* 2009; Butchart *et al.* 2010) and devising conservation priorities (Wilson *et al.* 2006; Isaac *et al.* 2007; Leader-Williams *et al.* 2010), but their sensitivity to uneven global data availability has received far less attention. Characterizing the distribution of conservation data deficiency is a first step in assessing the reliability of current biodiversity patterns and prioritizing areas for conservation research. In this chapter, I present the first species-level analysis of global patterns of conservation data deficiency.

Cross-taxa congruence in centres of Data Deficient species richness

I find that congruence in centres of Data Deficient species richness is low, and that no taxonomic groups act as a consistently good indicator of data deficiency in other groups. Determining the effect of uneven data availability on patterns of extinction risk and conservation priorities will therefore require taxon-specific approaches, given the low levels of congruence observed in both patterns of data deficiency and patterns of extinction risk among groups (Grenyer *et al.* 2006; Collen *et al.* 2014). Although reptiles and freshwater crabs exhibit moderate congruence with other groups due to their tropical distribution, marked differences in the distribution of centres of Data Deficient species richness exist within the tropics (Figure 2.2), refuting the existence of a homogeneous tropical data gap (Collen *et al.* 2008b). Western and equatorial Africa exhibit high levels of data deficiency in reptiles, but these regions are disproportionately well-known in freshwater crabs (Cumberlidge *et al.* 2009; Bland *et al.* 2012). Centres of data deficiency occur throughout most of the Neotropics in mammals, whilst hotspots are restricted to the Andes and Brazil's Atlantic forest in amphibians. Indeed, some taxonomic groups exhibit very localized centres of data deficiency: China is home to 174 Data Deficient freshwater crabs, mainly distributed in the southern provinces of Yunnan, Guangxi and Guangdong. The coarse geographical resolution of my study is likely to over-estimate true congruence, which may be particularly low at fine spatial scales relevant to conservation research. As a consequence, field studies directed towards Data Deficient species may not be transferable among taxonomic groups.

Why do I observe low congruence in centres of data deficiency? First, small geographical range sizes can result in low levels of overlap in species diversity (Grenyer *et al.* 2006). In this study however, groups with the smallest ranges do not consistently exhibit lower congruence with other taxonomic groups. Data Deficient mammals (median range size:

11,240 km²) and freshwater crabs (median: 8,906 km²) are wider-ranged than reptiles (median: 3,430 km²) and amphibians (median: 311 km²), but the former do not show higher levels of congruence than the latter (Table 2.2). Second, low congruence in centres of data deficiency may result from differences in the distribution of species richness among groups. Broad-scale patterns of species richness typically show low geographical overlap among groups both in the terrestrial and freshwater realms (Grenyer *et al.* 2006; Collen *et al.* 2014). Crayfish centres of species diversity show very little overlap with other groups, hence very little congruence in centres of data deficiency (2 – 5%). Third, uneven geographical availability of conservation data may result in low hotspot overlap. For example, the Indomalayan realm contains more Data Deficient freshwater crabs than expected by chance, whilst the Neotropical and Afrotropical realms contain fewer than expected (Bland *et al.* 2012). Disentangling the effects of species richness and relative availability of conservation data among regions is therefore crucial to understanding geographical patterns of data deficiency.

Geographical correlates of the spatial distribution of data deficiency

I find high levels of data deficiency in species-poor regions among all groups (Table 2.3), in contrast to coarse scale studies finding a negative correlation between the coverage of biodiversity data and species richness (Collen *et al.* 2008b). The IUCN states that “taxa that are poorly known can often be assigned a threat category on the basis of background information concerning the deterioration of the habitat and/or other causal factors” (IUCN 2001). It may therefore be possible to assign more species to data-sufficient categories when threat processes and ecological requirements can be inferred from multiple co-occurring species. Lower prevalence of data deficiency in species-rich areas may also result from aggregated survey patterns, in which scientists repeatedly select localities with desirable characteristics such as rarity and species richness hotspots (Dennis & Thomas 2000; Sastre & Lobo 2009). The low observed prevalence of data deficiency in species-rich regions has important consequences for conservation. First, low survey effort directed to species-poor localities containing a large proportion of Data Deficient species may efficiently improve estimates of extinction risk. Second, estimates of extinction risk may be more reliable in species-rich areas, which form the basis of many conservation prioritization schemes (Myers *et al.* 2000; Brooks *et al.* 2006).

Human population density and remoteness show inconsistent relationships with the prevalence of data deficiency among groups. Sparsely populated areas are consistently better-known in mammals; this is only the case in species-rich areas for amphibians and reptiles.

The negative effect of human population density on data availability in my global analysis may result from the high prevalence of data deficiency in poorly-sampled, but heavily populated areas of southern China, India and South East Asia. Characterizing the effect of human population density on data deficiency is difficult due to the dynamics of human settlement and biodiversity inventories. For example, human population size is positively correlated with anuran description date in the Brazilian Cerrado, due to temporal patterns of human colonization and description of species differing in range size (Diniz-Filho *et al.* 2005). Relationships between species richness and human population do not consistently persist when sampling effort is controlled for (Barbosa *et al.* 2010a; Cantarello *et al.* 2010; Luck *et al.* 2010; McKinney 2010), suggesting a complex interaction between human settlement, biodiversity study and spatial scale.

Because biological collections in inaccessible areas are often limited (Reddy & Davalos 2003; Tobler *et al.* 2007; Ficetola *et al.* 2013), remoteness may provide a more appropriate measure of global sampling effort. I find that remoteness interacts with species richness to determine patterns of data deficiency: in vertebrates, species-rich areas tend to be well-known unless those are inaccessible. Overall, the effect of human population density and remoteness in invertebrates was small, likely due to the low variability in these variables within the groups' distribution (e.g. 1 – 753 people/km² in crayfish, compared to 0 – 4,636 people/km² in mammals), or due to similar sampling efforts along population density and remoteness gradients.

Correlates and justifications of Data Deficient species status

Habitat specialist and narrow-ranged species are likely to experience low encounter rates with naturalists (Patterson 1994; Gaston *et al.* 1995), leading to high levels of data deficiency in vertebrates. However, low estimates of habitat breadth and range size for Data Deficient species may also result from lack of research itself. Small body size is often associated with late description date (Blackburn & Gaston 1994; Collen *et al.* 2004; Diniz-Filho *et al.* 2005) and low scientific attention (Brodie 2009), but small body size only predicts Data Deficient status in mammals. The role of body size may therefore vary with different metrics of species knowledge, or may only play a substantial role in groups that are extensively studied and show high body size variation, such as mammals and birds (Bonnet *et al.* 2002; May & Clark 2002).

The limited power of biological and geographical characteristics in predicting Data Deficient species status may result from differences in the application of the Data Deficient category within and among groups. Future studies could therefore investigate the role of species traits

in the application of Data Deficient species justification labels. Severe uncertainty (species known only from type series, few records, old records or records of unknown provenance) is involved in more than three quarters of freshwater crabs and odonate listings, and almost half of mammal and amphibian listings. Information is particularly scarce for species of uncertain geographical provenance (e.g. the dragonfly *Oligoaeschna speciosa* is only known from "Darjeeling, North East India"), or species which cannot be matched to wild individuals (e.g. the frog *Hyperolius fuscigula*). Continued investment in taxonomy is paramount to keeping the Red List up-to-date with recent species discoveries (Mace 2004), and reducing the proportion of species listed as Data Deficient due to taxonomic uncertainty. This proportion can be high even in well-known clades, such as birds and mammals (15 - 16%; Butchart & Bird 2010). In addition, precise geo-location data for specimens is essential to determining species' distributions, inferring threats and field surveying. Although the lack of information on threats and their impact on species has often been highlighted (Murray *et al.* 2014), I show that lack of natural history information is the main limiting factor in conducting conservation assessments. Only in crayfish, a relatively species-poor clade whose centres of diversity are located in developed countries (USA and Australia) did lack of information on population trends and threats justify most Data Deficient listings.

The Data Deficient category therefore reflects a spectrum of data deficiencies and shows considerable heterogeneity in its application. Transparent and consistent documentation for the category would increase comparability of assessments, and enable the prioritization of Data Deficient species for research and re-assessment to data-sufficient categories. I recommend the application of my Data Deficient justification labels to future assessments by the IUCN. Consistent information on date of first and last sightings and recent field surveys is also desirable to quantify information availability on Data Deficient species. Finally, I find that semantic uncertainties in the assessment of Data Deficient species (e.g. concerning the number, age and locality of records) considerably reduce the utility of Data Deficient assessments for conservation decision-making.

Limitations and prospects

I highlight that my results are conditional on the current state of knowledge in the groups investigated. Reptiles and odonates were assessed with the Sampled Red List approach (Baillie *et al.* 2008) rather than complete group assessments, which may increase uncertainty in observed patterns of data deficiency due to reduced sample sizes. Geographical range maps were not available for odonates, and near-complete species trait data were not available for amphibians and freshwater crabs. Lack of available phylogenetic information

precluded the use of comparative methods in reptiles and crayfish, but the similarity of results obtained in mammals with phylogenetic regressions and GLMMs may indicate the robustness of results derived from taxonomic information. Furthermore, estimates of species richness are sensitive to information availability, as taxonomies are revised and new species are discovered (Isaac *et al.* 2004; Scheffers *et al.* 2012). Only half of the estimated number of freshwater crab species have been discovered (Yeo *et al.* 2008), so conclusions for this group are subject to more uncertainty than for groups where fewer species remain to be discovered, such as mammals (Giam *et al.* 2012).

My study relies on the application of the Data Deficient category within the IUCN Red List system to quantify species knowledge. Species description dates, occurrence records and bibliometric information may reveal different patterns of knowledge availability, and should become the focus of systematic global comparisons among groups. Understanding the degree to which components of diversity are studied is crucial to strengthening indicators of biodiversity loss, and planning conservation research actions. Indicators of biodiversity knowledge should therefore be developed based on a wide range of data, such as data gaps within indicators of biodiversity loss (e.g. the Living Planet Index) and species trait databases (e.g. panTHERIA; González-Suárez *et al.* 2012). Such indicators should show desirable properties (Failing & Gregory 2003; Jones *et al.* 2011), including the ability to dynamically reflect diverse facets of biodiversity knowledge, and inform international targets of biodiversity data acquisition. Finally, explicitly considering the value of information (Dakins 1999; Yokota & Thompson 2004) would ensure acquired biological data meet the needs of applied conservation problems.

Conclusions

This study demonstrates that patterns of conservation data deficiency are not congruent among groups, and that these are primarily driven by spatial patterns of ecological research rather than species' biological characteristics. I conclude that integrating taxon-specific processes of biodiversity data collection is necessary to designing representative global conservation schemes. My study highlights the importance of taxonomic and fundamental ecological information in conservation assessments, and calls for renewed investment in taxonomy and field inventories globally. Creating indicators of biodiversity knowledge is paramount to designing robust conservation and data collection schemes, particularly for the world's poorly known and speciose taxa.

Chapter 3. Data deficiency and the selectivity of extinction risk in freshwater invertebrates

A version of this chapter is published as:

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

Introduction

With current species extinction rates exceeding geological background rates by several orders of magnitude, it is now clear that we are facing an extinction crisis comparable to mass extinctions of the paleontological past (May *et al.* 1995; Millenium Ecosystem Assessment 2005; Butchart *et al.* 2010). Yet species are not equally at risk of extinction (Purvis *et al.* 2000a). Extinction risk has been found to be non-randomly distributed among many groups, including families of birds (Bennett & Owens 1997), amphibians (Stuart *et al.* 2004; Bielby *et al.* 2006) and mammals (Purvis *et al.* 2000b). This phenomenon, termed taxonomic selectivity of extinction risk, has not only been observed in extant taxa but also in historical and paleontological patterns of extinction. For example, over the last 75 million years, some echinoid genera have been more likely to go extinct than others (McKinney 1997). The phylogenetically clumped nature of threat has severe implications for the loss of biodiversity. The loss of all constituent species of a clade and their evolutionary history is more likely under non-random extinction than under random extinction (Russell *et al.* 1998), and as a consequence, non-random extinction risk results in the more rapid loss of higher taxa and phylogenetic diversity than is predicted by random extinction (Nee & May 1997; Purvis *et al.* 2000a).

Extinction risk is also known to be non-randomly distributed across geographical areas; prevalence of extinction risk is higher where threatening processes such as habitat degradation, overexploitation, invasive species and diseases are more intense (Kerr & Currie 1995; McKinney 1997). Taxonomic and geographical selectivity are not independent, as evolutionary diversification within regions produces phylogenetic proximity that is often correlated with geographic proximity (Brooks *et al.* 1993). Geographical scale also affects patterns of taxonomic selectivity, and the concordance between extinction risk at a local and

global level can be low (Purvis *et al.* 2005). The persistence of taxonomic patterns of selectivity observed at the global scale at smaller spatial scales (where threatening processes are more homogenous) can indicate the importance of species biology in determining susceptibility to extinction risk (Bielby *et al.* 2006).

Studies of global extinction risk have primarily focused on birds, mammals and amphibians (Owens & Bennett 2000; Purvis *et al.* 2000b; Cardillo *et al.* 2004; Cardillo *et al.* 2005; Davies *et al.* 2006; Cardillo *et al.* 2008; Cooper *et al.* 2008; Lee & Jetz 2011), whilst the macroecology of invertebrates remains largely under-studied (Diniz-Filho *et al.* 2010). IUCN Red List assessments and global distribution maps have now been made for a range of invertebrate groups (Baillie *et al.* 2008), offering considerable scope for macroecological research and the development of conservation strategies based on broad-scale data. In particular, data on freshwater invertebrates provide a unique opportunity to characterize extinction risk patterns in highly threatened, yet neglected ecosystems (Millenium Ecosystem Assessment 2005; Revenga *et al.* 2005). The imperilment of freshwater systems also has direct links with human well-being and water security (Millenium Ecosystem Assessment 2005; Vorosmarty *et al.* 2010), and a representative picture of freshwater species conservation status is necessary for successful integrated water management and climate change adaptation (Strayer & Dudgeon 2010). Because global priorities for biodiversity conservation have been largely biased towards vertebrate species and terrestrial ecosystems, understanding the drivers of extinction risk in freshwater invertebrates will contribute to a more accurate picture of biodiversity as a whole.

However, high levels of data deficiency in IUCN Red List assessments for freshwater invertebrates could bias the results of broad scale studies based on these assessments. The Data Deficient (DD) category is assigned to a species “when there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status” (IUCN 2001). To date, all invertebrate taxa with systematic risk assessments show high proportions of Data Deficient species: 35% of dragonflies and damselflies (odonates; Clausnitzer *et al.* 2009), 49% of freshwater crabs (Cumberlidge *et al.* 2009), and 21% of crayfish (Samways & Böhm 2010) are currently listed as Data Deficient. Vertebrate groups are typically better known, with only 1% of birds, but 15% of mammals and 19% of reptiles listed as Data Deficient (Collen *et al.* 2009; Hilton-Taylor *et al.* 2009). Data Deficient species are often simply excluded from calculations on a taxon’s conservation status, but this approach disguises considerable taxonomic and geographical uncertainty in the distribution of risk. If the distribution of Data Deficient species is itself non-random among families and geographical regions, treating Data Deficient species as either all

threatened or non-threatened could dramatically alter observed patterns of extinction risk. Given the limited resources available for conservation, it is important to use all available information to prioritise taxa and geographical regions effectively (Leader-Williams *et al.* 2010). Disentangling the effects of the distribution of Data Deficient species from the observed distribution of threat is therefore crucial to obtaining a more accurate picture of biodiversity.

In this chapter, I investigate the selectivity of data deficiency and extinction risk in three groups of freshwater invertebrates: crayfish, freshwater crabs and odonates. I focus on four questions:

- i)* Is there evidence for taxonomic and geographical selectivity of data deficiency in freshwater invertebrates?
- ii)* Is there evidence for taxonomic and geographical selectivity of extinction risk in invertebrates under different treatments of Data Deficient species (Data Deficient species excluded, Data Deficient species considered non-threatened, and Data Deficient species considered threatened)?
- iii)* What are the effects of geographical scale on the taxonomic selectivity of data deficiency and extinction risk?
- iv)* Are there differences in the selectivity of data deficiency and extinction risk among invertebrate taxa, and among vertebrate taxa?

Methods

Data

I gathered species data from three recent freshwater invertebrate assessments: crayfish (all 586 species; IUCN 2011), freshwater crabs (all 1,281 species; Cumberlidge *et al.* 2009) and odonates (a randomly selected sample of 1,500 out of 5,680 species; Clausnitzer *et al.* 2009). I followed the IUCN taxonomy and identified the number of threatened (VU, EN and CR categories), non-threatened (LC and NT categories), and Data Deficient (DD) species in each group (Table 3.1). From the published assessments, I recorded for each species its taxonomic family and the biogeographic realm (Olson *et al.* 2001). I also used the justification for listing as Data Deficient to assign each species to one of eight categories (Figure 2.4).

Table 3.1 Number of threatened species, Data Deficient (DD) species and non-threatened species in crayfish, freshwater crabs and odonates.

Taxon	Number of threatened species	Number of DD species	Number of non-threatened species
Crayfish	146	125	315
Freshwater crabs	202	632	447
Odonates	135	527	838

Analyses

I tested for non-randomness in the distribution of both data deficiency and extinction risk within each invertebrate group, using Fisher tests on the number of species in each category due to the low diversity of some families and realms. First, I tested for global taxonomic non-randomness in the prevalence of Data Deficient and threatened species among families. I did not conduct analyses at the genus level due to the small size of some of the genera in the groups considered. Second, I tested for global geographic non-randomness in the prevalence of Data Deficient and threatened species among realms. Taxonomic and geographic selectivity are likely to be non-independent due to the strong biogeographic structure of families across realms and differences in threat pressure between realms. As a consequence, I tested for the presence of taxonomic selectivity of data deficiency and extinction risk within realms, only using data from realms that contained more than 30 species (crayfish: Nearctic; freshwater crabs: Afrotropical, Indomalayan and Neotropical; odonates: Afrotropical, Australasian, Indomalayan, Neotropical and Palearctic).

In each case, I tabulated the number of Data Deficient or threatened species against the number of data-sufficient and non-threatened species in each family or realm. I investigated the presence of non-random extinction risk under three scenarios representing uncertainty about the conservation status of Data Deficient species: Data Deficient species excluded (assumed as threatened as data-sufficient species), all Data Deficient species considered threatened and all Data Deficient species considered non-threatened. Despite testing distinct hypotheses, the analyses used different combinations and subsets of the same underlying tables. I therefore used Benjamini & Hochberg's (1995) correction for multiple hypothesis tests across all the Fisher's tests for each taxonomic group. All tests were conducted using R version 2.12.0 (R Development Core Team 2010). I interpreted significant associations in these tests by examining the magnitude and size of the difference between observed and

expected species number. I display observed and expected numbers of Data Deficient and threatened species for the 29 odonate families in Table S3.1.

Results

Taxonomic and geographical selectivity of data deficiency

The prevalence of data deficiency did not differ among families of crayfish, but differed among families of freshwater crabs and odonates (Table 3.2). Three families of freshwater crabs (Gecarcinucidae, Potamidae and Pseudothelphusidae) showed a higher proportion of Data Deficient species than other crab families (Table 3.2). All three groups exhibited geographic structure in the prevalence of Data Deficient species, with different realms showing higher numbers than expected in each group (Table 3.3).

Taxonomic and geographical selectivity of extinction risk

In crayfish, risk was unevenly distributed among families (Table 3.2) and biogeographical realms (Table 3.3). Global patterns of taxonomic and geographical selectivity in crayfish were robust to the different treatments of Data Deficient species and consistent with each other, as expected from the clumped geographic distribution of crayfish families. In freshwater crabs and odonates, the strength of the association between threat status and families (Table 3.2) or biogeographical realms (Table 3.3) varied with the treatment of Data Deficient species, with weaker associations observed when Data Deficient species were considered non-threatened. In both groups, the distribution of threatened species among realms was identical when Data Deficient species were excluded or considered non-threatened. The distribution varied in certain realms when Data Deficient species were considered threatened, although differences between the expected and observed levels of threat were small. The geographical distribution of risk among all scenarios was strikingly similar between freshwater crabs and odonates, except in the Australasian realm which contains few freshwater crab species.

Table 3.2 Global taxonomic selectivity of data deficiency and extinction risk in crayfish, freshwater crabs and odonates. The observed and expected distributions of DD (Data Deficient) or threatened species among families are indicated along with the results of the Fisher's exact tests. ns= non-significant. *p<0.05 ** p<0.01 *** p<0.001.

	Data Deficiency			DD species excluded			DD species non-threatened			DD species threatened		
	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend
Crayfish	ns			***			***			***		
Astacidae	3	2.2	+	3	2.4	+	3	2.6	+	6	4.7	+
Cambaridae	91	82.9	+	71	95.3	-	71	96.9	-	162	180	-
Parastacidae	31	39.9	-	72	48.3	+	72	46.6	+	103	86.3	+
Freshwater crabs	***			**			*			***		
Gecarcinucidae	28	26.5	+	9	8.1	+	9	8.5	+	37	35.2	+
Parathelphusidae	116	142.9	-	77	54	+	77	45.7	+	193	188.8	+
Potamidae	305	249.1	+	53	62.2	-	53	79.7	-	358	328.8	+
Potamonautidae	31	65.5	-	28	31.9	-	28	21	+	59	86.6	-
Pseudothelphusidae	144	124.5	+	31	33.7	-	31	39.8	-	175	164	+
Trichodactylidae	8	23	-	4	12.1	-	4	7.4	-	12	30.6	-
Odonates	***			***			*			***		

Table 3.3 Geographical selectivity of data deficiency and extinction risk in crayfish, freshwater crabs and odonates. The observed and expected distributions of DD (Data Deficient) or threatened species among biogeographical realms are indicated along with the results of the Fisher's exact tests. ns= non-significant; *p<0.05; ** p<0.01; *** p<0.001.

	Data Deficiency			DD species excluded			DD species non-threatened			DD species threatened		
	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend
Crayfish		***			***			***			***	
Afrotropics	4	1.5	+	2	0.95	+	2	1.7	+	6	3.2	+
Australasia	17	32	-	70	42.3	+	70	37.7	+	87	69.7	+
Nearctic	81	81.2	-	65	95.3	-	65	95.2	-	146	176.4	-
Neotropics	19	9.5	+	10	8.2	+	10	11.2	-	29	20.8	+
Palaearctic	6	2.6	+	2	1.9	+	2	3	-	8	5.5	+
Freshwater crabs		***			**			*			***	
Afrotropics	32	67.2	-	28	32.1	+	28	21.3	+	60	88.4	-
Australasia	21	16.3	+	2	3.7	-	2	5.1	-	23	21.1	+
Indomalaya	426	392.7	+	136	114.6	+	136	124.7	+	561	516.9	+
Neotropics	152	147.6	-	35	45.4	-	35	46.9	-	187	194.1	-
Palaearctic	5	12.3	-	1	6.2	-	1	3.9	-	6	16.3	-
Odonates		***			***			***			***	
Afrotropics	76	84.3	-	22	21.1	+	22	19.9	+	98	104.4	-
Australasia	84	69.7	+	21	14.7	+	21	16.5	+	105	86.4	+
Indomalaya	185	136.5	+	55	26	+	55	32.3	+	240	168.6	+
Nearctic	13	56.9	-	2	19	-	2	13.5	-	15	70.4	-
Neotropics	132	144.4	+	23	35.8	-	23	34.1	-	155	178.6	-
Oceania	15	7.3	+	0	0.7	-	0	1.8	-	15	9.1	+
Palaearctic	57	62.9	-	10	15.6	-	10	14.9	-	67	77.7	-

Table 3.4 Taxonomic selectivity of data deficiency and extinction risk of freshwater crabs within biogeographical realms. The observed and expected distributions of DD (Data Deficient) or threatened species among families are indicated along with the results of the Fisher’s exact tests. No significant taxonomic selectivity in data deficiency or extinction risk was detected in the Afrotropical realm. There was no significant taxonomic selectivity of extinction risk when DD species were considered threatened in the Indomalayan realm, or considered non-threatened in the Neotropical realm. ns= non-significant; *p<0.05; ** p<0.01; *** p<0.001.

Indomalaya	Data deficiency ***			DD species excluded **			DD species non-threatened **		
	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend
Gecarcinucidae	28	28.9	-	9	9.5	-	9	9.2	-
Parathelphusidae	94	137.4	-	75	59.8	+	75	43.9	+
Potamidae	304	259.8	+	52	66.7	-	52	82.9	-
Neotropics	Data Deficiency ***			DD species excluded *			DD species threatened ***		
	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend
Pseudothelphusidae	144	128.1	+	31	25.7	+	175	157.6	+
Trichodactylidae	8	23.9	-	4	9.3	-	12	29.4	-

Table 3.5 Taxonomic selectivity of data deficiency and extinction risk of odonates within biogeographical realms. There was no significant taxonomic selectivity in data deficiency or extinction risk in the Australasian realm. DD: Data Deficient. ns= non-significant; *p<0.05; ** p<0.01; *** p<0.001.

	Data Deficiency	DD species excluded	DD species non-threatened	DD species threatened
Afrotropics	***	ns	ns	***
Indomalaya	***	**	ns	***
Neotropics	***	ns	ns	***
Palaearctic	***	ns	***	***

Effect of geographical scale on taxonomic selectivity

Only the Nearctic realm was suitable for the investigation of the taxonomic selectivity of crayfish at the sub-global level, and showed no significant selectivity in data deficiency or extinction risk. For freshwater crabs, the Afrotropical, Indomalayan and Neotropical realms were suitable for the investigation of taxonomic selectivity at the sub-global scale. No significant taxonomic selectivity in data deficiency or extinction risk was detected in the Afrotropical realm. In the Indomalayan and Neotropical realms (Table 3.4), data deficiency was not randomly distributed among families, and selectivity of extinction risk varied with the scenarios considered. Selectivity of extinction risk was only significant in the Indomalayan realm when Data Deficient species were excluded or considered non-threatened, whereas in the Neotropical realm selectivity was only significant when Data Deficient species were excluded or considered threatened. In these cases, the distribution of threat among families was congruent among scenarios.

I investigated the presence of taxonomic selectivity of odonates in the Afrotropical, Australasian, Indomalayan, Neotropical and Palearctic realms (Table 3.5). The distribution of Data Deficient species and threatened species under the three scenarios was not significantly different from random in the Australasian realm. Data deficiency was non-randomly distributed in the remaining realms, and I generally detected no selectivity of extinction risk when Data Deficient species were excluded or considered non-threatened. On the other hand, taxonomic selectivity was highly significant when Data Deficient species were considered threatened.

Discussion

Freshwater invertebrate conservation faces huge challenges due to the increasing pressures humans are imposing on freshwater systems (Jackson *et al.* 2001; Malmqvist & Rundle 2002) and the very limited resources, both in terms of money and scientific effort, allocated to their conservation (Strayer 2006). For these species the analysis of global datasets, as opposed to the study of local species and populations, could provide more widely applicable results and recommendations whilst taking into account heterogeneity among phylogenetic and geographical subsets. In this study, I investigated the robustness of observed macroecological patterns of risk in freshwater invertebrates to high levels of data deficiency. The three taxa included the analyses all show large differences in ecology, geographical distribution, levels of data deficiency and risk, and give different perspectives on the selectivity of extinction risk in the freshwater realm.

Taxonomic and geographical selectivity of data deficiency

The Nearctic and Australasian realms are well-studied centres of crayfish diversity, containing 91% of all crayfish species. The other realms, despite containing 9% of species, are home to 23% of the Data Deficient species. In these realms information on population trends is especially scarce and there is little understanding of the effects of threats on crayfish populations. While Data Deficient crayfish typically lack detailed information on the impact of threats and trajectory of population trends, most Data Deficient freshwater crabs are only known from one or two geographical locations, with little or no information on their extent of occurrence, ecological requirements and population size (Figure 2.4). Data Deficient freshwater crab species are concentrated in species-rich clades and regions, such as the family Potamidae and the Indomalayan realm. Freshwater crabs exhibit high levels of endemism, have restricted ranges and often occupy remote habitats (Cumberlidge *et al.* 2009), which may be the reason why, in association with limited monitoring effort, many freshwater crab species are assessed as Data Deficient. While some Data Deficient freshwater crabs may be naturally rare and therefore more likely to be classified in a non-threatened category, there is still a high chance that some species, which have not been observed in decades (e.g. *Rouxana papuana* from Indonesia has not been observed in over a century) and have had their habitat transformed by human activity (e.g. *Thaipotamon siamense* from Thailand), may well be extinct. For such species, the only recourse is to initiate targeted surveys to confirm status.

Freshwater crabs and odonates exhibit similar patterns in the selectivity of data deficiency. First, the geographical distribution of data deficiency in freshwater crabs and odonates is consistent with the commonly observed tropical biodiversity data gap (Collen *et al.* 2008b). Second, most Data Deficient odonates are only known from a very few locations and specimens, as for freshwater crabs. The lack of information about Data Deficient odonates is particularly alarming, as records for 168 species are from unknown provenance; these species will be especially difficult to re-assign to data-sufficient categories in the future.

Taxonomic and geographical selectivity of extinction risk

The three taxonomic groups under consideration not only show considerable differences in patterns of selectivity of extinction risk, but also in the influence of data deficiency on these patterns. Australasian crayfish species (parastacids) remained over-threatened and Nearctic species (mostly cambarids) under-threatened in all scenarios, indicating that genuine differences in extinction risk exist between the two groups. Freshwater crayfish in the Australasian realm are particularly exposed to the threats of sedimentation due to agricultural and forestry effluents, habitat destruction, bush fires, droughts and over-

exploitation. Australian crayfish also appear to be extremely susceptible to the effects of climate change, including increasing temperatures, alterations to hydrological regimes and loss of suitable highland habitat (Chiew & McMahon 2002; Hughes 2003).

The strength of extinction risk selectivity varied across scenarios in freshwater crabs and odonates, with selectivity being reduced when Data Deficient species were considered non-threatened. This is due to the overall lower prevalence of threat in all families when Data Deficient species are considered non-threatened. The freshwater crab families Gecarcinucidae, Parathelphusidae and Trichodactylidae were consistently over or under-threatened across all scenarios of risk distribution, whereas patterns varied across scenarios for the remaining families. Similarly, the reliability of extinction risk trends among scenarios varied among biogeographical realms. My results indicate that the current understanding of risk patterns in freshwater crabs is heavily influenced by data deficiency. Additional work to determine the true extinction risk of these species is therefore needed before these data can inform conservation prioritisation with confidence. Formulating hypotheses concerning the relative roles of biological traits and threatening processes in determining extinction risk in the taxon is also problematic. Semi-terrestrial species, stenotopic species and island endemics are thought to be more susceptible to anthropogenic habitat disturbance (Cumberlidge *et al.* 2009) but evidence is scarce and mostly derived from other groups such as amphibians (Sodhi *et al.* 2008). The extent of congruence in predictors of extinction risk among freshwater groups is therefore a useful avenue for further research.

Odonates showed very strong geographical non-randomness of extinction risk in all scenarios, and qualitatively consistent trends among biogeographical realms. Species in the Indomalayan realm were more threatened than expected by chance regardless of how Data Deficient species were assigned, which may be due to the high number of endemic species in the Indomalayan islands and large-scale logging of lowland forests (Clausnitzer *et al.* 2009). On the other hand, the Nearctic and Neotropical realms were consistently under-threatened. Temperate species, including Nearctic species, have suffered declines in the second half of the 20th century but many are recovering due to improved water management (Kalkman *et al.*, 2008). These species also tend to have wider distributions than their tropical counterparts, and may be more able to recover from local scale population decline (for a global review of threats affecting odonates, see articles in Clausnitzer & Jödicke 2004). My results indicate that the geographical distribution of risk in crayfish and odonates is reliable and could be used to efficiently determine the allocation of conservation resources to certain geographical regions.

Effect of geographical scale on taxonomic selectivity

Taxonomic selectivity of extinction risk at the global level may simply be a by-product of geographical selectivity, as clades endemic to certain regions may experience different intensities of threatening processes (Russell *et al.* 1998; Bielby *et al.* 2006). If this were the case, families should display similar levels of risk among geographical scales. On the other hand, the persistence of taxonomic selectivity at smaller scales would suggest that biological differences are at least partially responsible for the observed selectivity. The effect of taxonomic selectivity within geographic regions could not be investigated in detail in crayfish as crayfish families do not often co-occur in biogeographical realms. However, there is evidence for significant differences in extinction risk among genera of crayfish living in the same USA state (Adamowicz & Purvis 2006). While at the global level it may be difficult to disentangle the effects on crayfish extinction risk of common evolutionary history from the geographical distribution of threatening processes, at smaller geographical scales species traits seem to be important. This finding has substantial implications for the creation of predictive models of extinction risk in crayfish, in which biological traits, geographical factors and their interaction are likely to determine risk.

The analysis of taxonomic selectivity at sub-global scales in freshwater crabs and odonates revealed some complex patterns in information availability and risk prevalence. Data deficiency was generally unevenly distributed among freshwater crab and odonate families within realms, hence poorly-known families should become the target of conservation research at sub-global scales. There was little consistent evidence for taxonomic selectivity of extinction risk at sub-global scales in both groups, either among biogeographical realms or among scenarios. I therefore can neither accept nor refute the hypothesis that geographic differences in threat intensity are responsible for the non-random pattern in extinction risk at the global level. My results indicate that understanding the effects of family-specific attributes on extinction risk is difficult when the proportion of Data Deficient assessments in a group is large.

Limitations

My study constitutes a coarse analysis of the factors that determine global extinction risk in invertebrates, and as such observed patterns cannot be easily attributed to processes that occur at a local scale. Additionally, different proportions of Data Deficient species from each family or realm may be threatened, rather than none or all, and as a consequence my analyses are likely to underestimate the effect of Data Deficient species on the distribution of threatened species. Any statements made on the reliability of observed trends are limited by

the scenarios considered and the best information available to date. However, the analyses presented here encompass a wide range of scenarios, and my conservative approach further highlights the crucial influence of Data Deficient species in the selectivity of extinction risk.

Overall findings for each group could also be sensitive to the assessors' attitude to Red Listing. IUCN's quantitative criteria for the assignment of Red List categories ensures that in threatened categories at least, subjectivity of assessors plays a minor role in differences across taxa and regions. The case is less clear for Data Deficient species listings, where this category marks a lack of information, or understanding, on a given taxon's status. The attitude of the assessors involved in the crayfish, freshwater crab and odonate assessments towards Data Deficient listings may vary; however, clear guidelines were used to assign this category, all the assessments I used were coordinated by one Red List assessor (B. Collen), and were passed through the IUCN verification system, which should go some way to minimising any potential effect on my results.

Conclusions

My study shows that the effect of Data Deficient species on the selectivity of extinction risk is not only dependent on the absolute number of Data Deficient species in the taxon, but also on the distribution among families and realms of these Data Deficient species. Global patterns of taxonomic selectivity and geographical selectivity were generally consistent with one another, and robust to different treatments of Data Deficient species. At sub-global scales it was not possible to disentangle the effects of common evolutionary descent from those of information availability on extinction risk selectivity. Taxonomic selectivity in amphibians has been shown to be independent of both geographical effects and differences in knowledge of species conservation status (Bielby *et al.* 2006). However, given the current amounts of data deficiency, the relative importance of family-specific characteristics and threatening processes in driving extinctions in freshwater invertebrates cannot be determined. While the understanding of extinction risk in freshwater invertebrates remains compromised by high levels of data deficiency, prioritisation of freshwater invertebrates for conservation at the sub-global scale remains a challenge.

Given the significant impact of Data Deficient species on the understanding of patterns of risk of invertebrates, Data Deficient species should be given high research priority to determine their true status. Ideally this should be done through field assessments, but the use of contextual information, expert opinion, techniques combining information about collection efforts with the geographical location of specimens (Good *et al.* 2006) or the outputs of predictive extinction risk modelling may also allow the preliminary re-assignment

of a large number of Data Deficient species. Broadening the coverage of biodiversity assessments to under-studied taxa and systems is essential to developing a more representative picture of biodiversity. Despite recent efforts toward achieving this goal, my study shows that high level of data deficiency challenge the integration of these assessments into conservation decision-making, and supports the need for increased efforts in invertebrate study and conservation.

Chapter 4. Predicting the conservation status of Data Deficient species

A version of this chapter is in revision in *Conservation Biology*.

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 towards targets for the Convention on Biological Diversity (Jones *et al.* 2011), and one of the most widely used is the IUCN Red List (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 70,000+ 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 2013b). 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 Data Deficient. Uncertainty within many groups about the true level of extinction risk of Data Deficient species considerably influences our understanding of patterns of threat and risk (Bland *et al.* 2012), as the distribution of Data Deficient 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 Data Deficient species were non-threatened to 36% if all Data Deficient species were threatened (Hilton-Taylor *et al.* 2009). In addition, genuinely threatened Data Deficient species may be neglected by conservation programmes due to their uncertain extinction risk status.

Determining the true conservation status of Data Deficient species is essential to developing an accurate picture of global biodiversity and enabling the protection of threatened species.

Re-assessment of the 10,673 species currently classified as Data Deficient 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 Data Deficient species. The distribution of many Data Deficient 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 (Cardillo & Meijaard 2012; Purvis 2008), and could enable the preliminary re-assessment of Data Deficient species.

Comparative datasets frequently contain many variables, with non-linear relationships, 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 use 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 (No Free Lunch Theorem: see Webb 2002 and 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 deriving rules-of-thumb to predict the level of risk faced by Data Deficient species.

In this chapter I investigate the performance of ML algorithms in predicting extinction risk and in estimating the prevalence of risk in Data Deficient terrestrial mammals. Terrestrial mammals are a well-suited model taxon for the purposes of this study: they contain a high proportion of species of known conservation status (85%) and previous studies (Purvis *et al.* 2000a; Cardillo *et al.* 2005, 2008; Davidson *et al.* 2009) provide 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 Data Deficient species. I 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, I address the following questions:

- i)* 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?
- ii)* How accurately can those methods predict current geographical patterns of extinction risk?
- iii)* Using the models obtained, what is the predicted level of extinction risk faced by Data Deficient species?
- iv)* How do my findings change current geographical patterns of extinction risk for terrestrial mammals?

Methods

Data

I 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) (Table 4.1). For each species, I 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. Since some ML methods require complete data, missing data was either phylogenetically imputed (Bruggeman *et al.* 2009; Fritz *et al.* 2009), or assigned the genus or family median (mode for categorical variables) for species missing from the phylogeny. I 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 set 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, I recorded biogeographical distribution (IUCN 2010), External Threat Index (Cardillo *et al.* 2004) and habitat suitability (Rondinini *et al.* 2011a) for each species. All geographical variables were 100% complete for each species. See Table S4.1 for details on explanatory variables.

Previous studies have reached inconsistent conclusions about the primary traits explaining variation in extinction risk among species (Cardillo & Meijaard 2012). Uninformative explanatory variables are unlikely to affect predictive performance in problems with fewer variables than species (Webb 2002; Kuhn 2008). I therefore do not undertake variable selection, but instead focus on using all available traits implicated in determining extinction risk to make the best predictions.

Table 4.1 Characteristics of the datasets used to model extinction risk in mammals.

Dataset	Number of data-sufficient species	Percentage 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

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 (Table 4.2). I also computed decision stumps using 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. I 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 I included taxonomic order, family and genus in all models to partially account for shared evolutionary history. For each taxonomic dataset, I removed highly correlated ($r > 0.9$) and low variance variables, which can lead to collinearity and zero variance in cross-validation partitions. All numeric predictors were centred to a mean of zero and scaled to a standard deviation of one before analysis (Kuhn 2008).

Table 4.2 Characteristics of different machine learning methods, adapted from Hastie *et al.* (2009) and Kampichler *et al.* (2010). Trees include decision stumps, classification trees, random forests and boosted trees. My study only found a significant difference in predictive performance between decision stumps and other methods. Key: +: good, =: fair, -: poor.

	Trees	Neural Networks	Support Vector Machines	K-Nearest Neighbours
Handling of multinomial 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	+	-	-	-

I set aside Data Deficient species and, within each taxonomic group, randomly divided the remaining species into a 25% validation set and 75% training set to assess the performance of different ML methods. For each ML method, I 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 (AUC), 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 AUC values of various tuned models on the 25% validation set. In all models, I identified a probability threshold above which species are identified as threatened by maximizing the Youden index ($Y = \text{sensitivity} + \text{specificity} - 1$; Youden 1950). The Youden index effectively lends equal weight to detecting threatened and non-threatened species whilst accounting for class imbalance (Youden 1950; Perkins & Schisterman 2006), a reasonable attitude given the importance of accurately identifying threatened species (IUCN 2001).

Multiple classification performance measures are commonly used among different research fields, reflecting varying attitudes towards misclassification costs (Hand 2012). To investigate the role of performance measure on my results, I 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). The prior distribution of misclassification costs is a beta distribution taking its mode at the same cost

as the Youden index (Hand 2012). Assessing model performance with the H measure did not qualitatively affect the results, and I present those in Tables S4.5-4.7. All analyses were conducted in R version 2.14.1, using the *caret* package (Kuhn 2008) to optimize model parameters. For further details on the methods see Appendix III.

Spatial analysis of predictions

I assessed the ability of the best global ML model to predict known patterns of extinction risk. Using species' range maps (IUCN 2010), I 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. I 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 (Lee & Jetz 2011). I also fitted simultaneous autoregressive models to account for spatial autocorrelation (Figure S4.2). I produced maps in ArcGIS 9.3 and conducted all analyses in R version 2.12.0 (R Development Core Team 2010).

Predictions for Data Deficient species

I predicted the status of 493 Data Deficient species from the best performing global model using the same threshold as for the validation dataset, and tabulated the number of Data Deficient species predicted to be threatened and non-threatened in 6,593 hexagons. I then compared the proportion of threatened species in cells with and without incorporating the predictions for Data Deficient species. Finally, I 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 models and taxonomic levels

Area under receiver operator characteristic curve (AUC) for best models ranged between 0.873 and 0.961 (Table 4.3), 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 (Table S4.3). 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, 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 AUC than predictions from the order-specific models (Table S4.3). Near Threatened species showed lowest classification accuracy, with 66% of Near Threatened species correctly classified as non-threatened (Table 4.4). Classification accuracy was homogeneous among other Red List categories (87-98%), and among threat types (94-100%). Threatened species with ranges larger than a million km^2 were less likely to be correctly classified (87%); conversely, non-threatened species with very small ranges ($<20,000 \text{ km}^2$) were less likely to be correctly classified (74%).

Table 4.3 Area under the receiver operator characteristic curve (AUC) for each combination of tool and dataset on the validation sets. CT: classification tree, RF: random forests, BT: boosted trees, KNN: k-nearest neighbours, SVM: support vector machines, NNET: neural networks, DS: decision stump.

	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

Table 4.4 Proportion of species in the validation set correctly identified as threatened or non-threatened by the best machine learning model. Species subdivided according to Red List categories, threat type and range size. Threat type was obtained from the IUCN global mammal assessment (Schipper *et al.* 2008). The validation set contains 991 species.

Criterion	Proportion correctly identified	Number of species
IUCN Red List categories		
CR	0.96	28
EN	0.98	97
VU	0.87	94
NT	0.66	67
LC	0.91	705
Threat type		
Habitat loss	0.94	186
Invasive species	1	36
Utilisation	0.94	88
Threatened species by range size (km ²)		
0-20,000	0.97	75
20,000-100,000	0.90	39
100,000-1,000,000	0.95	60
1,000,000+	0.87	45
Non-threatened species by range size (km ²)		
0-20,000	0.74	97
20,000-100,000	0.87	108
100,000-1,000,000	0.93	287
1,000,000+	0.9	280

Spatial analysis of predictions

Observed and predicted proportions of threatened species in assemblages of the validation set were broadly consistent (Figure 4.1), indicating that ML tools can correctly predict macroecological patterns of extinction risk. In both ordinary least squares (OLS) and spatial regression (SAR) models, I found a strong positive association between predicted assemblage threat and 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, $p\text{-value} < 0.001$, $t_{1,4501} = 3.865$; SAR: slope=0.096, $p\text{-value} < 0.0001$, $t_{1,4499} = 5.448$), with model fit improving with larger assemblage size (Figure S4.3). 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%).

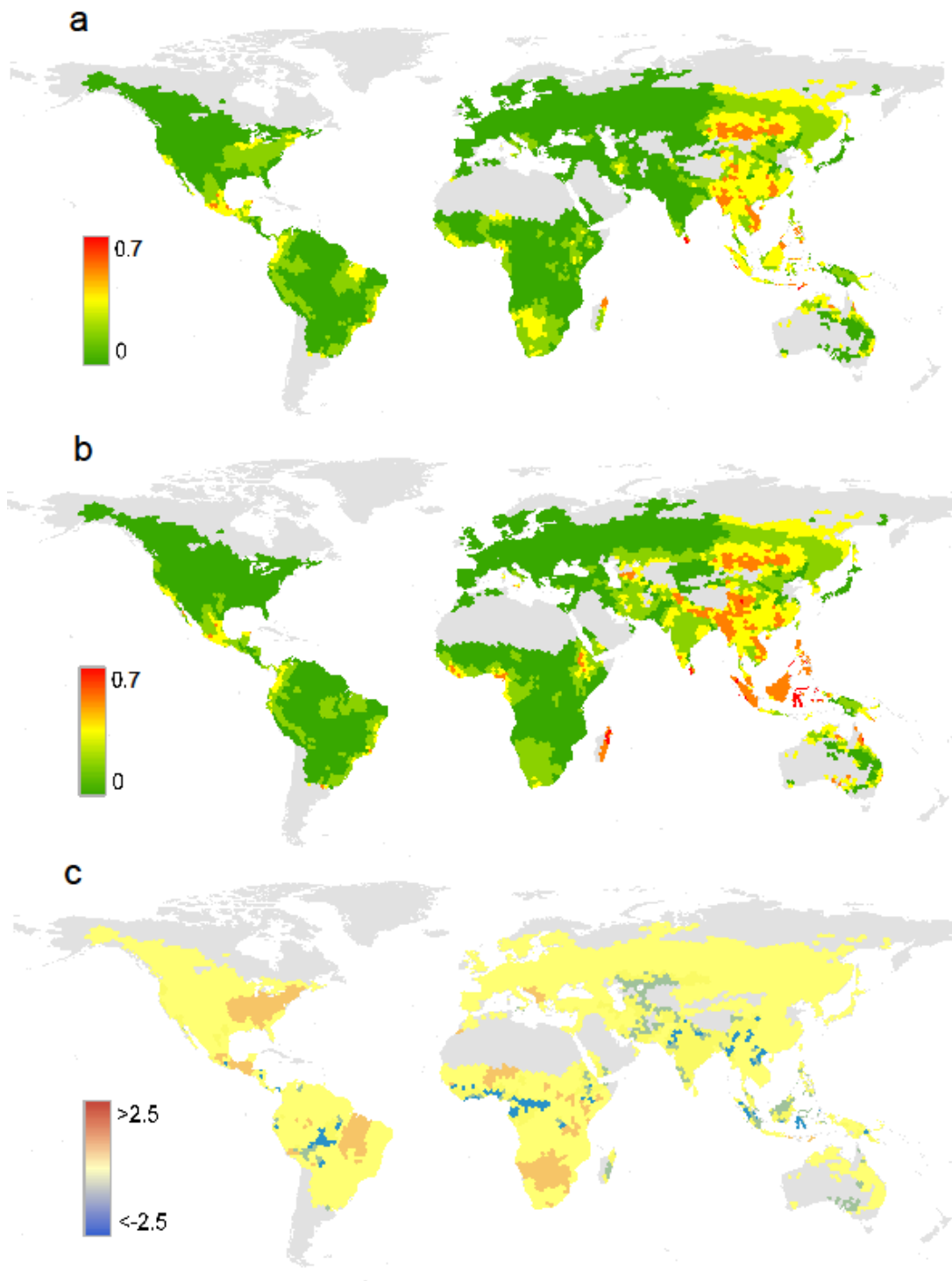


Figure 4.1 Global distribution of the proportion of threatened terrestrial mammals in the validation set. Proportion observed (a) and predicted (b) from the best machine learning model. Standardized residuals (c) display the observed-predicted difference scaled to a standard deviation of one. The validation set contains 991 species.

Predictions for Data Deficient species

Model outputs predict 313 of 493 Data Deficient species to be threatened with extinction, implying that underlying risk levels are much greater in Data Deficient species (63.5%) than data-sufficient species (22.1%). The spatial congruence between threat hotspots identified using only data-sufficient species and hotspots incorporating the Data Deficient species predictions was very high (spatial rank correlation= 0.987, $p < 0.001$; Figures 4.2 and 4.3). Additionally, the levels of threat in centres of threatened species richness may previously have been underestimated according to the regression model of observed vs. predicted threat (testing for slope \neq 1: OLS: slope=1.036, $p < 0.0001$, $F_{1,6591}=242.96$; SAR: slope= 1.043, $p < 0.0001$, $\chi^2_{1,6589}=214.15$).

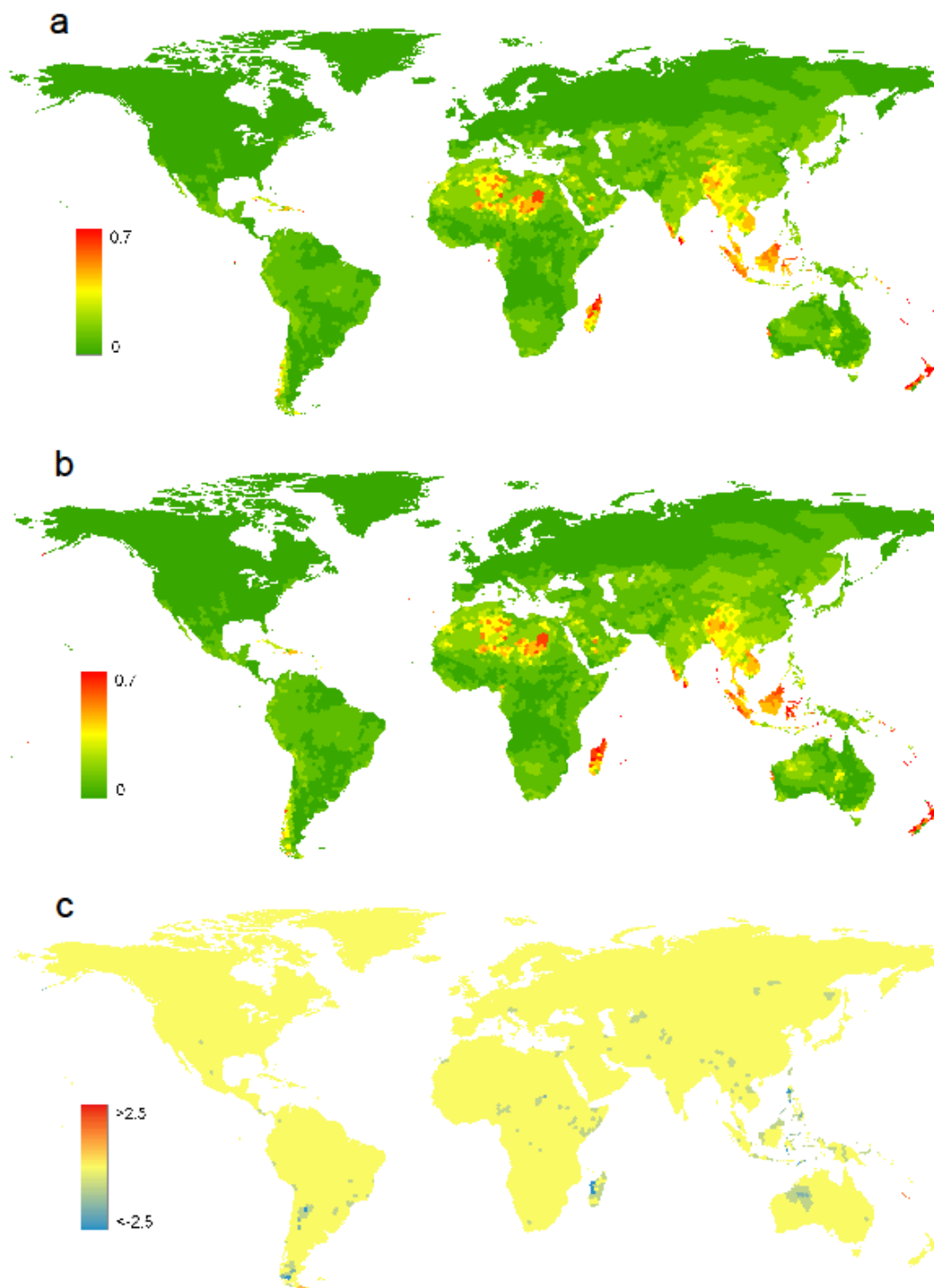


Figure 4.2 Global distribution of the proportion of threatened species for all terrestrial mammals. Proportion of threatened species when Data Deficient species are excluded from calculations (assumed as equally threatened as data-sufficient species) (a), and when Data Deficient species model predictions are included (b). Standardized residuals (c) display the observed-predicted difference scaled to a standard deviation of one. Distribution based on 4,461 species.

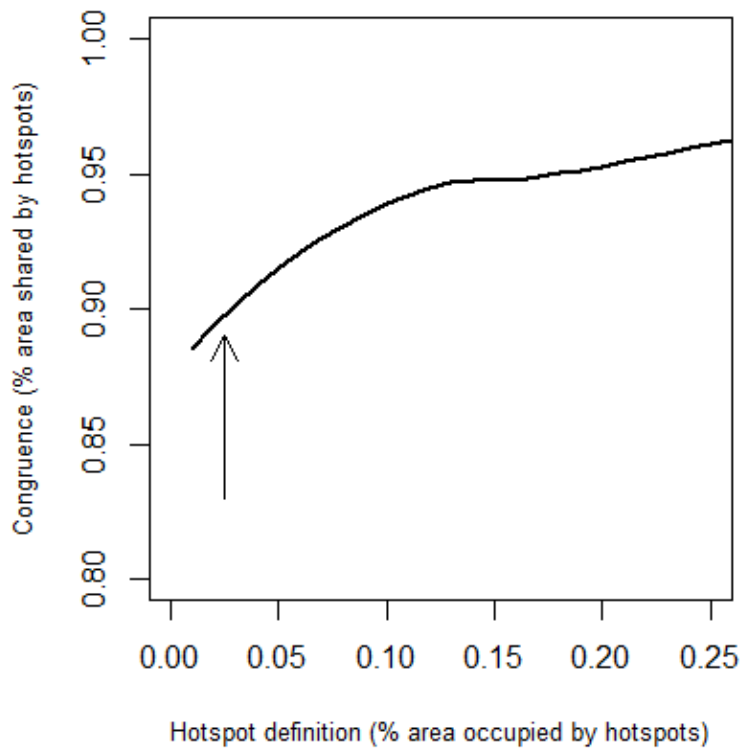


Figure 4.3 Extent of congruence in hotspots of proportion of threatened species between two scenarios, shown across a range of hotspot definitions. The two scenarios are: 1) exclusion of Data Deficient species (assuming Data Deficient species as equally threatened as data-sufficient species) and 2) inclusion of Data Deficient species model predictions. A value of one indicates perfect congruence; the vertical arrow indicates the 2.5% hotspot definition.

Discussion

We have little 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 (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

I predict 313 of 493 (63.5%) Data Deficient species are threatened with extinction. A recently published prediction of species extinction risk using eigenvector methods predicted 35% of 481 Data Deficient 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). A previous random forests model (Davidson *et al.* 2009) predicted only 28 of 341 (8.2%) Data Deficient terrestrial mammals to be at risk, perhaps reflecting the low sensitivity of the model to the detection of threatened species (sensitivity of 47.7% compared to 93.5% in my best model; Table S4.3). My 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. My 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 show considerably higher levels of species risk than previously acknowledged. Additionally, areas containing Data Deficient 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 Data Deficient 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.

My results suggest that Data Deficient species are of great conservation concern. Data Deficient species have smaller ranges (median=9,891 km²) than their data-sufficient counterparts (median= 1,666,107 km²), which contributes to their high extinction risk. Maps of Data Deficient species ranges may be uncertain and underestimated when collection effort is low. Nonetheless, the data suggest that many Data Deficient species are likely to be range-restricted and that geographical measures derived from the species' range maps are broadly

representative of the species' environment. I make the best use of the information available for each species, and note that risk predictions for individual Data Deficient species should be interpreted in the context of their IUCN Red List documentation. Since 2008, two Data Deficient 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 my model. These cases, along with the high consistency between predicted probability of threat and Red List category in the validation set (Figure S4.4), indicate that Data Deficient species which are assigned a high probability of threat are likely to be at imminent risk of extinction.

Worryingly, nearly 40% of Data Deficient species are only known from few specimens, old records or from unknown provenance (Figure 2.4), 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 Data Deficient 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 of time remain extant (particularly those that are only known from type specimens), but show considerably higher levels of threat than other species (Scheffers *et al.* 2011). We may therefore expect very poorly-known Data Deficient species to be extant, but on the brink of extinction.

Ninety-one species listed as Data Deficient 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. I predict 53 out of 90 species (59%) listed as Data Deficient 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 Data Deficient on the 2008 Red List. Hence, I expect threatened Data Deficient 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 Data Deficient species for field surveys and re-assessment. Collection of life-history and distribution information is especially urgent for the 184 Data Deficient species excluded from the analysis due to insufficient data.

Comparison of Machine Learning models 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; I 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). 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. I 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. Geographical range size can provide misleading information on conservation status: my model was less likely to assign narrow-ranging non-threatened species and wide-ranging threatened species to their correct status (Table 4.4). 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), my 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 my 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. The study also lacked a phylogenetic framework, although I took into account taxonomy in the 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 (Table S4.3), 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. (2000a) identified population density as a significant predictor of elevated extinction risk in primates, but I was unable to use this variable due to its poor coverage among 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.* 2011b), 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 Red List assessments. Three of the 15 species incorrectly classified as non-threatened by my model (*Proechimys roberti*, *Reithrodontomys microdon* and *Scotoonycteris 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 (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 Data Deficient 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 my study leaves global mammalian conservation priorities generally unaffected, I conclude risk levels in terrestrial mammals are likely to have been considerably underestimated. Predicting the conservation status of Data

Deficient species can reduce uncertainty in global patterns of risk, and enable the transparent prioritization for field surveys of potentially threatened Data Deficient species. Such an approach could be particularly cost-effective for taxa containing large numbers of Data Deficient species, such as invertebrates (Samways & Böhm 2010). Finally, Data Deficient species may be indicative of spatial knowledge deficiency and could inform species inventories. Taking into account information on Data Deficient species may therefore help tackle data gaps in biodiversity indicators, as well as conserve the earth's poorly-known biodiversity.

Chapter 5. Cost-effective assessment of extinction risk with limited information

A version of this chapter is submitted to *Ecology Letters*.

Introduction

Global indicators of biodiversity change are central to monitoring progress towards the 2020 Aichi biodiversity targets, and assessing the success of conservation actions globally. Resources for conservation are orders of magnitude below what is needed to reverse declines in biodiversity (McCarthy *et al.* 2012), so biodiversity monitoring should cost-effectively inform conservation decisions (McDonald-Madden *et al.* 2010). Representativeness and reliability have been identified as desirable properties of indicators (Dobson 2005; Jones *et al.* 2011), but the costs of achieving these are not well understood. Developing reliable biodiversity indicators with limited funds is therefore a pressing challenge for conservation science.

The taxonomic coverage of the IUCN Red List has improved in recent years (Collen & Bailie 2010; Böhm *et al.* 2013), with more than 70,000 species assessed as of 2013 (IUCN 2013a). However, one in six species on the Red List are too poorly-known to assign to a category of extinction risk, and are then listed as Data Deficient (DD). This gap in knowledge contributes to considerable uncertainty in global patterns of extinction risk (Figure 2.1 and Bland *et al.* 2012) and subsequent conservation prioritization (Trindade-Filho *et al.* 2012). The prevalence of Data Deficient species is particularly high in recently assessed invertebrate groups (e.g. 49% of freshwater crabs; Cumberlidge *et al.* 2009), hindering IUCN's efforts to broaden the coverage of biodiversity assessments (Collen *et al.* 2009). Re-assessment of the 10,673 species currently listed as Data Deficient to data-sufficient categories will require considerable resources, given the costs of biodiversity surveys (Balmford & Gaston 1999) and Red List assessments (Stuart *et al.* 2010). Determining the most cost-effective strategy to reduce uncertainty on the IUCN Red List is therefore a crucial task.

Comparative studies of extinction risk based on species trait data have previously yielded insight into the determinants of risk among groups (Purvis *et al.* 2000a; Cardillo *et al.* 2008;

Cooper *et al.* 2008; Lee & Jetz 2011), so trait data alone could underpin a preliminary re-assessment of Data Deficient species (Davidson *et al.* 2009; Safi & Pettorelli 2010; Jones & Safi 2011). Good coverage of species' trait data is available for a large number of Data Deficient species and includes life-history, ecological and phylogenetic information (Chapter 4). The geographic distribution of many Data Deficient 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 potentially be used to inform global estimates of extinction risk. Recently developed Machine Learning models of extinction risk based on species trait data have shown excellent predictive performance, and have been used to predict the likely status of Data Deficient terrestrial mammals (Chapter 4). Such models may be cheaper to apply than collecting field-based data to update Red List assessments; yet model predictions may be inaccurate, biasing estimated extinction risk levels in a group.

Given the importance of reducing uncertainty in global biodiversity indicators, and the trade-off between the cost of a monitoring method and its reliability (McDonald-Madden *et al.* 2010), how can we cost-effectively estimate extinction risk levels in Data Deficient species? Should we determine species extinction risk with field surveys and updated Red List assessments, predictive models or a combination of the two approaches? I use sampling theory to answer these questions. Specifically, I compare the variance in estimates of risk prevalence using two methods:

- i) Single sampling.* The proportion of Data Deficient species at risk of extinction is inferred from surveying and updating Red List assessments for a random subset of Data Deficient species. The available financial resources determine the size of the subset, hence the variance in the estimated prevalence of threat in Data Deficient species.
- ii) Double sampling.* The same financial resources are shared between predictive models of extinction risk based on biological trait data and a smaller set of updated Red List assessments. Given the relative costs of these two procedures and expected accuracy of a model classification, double sampling theory (Tenenbein 1970) identifies both the optimal allocation of funds to each process and the resulting variance in estimated prevalence. If predictive model development is sufficiently cheap and accurate, double sampling can give more precise estimates of threat prevalence than single sampling (Tenenbein 1970). On the other hand, if models are expensive and inaccurate predictors of extinction risk, it may be more cost-effective to only collect field data and update Red List assessments. Double sampling theory is frequently used in medical diagnostics (Baker 1991; Zhou *et al.*

2002) and quality control (Poduri 2005), but few ecological applications exist (Bart & Earnst 2002; Harper *et al.* 2004; Rayner *et al.* 2011).

IUCN Red List assessments may not perfectly reflect the actual risk of extinction of a species due to their coarse resolution (Butchart *et al.* 2005) and potential errors in assessments (Butchart *et al.* 2004). However, I am seeking to replicate Red List assessments with predictive models to ensure compatibility with the thousands of available assessments, so I am not concerned with possible errors in the Red List. In the context of this paper, species can be assessed as threatened or non-threatened with Red List assessments or with predictive models of extinction risk, as defined within the Red List categories (threatened: VU, EN and CR. Non-threatened: NT and LC; IUCN 2001).

I use four taxonomic groups with varying levels of data deficiency as case studies: terrestrial mammals ($n = 4,997$; 22.1% DD), amphibians ($n = 4,449$; 41.7% DD), reptiles ($n = 1,500$; 20.1% DD) and crayfish ($n = 586$; 31.3% DD). For each group, I calibrate Machine Learning models of extinction risk on species of known conservation status, and assess their reliability compared to Red List assessments. I compute the costs of field data collection and updating Red List assessments, and compare them with the costs of developing predictive models of extinction risk. I then devise the most cost-effective strategy for determining the extinction risk of Data Deficient species for each group.

Methods

Double sampling

I follow Tenenbein (1970) to estimate the proportion of threatened species (p) using double sampling theory. I give details for maximizing precision for a fixed cost, in which case I compare two estimates of variance:

$$V_s = \frac{pq}{n_s} \quad \text{Equation 1}$$

$$V_d = \frac{pq}{n_d}(1 - K) + \frac{pq}{N}K \quad \text{Equation 2}$$

First, the variance under single sampling (V_s) is simply the binomial variance: I conduct a small set of expensive assessments of size n_s and find the proportion of threatened (p) and non-threatened species ($q = 1 - p$). Second, for the variance under double sampling (V_d), I share the cost between cheap modelling for a larger set of species (N) and assessments for a small subset of modelled species (n_d) and again find the proportions of threatened and non-

threatened species. Note that $n_d < n_s < N$: by modelling some species, I cannot afford to assess as many species.

The comparison of these two variances hinges on K , the coefficient of reliability of the model, which lies in the range $[0, 1]$. If the model is perfect ($K = 1$), then $V_d = (pq)/N$, and since $n_s < N$, I gain a more precise estimate of p than from single sampling. If the model is useless ($K = 0$), I only have $V_d = (pq)/n_d$, and since $n_d < n_s$, I have a less precise estimate of p . For intermediate values of K , V_d is an average weighted by these two extremes.

In order to use this approach in practice, I need to know three things:

i) The coefficient of reliability of the model (K). Below, I estimated K from machine learning predictions of the conservation status of data-sufficient species, based on species trait data. I assumed that these models are similarly reliable for Data Deficient species. Where no previous assessments are available for estimating K , it would be necessary to conduct a small pilot study of assessments and modelling.

The calculation of K makes use of key values calculated from a confusion matrix: the assessed proportions of threatened (p) and non-threatened species (q), the model misclassification probabilities for threatened (ϕ) and non-threatened (ϑ) species, and the modelled proportion of threatened species (π). From these values, Tenenbein (1970) derives:

$$K = \frac{pq(1-\theta-\phi)^2}{\pi(1-\pi)} \quad \text{Equation 3}$$

The example below shows a confusion matrix for the classification of 202 species by assessment (rows) and predictive modelling (columns):

	T	NT				
T	120	12	132	$p(1-\phi)$	$p\phi$	p
NT	5	65	70	$q\vartheta$	$q(1-\vartheta)$	q
	125	77	202	1- π	π	

From this I calculate: $p = 132/202 = 0.65$, $q = 70/202=0.35$, $\phi = 12/132 = 0.09$, $\vartheta = 5/70 = 0.07$, $\pi = 125/202 = 0.62$ and hence $K = 0.67$.

ii) The costs of risk assessments (c_1) and modelling (c_2) per species, and their cost ratio ($R=c_1/c_2$). Below, I estimated these values from the cost of previous assessments and the combined costs of collating life history databases and performing modelling.

iii) The sampling ratio (f_o), giving an optimal division of costs between modelling and assessment ($n_d = N^*f_o$) that minimizes the variance V_d . This is derived by Tenenbein (1970) as:

$$f_o = \min \left[\sqrt{\frac{1-K}{KR}}, 1 \right] \quad \text{Equation 4}$$

If f_o is close to 1, it is unlikely that double sampling will be cost effective since nearly all modelled species must also be assessed, but if $f_o < 1$ then double sampling may generate more precise estimates for the same cost. A crucial metric is the proportional reduction in cost achieved by double sampling (λ):

$$\lambda = 1 - \frac{\left(R + \frac{1}{f_o}\right)(1-K-Kf_o)}{R} \quad \text{Equation 5}$$

The threshold $\lambda > 0$ (Figure 5.1) gives the region in which double sampling is a cost effective alternative to assessment alone.

Estimating the coefficient of reliability K

I developed predictive models of extinction risk for four taxonomic groups: terrestrial mammals (hereafter, mammals), amphibians, reptiles and crayfish. These datasets vary markedly in size, ratio of data-sufficient to Data Deficient species, and availability species data (Table 5.1). I defined data-sufficient species as either threatened (CR, EN or VU) or non-threatened (NT or LC). For each group, I predicted the conservation status of data-sufficient species for which trait data were available. I included taxonomic order, family and genus in all models to partially account for shared evolutionary history.

Table 5.1 Description of IUCN Red List assessments and predictive models of extinction risk for terrestrial mammals, amphibians, reptiles and crayfish. Data-sufficient species are listed as Least Concern, Near Threatened, Vulnerable, Endangered or Critically Endangered on the Red List. *: Sampled Red List of 1,500 randomly selected reptiles.

	Number of data-sufficient species	Number of Data Deficient species	Percentage of threatened data-sufficient species	Number of data-sufficient species in models	Number of predictors of extinction risk	Number of models of extinction risk
Mammals	4,300	677	22.1	3,967	35	7
Amphibians	4,449	1,294	42	478	15	4
Reptiles*	1,199	301	20.1	982	29	7
Crayfish	467	125	31.3	440	24	4

Species data included life-history, ecological, environmental and threat exposure information. Species data varied among groups, due to differences in variable measurement, variable availability, and relevance of variables to extinction risk prediction. Datasets are comparable in the sense that they use the best macroecological data available to date to predict extinction risk. I used the mammal dataset from Chapter 4, and collated a similar amphibian dataset based on Bielby *et al.* (2008) and Cooper *et al.* (2008)(Table S5.1). For reptiles, I collated the following life-history and ecological traits: maximum snout-vent length, reproductive mode, trophic level, habitat type, and number of IUCN-listed habitats (Böhm *et al.* 2013). For crayfish, I collected: maximum carapace length, habitat type, and number of IUCN-listed habitats (IUCN 2010) (Table S5.1). Using mean values from within species' geographic ranges, I also compiled spatial data on both species' environmental niche and threat exposure with ArcGIS 9.2 as follows:

i) Niche. For both reptiles (Böhm *et al.* 2013) and crayfish (IUCN 2010), I extracted: temperature, temperature seasonality, precipitation, precipitation seasonality, minimum elevation, and elevation range (all from Hijmans *et al.* 2005). I also extracted the latitude of the range centroid and extent of occurrence.

ii) Threat exposure. For reptiles, I extracted: Human Footprint (CIESIN 2005b), mean and minimum human population density for the year 2000 (CIESIN 2005a). For crayfish, I extracted: water consumption, wetland disconnectivity, river fragmentation, mercury deposition, pesticide loading and sediment loading (all from Vorosmarty *et al.* 2010).

Machine Learning (ML) tools are increasingly used in ecology for statistical pattern recognition (De'ath & Fabricius 2000; Prasad *et al.* 2006; Cutler *et al.* 2007; Olden *et al.* 2008). A wide range of ML algorithms are available, and their predictive performance

depends on the study objectives and available data (Duda *et al.* 2001; Hastie *et al.* 2009). For mammals and reptiles, I trained classification trees, boosted trees, random forest, k-nearest neighbours, support vector machines and neural networks as in Chapter 4. For amphibians and crayfish, I trained classification trees, random forests and boosted trees. I did not train additional ML techniques for amphibians and crayfish as the necessary data pre-processing (encoding of multi-level categorical variables as dummy variables) reduced model predictive performance (Table S5.2). For all groups, I trained decision stumps based on geographical range size alone to assess the power of geographical range size in predicting threatened status (IUCN criterion B). Range boundaries may be more uncertain for Data Deficient species than data-sufficient species. To assess the influence of uncertainty in range size on model predictions, I coarsened species range sizes by rounding log-transformed range sizes to the nearest higher integer (e.g. 1 = 0 to 1 km², 8 = 10,000,000 to 100,000,000 km²). I then recalibrated all models of extinction risk.

I partitioned data-sufficient species into a training set comprising 75% of species and a validation set comprising 25% of species. The validation set was set aside for comparison of different model types. For each ML tool and dataset in turn, I optimized tuning parameters using ten-fold cross-validation on the training set. For each combination of tuning parameters, I measured area under the receiver operating characteristic curve (AUC) in the cross-validation test folds. AUC provides a tool for model selection that is insensitive to class imbalance and does not require the specification of misclassification costs (Fawcett 2006). ML tools were compared independently on the validation sets previously set aside. As predictions of risk were probabilistic, predicting the risk category of a species required a threshold of predicted risk above which a species should be classified as threatened. For each trained model I calculated the reliability coefficient K among all possible thresholds and selected the threshold maximizing K .

Estimating the cost ratio R

For each taxon I calculated the cost of risk assessments (c_1) and the cost of predictive models (c_2), expressed in US dollars (\$) per species.

i) Assessment costs (c_1). The cost of a risk assessment includes the cost of collecting information to a level suitable for the application of IUCN Red List criteria, and re-assessment by the IUCN. I only accounted for field survey costs and not putative costs of resolving taxonomic uncertainty, if any taxonomic issues were present. A species can be classified by the IUCN as threatened according to five criteria (IUCN 2001; Mace *et al.* 2008): population size (criteria C and D), population size reduction (criteria A and C),

geographical range size (criterion B) and quantitative analysis (criterion E). Collecting sufficient data for poorly-known species to estimate population size or conduct quantitative analyses will be difficult, considering the short timeframe relevant to most global conservation targets (e.g. Aichi Targets: Convention on Biological Diversity 2010). Therefore, I focused on criterion B (range size), which can be predominantly investigated with presence/absence surveys. I computed three survey costs for mammals, as these vary markedly with geographical range size hence survey effort. I computed one survey cost for amphibians, reptiles and crayfish as survey effort is less variable among species (amphibians: J. Rowley, *pers. comm.*; S. Loader, *pers. comm.*; reptiles: M. Martins, *pers. comm.*; crayfish: Z. Loughman, *pers. comm.*). I estimated costs for a range of species varying in life-history, conspicuousness and remoteness of geographical location through consultation with experts from the IUCN/SCC Specialist Groups and a range of funding bodies for threatened species research (Table S5.5). I derived IUCN Red List assessments costs from published sources (Stuart *et al.* 2010) and consultation with IUCN assessors (mammals: B. Collen, amphibians: A. Angulo, reptiles: M. Böhm, crayfish: N. Richman)(Table S5.4).

ii) Predictive model costs (c_2). Predictive model building involves the following stages: collection of species trait data, GIS extractions of species range maps, data cleaning, and Machine Learning model calibration and interpretation. I computed the project and staff costs of collecting life-history traits from database compilers for mammals (Jones *et al.* 2009; Chapter 4), amphibians (Bielby *et al.* 2008), reptiles (M. Böhm, *pers. comm.*), and crayfish (this study). The cost of the four mammal life-history variables included in the analysis is uncertain: the true cost of collecting four out of 44 variables in the panTHERIA database is likely to be more than 9% of the total cost of the database, as the cost of collecting additional data is likely to diminish after a certain number of variables. I therefore computed three costs of mammal trait data: 1. Cost per species of only the four panTHERIA variables used in the analysis; 2. Cost per species of all panTHERIA variables; and 3. Cost per species of mammal trait data identical to cost of crayfish trait data, which I personally collected. I computed the cost of GIS extraction per species by accounting for the processing time of each map and the staff costs of a postgraduate research assistant at a standard UK university rate. For amphibians, I combined the costs of life-history trait collection and GIS extraction. I computed the cost of data cleaning, Machine Learning model calibration and interpretation based on the recorded task time and the staff costs of a postdoctoral researcher. Details of costs for both risk assessments and predictive models are available in Table S5.4.

Results

Estimating the coefficient of reliability K

Machine Learning tools achieved very high classification performance in mammals, amphibians, reptiles and crayfish as measured by AUC (Table 5.2). Values of the coefficient of reliability K ranged between 0 and 0.7 among models and taxa (Table 5.2), where 1 indicates perfect congruence between predictive models and IUCN assessments. The best models were random forests in mammals ($K = 0.7$) and crayfish ($K = 0.555$), boosted trees in amphibians ($K = 0.629$), and neural networks in reptiles ($K = 0.485$). Models calibrated on a coarse measure of range size achieved lower maximum K values (mammals: 0.497; amphibians: 0.587; reptiles: 0.364; crayfish: 0.467) than models calibrated on raw range size (Table 5.2). Decision stumps based on geographical range size alone achieved lowest K values in all taxa (mammals: 0.32; amphibians: 0.467; reptiles: 0.248; crayfish: 0.157).

Table 5.2 Model performances among predictive models and taxonomic groups, for (a) models calibrated on fine geographical range size, and (b) models calibrated on coarsened geographical range size. AUC: area under receiver-operator characteristic curve, cutoff: predicted probability of risk above which a species is classified as threatened, ϑ : probability of misclassification for genuinely threatened species, φ : probability of misclassification for genuinely non-threatened species, π : proportion of threatened species estimated by the model, K : coefficient of reliability of the model. The true proportion of threatened species in the sample (p) for each group is: mammals = 0.221, amphibians = 0.568, reptiles = 0.169, crayfish = 0.314.

	AUC	Cutoff	ϑ	φ	π	K
i) Fine geographical range size						
Mammals						
Classification Tree	0.895	0.3	0.102	0.233	0.249	0.406
Random Forests	0.971	0.604	0.014	0.196	0.189	0.7
Boosted Trees	0.935	0.317	0.069	0.201	0.231	0.515
Support Vector Machines	0.932	0.385	0.059	0.21	0.221	0.533
Neural Networks	0.922	0.448	0.082	0.242	0.231	0.443
K-Nearest Neighbours	0.906	0.345	0.069	0.333	0.201	0.383
Decision Stump	0.75	0.731	0.05	0.447	0.161	0.32
Amphibians						
Classification Tree	0.898	0.846	0.1	0.196	0.5	0.485
Random Forests	0.953	0.428	0.045	0.18	0.621	0.625
Boosted Trees	0.949	0.269	0.03	0.2	0.638	0.629
Decision Stump	0.842	0.731	0.136	0.18	0.569	0.467
Reptiles						
Classification Tree	0.895	0.192	0.196	0.049	0.322	0.367
Random Forests	0.916	0.354	0.107	0.219	0.22	0.369
Boosted Trees	0.928	0.164	0.147	0.073	0.277	0.426
Support Vector Machines	0.925	0.214	0.113	0.171	0.233	0.403
Neural Networks	0.943	0.283	0.108	0.097	0.24	0.485
K-Nearest Neighbours	0.894	0.255	0.117	0.268	0.22	0.308
Decision Stump	0.726	0.731	0.059	0.488	0.135	0.248
Crayfish						
Classification Tree	0.874	0.828	0.053	0.382	0.229	0.388
Random Forests	0.919	0.456	0.08	0.176	0.312	0.555
Boosted Trees	0.927	0.38	0.093	0.176	0.321	0.527
Decision Stump	0.698	0.731	0.026	0.706	0.11	0.157
ii) Coarse geographical range size						
Mammals						
Classification Tree	0.875	0.75	0.045	0.411	0.165	0.368
Random Forests	0.927	0.408	0.046	0.279	0.196	0.497
Boosted Trees	0.912	0.456	0.062	0.297	0.204	0.436
Support Vector Machines	0.915	0.394	0.058	0.301	0.199	0.441
Neural Networks	0.892	0.36	0.096	0.292	0.231	0.363
K-Nearest Neighbours	0.897	0.276	0.124	0.228	0.267	0.368
Decision Stump	0.718	0.731	0.038	0.525	0.135	0.28
Amphibians						
Classification Tree	0.9	0.286	0.12	0.12	0.551	0.571
Random Forests	0.946	0.666	0.06	0.167	0.5	0.587
Boosted Trees	0.94	0.69	0.08	0.151	0.517	0.58
Decision Stump	0.769	0.731	0.4	0.06	0.706	0.344
Reptiles						
Classification Tree	0.854	0.09	0.147	0.219	0.253	0.298
Random Forests	0.89	0.242	0.157	0.146	0.273	0.343
Boosted Trees	0.901	0.162	0.152	0.171	0.265	0.331
Support Vector Machines	0.907	0.207	0.142	0.171	0.257	0.347
Neural Networks	0.919	0.427	0.064	0.341	0.163	0.364
K-Nearest Neighbours	0.88	0.246	0.122	0.293	0.22	0.279
Decision Stump	0.5	0	0	0	0	0
Crayfish						
Classification Tree	0.823	0.727	0.12	0.323	0.294	0.322
Random Forests	0.883	0.38	0.2	0.088	0.422	0.447
Boosted Trees	0.868	0.432	0.107	0.206	0.321	0.467
Decision Stump	0.633	0.731	0.133	0.471	0.256	0.141

Estimating the cost ratio R

The cost ratios of risk assessments to predictive models were generally very high, indicating that the cost of collecting data and reassessing species was much greater than running predictive models on pre-existing data. I computed cost ratios R (c_1/c_2) of 233, 1,877 and 2,489 for mammals, contingent on the three cost estimates of life-history data. I computed cost ratios of 836 for amphibians, 1,375 in reptiles, and 1,401 for crayfish. As the cost ratios were uncertain due to the valuation of field surveys among a large number of species, I set three realistic cost ratios for the analysis: low ($R = 250$), medium ($R = 1,500$), and high ($R = 3,000$). Models based on geographical range size alone achieved very low costs relative to risk assessments (mammals: $R = 2,409,673$; amphibians: $R = 235,902$; reptiles: $R = 481,397$; crayfish: $R = 272,131$). There again, I generalized the analysis to low ($R = 250,000$) and high ($R = 2,500,000$) estimates.

Double sampling

Under the cost ratios and model performances estimated in the studied groups, it was always more cost-effective to determine the status of all Data Deficient species with predictive models and assess a small sample of species with risk assessments (double sampling), rather than spend the same resources on risk assessments alone (single sampling) (Figure 5.1). For example, I set the minimum cost ratio $R = 250$, with the cost per species of a risk assessment at US \$50,000 and the cost per species of a predictive model at US \$200. One could estimate the proportion of threatened Data Deficient mammals by sampling 100 Data Deficient species at random with risk assessments, for a total cost of \$5,000,000.

Alternatively, one could assess all 677 Data Deficient species with the best predictive model, for a cost of \$135,400. Double sampling theory shows that, to achieve precision in the estimate of risk identical to the single sampling scheme, one would assess an additional 34 species with risk assessments at a cost of \$1,700,000. Therefore the total cost of the double sampling scheme is US \$1,835,400, achieving a 63.3% reduction in cost compared to the single sampling scheme (Figure 5.1 and Appendix IV).

For the best model calibrated on a fine measure of range size, the percentage of Data Deficient species to assess with risk assessments across cost ratios ranged from 1.2 – 4.1% in mammals, 1.4 – 4.9% in amphibians, 1.9 – 6.5% in reptiles, and 1.6 – 5.7% in crayfish (Figure 5.1). This corresponds to updating IUCN Red List assessments for 8 to 28 mammals, 18 to 63 amphibians, 6 to 20 reptiles, and 2 to 7 crayfish. The number of risk assessments to update increased when models were calibrated on a coarse measure of range size, requiring the selection of 12 to 43 mammals, 21 to 68 amphibians, 8 to 25 reptiles, and 2 to 9 crayfish.

For the best model calibrated on precise data, reduction in cost achieved by double sampling across cost ratios ranged between 63 – 69% for mammals, 56 – 61% in amphibians, 42 – 47% in reptiles, and 49 – 54% in crayfish (Figure 5.1). Reduction in cost decreased when coarsening range size: among cost ratios, the best model for each group achieved a reduction in cost of 43– 48% in mammals, 52 – 59% in amphibians, 30 – 35% in reptiles, and 40 – 45% in crayfish. Reductions in cost achieved by models calibrated on range size alone were low, and approximated 32% in mammals, 47% in amphibians, 25% in reptiles and 16% in crayfish for the two cost ratios considered (Figure 5.1).

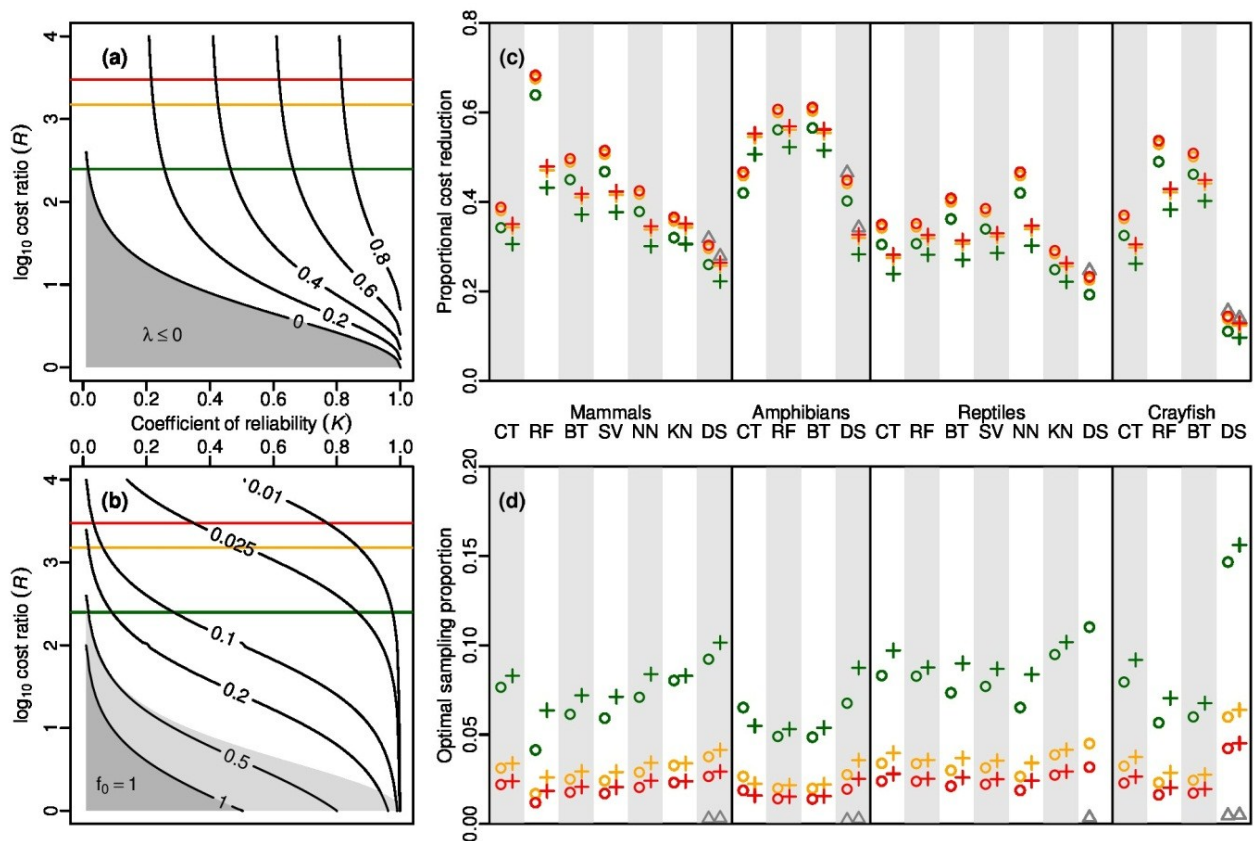


Figure 5.1 Proportional reductions in cost and optimal sampling proportion for double sampling assessments of extinction risk. (a) Proportional reduction in cost for double sampling assessments, given model reliability (K) and cost ratio (R), showing $\lambda < 0$ (light grey). (b) Optimal sampling proportion given model reliability (K) and cost ratio (R), showing $f_0 = 1$ (dark grey) and $\lambda < 0$ (light grey). (c) Proportional reduction in cost and (d) optimal sampling proportion among predictive models and taxonomic groups. Circles: models calibrated on fine geographical range size. Crosses: models calibrated on coarsened geographical range size. Grey triangles: models calibrated on range size alone with $R = 2,500,000$. Red symbols: $R = 3,000$. Yellow symbols: $R = 1,500$. Green symbols: $R = 250$. CT: classification tree. RF: random forests. BT: boosted trees. SV: support vector machines. NN: neural networks. KN: k-nearest neighbours. DS: decision stumps.

Discussion

Reliable biodiversity indicators are central to monitoring progress towards Aichi targets and the effectiveness of conservation actions globally. Yet, even the most comprehensive and authoritative indicators suffer from data gaps: one in six species on the IUCN Red List are assessed as Data Deficient (IUCN 2013a), severely limiting our understanding of patterns and trends in extinction risk. Given the limited resources available for conservation and the high costs of biodiversity surveys and assessments, the most cost-effective strategy for estimating extinction risk prevalence in Data Deficient species should be devised. Double sampling with IUCN Red List assessments and predictive models cost-effectively estimates the proportion of Data Deficient species at risk of extinction, offering a solution for resolving a substantial gap in our knowledge of biodiversity. I find that it is always more cost-effective to model the risk status of all Data Deficient species in a group and update IUCN Red List assessments for a small number of Data Deficient species (double sampling), compared to allocating all financial resources to updating IUCN Red List assessments (single sampling). This is because identical or lower variance in the estimate of the proportion of Data Deficient species at risk can be obtained from combining model predictions and Red List assessments, rather than using Red List assessments alone.

Using the best model for each group and a medium estimate of costs, 12 Data Deficient mammals, 26 amphibians, 9 reptiles and 3 crayfish should be randomly surveyed in the field and re-assessed according to IUCN criteria. Double sampling reduces the cost of determining the proportion of Data Deficient species at risk of extinction by up to 69% compared to assessing species with Red List assessments alone, as pre-existing biological data are used to minimize the number of field surveys to perform. Applying the smallest coefficient of reliability found for the best model in a group ($K = 0.485$ in reptiles), and a medium cost ratio ($R = 1,500$) to all Data Deficient species not included in this study, a double sampling scheme reduces costs on average by 50%. If estimates of species geographical range size are only available to the nearest order of magnitude, a double sampling scheme reduces costs by 38%.

Determining the conservation status of Data Deficient species is an expensive endeavour as it encompasses the cost of additional research and field surveys in addition to the cost of the risk assessment process (Stuart *et al.* 2010). Survey costs vary widely among survey types and species, and I made use of the best information available to estimate a mean cost of surveying and assessing Data Deficient species according to IUCN criteria (Table S5.5). Assuming Data Deficient species in groups not included in this study can be surveyed and

assessed for similar costs as amphibians, reptiles and crayfish (US \$25,400 per species), I estimate the total cost of surveying and risk assessments for all 10,673 Data Deficient species on the Red List (IUCN 2013a) to US \$284 million. The figure may overestimate cost as it does not reflect efficiencies in surveying multiple species simultaneously (Gardner *et al.* 2008), but highlights the cost of addressing data deficiency on the Red List, and the value of biological data accumulated over the last centuries.

The cost of increasing the number of species on the Red List to 160,000 has been estimated at US \$60 million (the Barometer of Life: Stuart *et al.* 2010; also see Collen & Bailie 2010; Knight *et al.* 2010). Many invertebrate, plant and fungi species to be included in the Barometer of Life are not well studied and may be assessed as Data Deficient. As a consequence, creating a meaningful Barometer is likely to require considerable additional investment in field surveys and natural history research. Under current funding of the Red List, more than 90% of the Barometer of Life assessments will become outdated in the next decade (Rondinini *et al.* 2013). Limited resources for tracking the status of biodiversity therefore create a trade-off between expanding the coverage of biodiversity assessments (Collen *et al.* 2009), keeping assessments up-to-date (Rondinini *et al.* 2013), and ensuring reliability and utility of risk assessments (Bland *et al.* 2012). These are the costs merely for understanding extinction risks; the cost of reducing the extinction risk of all globally threatened species was estimated at US \$3.41 to \$4.76 billion, of which only 12% is currently funded (McCarthy *et al.* 2012).

Data Deficient species receive very little conservation investment, despite their prevalence on the Red List, and their influence on global conservation prioritization (Bland *et al.* 2012) and reserve selection (Trindade-Filho *et al.* 2012). For example, less than 1% of the awards from the People's Trust for Endangered Species (People's Trust for Endangered Species 2013), and 3% of the awards from the Mohamed Bin Zayed Species Conservation Fund are directed toward Data Deficient species (Mohamed bin Zayed Species Conservation Fund 2013). I show that increased investment in desktop research and risk assessments for poorly-known species could cost-effectively reduce uncertainty in estimates of extinction risk, enabling the adequate monitoring of progress towards international biodiversity targets.

Predictive models of extinction risk are not only a cheaper option than risk assessments, but are more likely to be developed within time scales relevant to biodiversity targets. Group assessments require extensive workshops, administration and training (Rondinini *et al.* 2013) and typically take several years to complete. On the other hand, predictive models require collection of data from species descriptions, museum specimens and other natural history

resources, which can be carried out by non-experts. Whilst calibration of Machine Learning tools requires statistical expertise, accessibility could be improved by developing user-friendly platforms. Development of predictive models of extinction risk could be achieved alongside other trait-focused research on a wide range of organisms, including speciose groups of invertebrates.

Double sampling may not be cost-effective under certain conditions. With levels of congruence between predictive models and IUCN Red List assessments of $0.4 < K < 0.7$, double sampling is not cost-effective when the costs of modelling and updating Red List assessments are about equal ($R < 1.5$). Such a cost ratio is unlikely to occur given the costs of biodiversity surveys (Gardner *et al.* 2008) and risk assessments (Stuart *et al.* 2010). With poor models ($K < 0.1$), reductions in cost are small (<10% among cost ratios) so managers may decide the overhead costs of calibrating models are not worthwhile. In a nutshell, if risk assessments are at least 250 times more expensive than predictive models, and models relatively reliable ($K > 0.4$), double sampling reduces cost by 40% or more – a good rule of thumb for managers wishing to use predictive models.

Comparison of Machine Learning models and datasets

The utility of predictive models of extinction risk for conservation depends on their accuracy and cost relative to risk assessments. In this study's focal groups, models achieved very high AUC in independent validation sets (Table 5.2), indicating excellent discrimination between threatened and non-threatened species. The level of congruence between predictive models and IUCN assessments varied among ML tools and datasets, creating differences in savings with double sampling compared to single sampling (Figure 5.1). Exactly why models differ in predictive performance within and among datasets depends on the link between the ML algorithm, fitted functions and the data. The question is best investigated by simulating data (see Elith & Graham (2009) for an example in species distribution modelling), but I include information on variable importance in random forests, boosted trees and classifications trees (Figure S5.3). Most models performed better on mammals and amphibians than crayfish or reptiles (Table 5.2), likely due to the high amount of species-level data available for mammals and the high prevalence of risk in amphibians (Table 5.1). Classification trees and k-nearest neighbour models are conceptually simpler and computationally less intensive than other tools (Hastie *et al.* 2009), and achieved low coefficients of reliability in mammals and reptiles (Table 5.2). Boosted trees and random forests performed well, and seem particularly suited for extinction risk modelling on macroecological datasets, which may present non-linear relationships between trait and risk values, missing data and multi-level categorical

variables. However, I recommend testing multiple ML techniques and modelling parameters to achieve maximum classification performance (Hastie *et al.* 2009).

The biggest savings were achieved by improving model performance, whilst savings were less sensitive to the estimate of risk assessment and model costs (Figure 5.1). Approximate cost ratios may therefore be sufficiently informative when designing double sampling schemes. Improving model performance could be achieved by including more species in a group; collecting additional species traits; or improving precision of species traits included in the model. Precise estimates of range size may provide the most effective way of developing powerful models of extinction risk, as range size is explicitly incorporated in risk assessments (IUCN 2001) and consistently predicts extinction risk among groups (Cardillo & Meijaard 2012). Reliable range maps may not be available for Data Deficient species (though in a recent assessment, only 4 species of Data Deficient crayfish could not be mapped; B Collen *pers. comm.*), which may affect model predictions and cost-effectiveness. For a medium cost ratio, models calibrated on a coarse measure of range size still achieved 34 – 56% reduction in cost among groups (Figure 5.1).

Collection of species data forms the bulk of the cost of developing predictive models, particularly life-history traits (Table S5.4). Calibrating models based on range size alone could be more cost-effective if reduction in predictive performance is offset by a smaller cost of data collection. Reductions in cost achieved by models calibrated on range size alone were small, ranging from 15 to 47% among groups, depending on the predictive power of range size. Range size alone was a particularly good predictor of risk in amphibians, for which 58% of modelled threatened species were listed under IUCN Criterion B. On the other hand, threatened reptiles could not be identified with decision stumps calibrated on coarse geographical range size. All in all, complex interactions among taxon size, risk prevalence and availability of life-history and threat information determine the ability to predict extinction risk among groups. Double sampling may therefore have limited utility for determining in advance the quantity and quality of information necessary for risk predictions in new groups. Preposterior Bayesian techniques can assess the value of information in risk assessments (Sahlin *et al.* 2011), and could complement a double sampling scheme.

Limitations

The study has a number of limitations. First, I modelled binomial threat status (threatened vs. non-threatened) rather than Red List categories, due to difficulties in modelling highly imbalanced response categories with the available data (Hastie *et al.* 2009). The raw results of the ML models are probabilistic, indicating the probability of belonging to a threatened

category. A multinomial double sampling scheme (Tenenbein 1972) could investigate the cost-effectiveness of estimating the prevalence of specific categories, or varying categorisation thresholds from probabilistic results, but the approach may result in poor classification performance. Second, I assume that the relationship between predictor variables and extinction risk is similar in data-sufficient and Data Deficient species. I believe this assumption is fair and warranted given the urgency of addressing data deficiency on the Red List. Accurate predictions also require the range of trait values exhibited by Data Deficient species to be represented by modelled data-sufficient species (Figures S4.1, S5.1 and S5.2; Data Deficient data not available for amphibians).

Estimation of predictor variables may be less accurate in Data Deficient species and could affect model performance. I used the best available data and investigated the role of uncertainty in geographical range size, and show that data uncertainty can be readily incorporated into a double sampling scheme. I also assume that geographical range maps are available for all species to assess in a sample, which may not be the case for all Data Deficient species, or for species not assessed by the IUCN (e.g. species not selected in the Sampled Red List assessment of their taxonomic group). For such species, the cost of constructing a range map from occurrence records and atlases would need to be incorporated in the costs of predictive models.

A necessary assumption of double sampling theory is the availability of an error-free measurement method. Current Red List assessments may not accurately measure extinction risk, as species can change Red List status due to genuine improvements and deterioration in conservation status, and non-genuine reasons (Butchart *et al.* 2004). Species may undergo non-genuine changes in Red List status as a result of criteria revision, improved knowledge, changes in taxonomy or changes in the risk attitude of assessors (Hoffmann *et al.* 2010). For example, of the 4,828 mammals evaluated on the 1996 Red List, 2,939 species changed status on the 2008 Red List. Most changes were due to the revision of the IUCN criteria in 2001 (IUCN 2001), rather than genuine differences in conservation status (171 species) (Hoffmann *et al.* 2010). Fewer changes would be expected in the future for the groups investigated in this study, but it should be noted that the utility of models of extinction risk is contingent on the quality of IUCN Red List assessments.

Finally, the double sampling scheme relies on a binomial sampling distribution for an infinite population (Tenenbein 1970). In reality Data Deficient species represent populations of finite size, which will be more adequately modelled by a hypergeometric distribution. As the single sample size approaches the total population size, the variance in the estimated proportion of

Data Deficient species at risk decreases faster for a hypergeometric distribution than for a binomial distribution, eventually reaching zero when all species have been red listed. To date, double sampling theory has not been extended to the hypergeometric distribution, but I provide in Appendix IV estimates of variance for single sampling with a binomial distribution, single sampling with a hypergeometric distribution, and double sampling with a binomial distribution. I show that under realistic conditions ($K = 0.4$ and $R = 1,500$), double sampling with a binomial distribution performs better than single sampling with a hypergeometric distribution if red listing is financially constrained to fewer than 188 out of 500 Data Deficient species, or 376 out of 1,000 Data Deficient species. These numbers provide an upper boundary for the use of double sampling with a binomial distribution; double sampling with a hypergeometric distribution may prove cost-effective for higher budgets. Double sampling as implemented in this study will therefore yield adequate results under limited budgets, which are commonplace in conservation biology (McCarthy *et al.* 2012).

Conclusions

To measure progress towards international targets and halt the current loss biodiversity, reliable indicators of biodiversity change are needed. For the first time, I test the cost-effectiveness of reducing uncertainty in a major biodiversity indicator. I show that double sampling with predictive models cost-effectively estimates the proportion of IUCN Data Deficient species at risk of extinction, and reduces assessments costs by up to 69% compared to single sampling with IUCN Red List assessments alone. Double sampling remains cost-effective under poor data quality and availability, demonstrating the method's capacity to cheaply inform the conservation status of poorly-known groups of plants and invertebrates. Double sampling could be applied more widely in ecology and conservation to formally compare the cost-effectiveness of sampling methods differing in cost and reliability. Double sampling schemes are also available for multinomial data (Tenenbein 1972), continuous data (Gilbert 1987), and for designing pilot studies in multiple stages (Tenenbein 1971). Given the urgency of the biodiversity crisis and the limited availability of conservation funds and biological data, designing efficient monitoring schemes is imperative.

Chapter 6. Conclusions

In this thesis, I aimed to identify and resolve the effects of Data Deficient species on the estimation of global patterns and levels of extinction risk. I asked four questions:

- i)* What factors determine the availability of species conservation data?
- ii)* What is the effect of data deficiency on global patterns of extinction risk and conservation prioritization?
- iii)* Can the likely conservation status of Data Deficient species be determined?
- iv)* Can it be determined cost-effectively?

In this chapter, I review the main findings of my thesis, and highlight key challenges for resolving the effects of data gaps in biodiversity patterns. I also provide specific recommendations for the implementation of the Data Deficient category by IUCN.

Summary of research findings

Characterizing the geographic distribution of conservation data deficiency is a first step in assessing the reliability of estimates of species extinction risk, and prioritizing areas for conservation research. In Chapter 2, I assessed the cross-taxa congruence in the spatial distribution of Data Deficient species richness, and investigated the relative roles of species biology and human sampling effort in driving patterns of data deficiency. Centres of Data Deficient species richness exhibited low levels of congruence among groups. Determining the effect of uneven data availability on patterns of extinction risk will therefore require taxon-specific approaches, given the low levels of congruence observed in both patterns of data deficiency and patterns of extinction risk among groups (Grenyer *et al.* 2006; Collen *et al.* 2014). My finding also implies that conservation research actions directed towards poorly-known species may not be transferable among taxonomic groups. In addition, my study suggested that patterns of global conservation data deficiency were primarily driven by spatial patterns of ecological research. Data Deficient species shared few biological characteristics, and represented a range of data deficiencies rather than a homogenous group. I concluded that many species are only known from old or very few records, highlighting the importance of basic taxonomic and natural history research for improving conservation assessments.

In Chapter 3, I investigated the impact of Data Deficient species on taxonomic and geographical patterns of extinction risk in crayfish, freshwater crabs and dragonflies. I

evaluated three scenarios accounting for the range of uncertainties conferred by Data Deficient species: excluding Data Deficient species, treating all Data Deficient species as non-threatened, and treating all Data Deficient species as threatened. I found that global taxonomic and geographical patterns of extinction risk were generally robust to the different treatments of Data Deficient species. Data deficiency significantly affected extinction risk patterns within biogeographical realms, severely limiting our ability to prioritize freshwater invertebrates for conservation. Given current levels of data deficiency, understanding the relative importance of biological traits and threatening processes in driving extinctions is also difficult. I concluded that Data Deficient species should be given high research priority to determine their conservation status, either through field surveys or predictive modelling.

To reduce the uncertainty in estimates of extinction risk contributed by Data Deficient species, I predicted the conservation status of Data Deficient mammals from widely available life-history, environmental and threat information (Chapter 4). I found that Machine Learning models could accurately predict species conservation status and centres of threatened species richness. Support vector machines, neural networks, boosted trees and random forests performed well. I therefore recommended testing multiple Machine Learning models to achieve highest performance when predicting extinction risks in new species groups. I found Data Deficient mammals to be at high risk of extinction, increasing the estimated proportion of threatened mammals from 22 to 27% globally. Regions predicted to contain large numbers of threatened Data Deficient species were already conservation priorities, but showed considerably higher levels of risk than previously recognized. I concluded that unless directly targeted for field surveys and conservation actions, species classified as Data Deficient were likely to slide towards extinction unnoticed.

In Chapter 5 I investigated whether predictive models could cost-effectively estimate extinction risk levels in Data Deficient mammals, amphibians, reptiles and crayfish. I showed that regardless of model type used or species group examined, it was always more cost-effective to determine the conservation status of all species with predictive models and assess a small proportion of species with Red List criteria (double sampling), rather than spend the same resources on field surveys and Red List assessments alone (single sampling). Double sampling reduced assessments costs by up to 69%, and remained cost-effective under simulations of poor data quality and availability. I concluded that double sampling could be used to reduce the impact of uncertainty in the Red List and Red List Index to monitor progress towards the Aichi targets of the Convention on Biological Diversity.

Limitations of the research and future prospects

Patterns in biodiversity data collection

In Chapter 2 I assessed the relative roles of total species richness, human population density and remoteness in explaining spatial patterns of data deficiency, but their interactions remained unclear. Correlative approaches show limited ability to determine the causes of knowledge deficiency (Diniz-Filho *et al.* 2005; Vale & Jenkins 2012), so links between sampling effort and biodiversity patterns may be best addressed with process-based approaches, such as simulation studies of survey patterns (Sastre & Lobo 2009) or mechanistic models of taxonomic effort (Joppa *et al.* 2011). Developing mechanistic models capable of describing current patterns of biodiversity knowledge, and predicting future patterns under different scenarios of information collection is therefore a useful avenue for future research.

All models of biodiversity knowledge, whether correlative or mechanistic, require an accurate measure of sampling effort. In Chapter 2 I demonstrated that global measures of remoteness and human population density do not completely capture the complex processes of human exploration and biological collection. Near-exhaustive datasets on sampling effort typically cover small geographical areas of importance to particular museums (Good *et al.* 2006). Synthesizing the density of records within geographical areas can provide a coarse measure of sampling effort, but does not provide an indication of the quality or bias of sampling (Soberón *et al.* 2007). Global datasets of collection records suffer from poor representativeness and data quality (e.g. GBIF; Yesson *et al.* 2007), so their capacity to accurately reflect sampling effort must be ground-truthed. Developing a global, representative indicator of sampling effort is therefore paramount to accurately estimating biodiversity patterns.

In addition, measures of sampling effort must reflect the dynamic nature of biodiversity data collection. Species geographical range maps derived from records collected in different time periods can differ substantially (Lobo *et al.* 2007), whilst temporal patterns in species discoveries form the basis of extrapolations of the number of undiscovered species (Scheffers *et al.* 2012). As such, past patterns of data collection can inform extrapolations of future rates of data collection, and inform data collection to resolve biases in biodiversity patterns. A further 480 years may be required to describe all species on earth (May 2011), or up to 1,000 years for fungi alone (Blackwell 2011). Similarly, given current patterns of data collection, how many years will be required to obtain representative patterns of extinction risk in poorly known groups? Investigating the number and characteristics of species coming

in and out of the Data Deficient category can provide some insights into temporal patterns of data collection (Chapter 4), but comparisons are limited in many groups by the low availability of IUCN data among time periods. Reducing uncertainty in biodiversity patterns will therefore require the development of indicators of biodiversity knowledge capable of accurately reflecting the temporal and spatial processes driving biological data collection.

Predictive modelling of biodiversity patterns

Predictive modelling may be necessary to account for uncertainty in biodiversity patterns and inform conservation objectives. Chapters 4 and 5 highlight a number of challenges in developing predictive models of extinction risk.

i) Data uncertainty: First, predictive models rely on limited amounts of information representing the current state of knowledge of the system. Developing predictive models of extinction risk in mammals required phylogenetic imputation of missing life-history data, a process affected by the accuracy of the phylogeny, and the quantity and phylogenetic distribution of available trait data. For example, exploratory analyses with phyloPARS suggest that imputation based on skewed body mass data may generate biased datasets, so inferences should be made with caution (González-Suárez *et al.* 2012). Estimating the sensitivity of models of extinction risk to imputation of life-history data should therefore be the focus of future research. Some Machine Learning models cannot cope with missing data, whilst others account for missing data through different means (e.g. surrogate splits in classification trees and boosted trees vs. imputation in random forests; Hastie *et al.* 2009). As a consequence, missing data in models of risk are best avoided and the best efforts should be made to acquire near complete data. Species' range maps may also be uncertain due to omission and commission errors (Boitani *et al.* 2011; Ficetola *et al.* 2014). I investigated the effects of uncertainty in species' range maps by calibrating models of risk with a coarse measure of geographical range size (Chapter 5), but further research should focus on estimating the effects of uncertainty in characterizing species' niches and exposure to anthropogenic threats.

ii) Model uncertainty: Predictions of biodiversity patterns are sensitive to model form, model implementation and model evaluation. To account for uncertainty in model form, I calibrated seven Machine Learning model types in Chapters 4 and 5. I showed that some model types performed consistently well, but that different model types performed best in different datasets. Differences in predictive performance depend on the link between the algorithm, fitted functions and data distribution, which can be investigated by simulating data with known distributions and relationships. This approach has been applied to species

distribution modelling (Elith & Graham 2009) and could be applied to extinction risk modelling. Models calibrated in Chapter 4 and 5 focused on predictive performance rather than explanatory power, so future studies could focus on determining the role of extinction risk drivers and could undertake variable selection and model simplification. In addition, my models took into account taxonomic information rather than phylogenetic information. Efficiently incorporating phylogenetic information into Machine Learning models is therefore crucial to broadening the applicability of Machine Learning to ecology.

Model evaluation remains a challenge for predictive ecological modelling (Araújo & Guisan 2006). To date most studies of extinction risk have focused on ecological explanation and selected models with hypothesis testing (Cardillo *et al.* 2005, 2008; Cooper *et al.* 2008), or assessed classification performance on training data (Safi & Pettorelli 2010). To independently compare seven different Machine Learning model types in Chapters 4 and 5, I undertook ten-fold cross validation followed by evaluation on a validation set. The approach is conservative and prevents underestimation of the expected error rate (Hastie *et al.* 2009), but is data-intensive and may reduce predictive performance on small datasets.

A large number of model performance measures are available, representing different modelling paradigms among fields and varying attitudes towards misclassification costs (Hand 2012). I trained models based on the area under the curve (AUC), the H value and the K coefficient of reliability (Chapters 4 and 5), to select the best models to address conservation objectives and assess the effect of model performance measures on model selection. I showed that performance measures did not qualitatively impact model selection, but results may differ in further studies. Finally, selecting a threshold above which a numeric prediction is classified as positive (e.g. species classified as threatened) is problematic (Liu *et al.* 2005; Lawson *et al.* 2014). This is particularly the case when the numbers of threatened and non-threatened species are highly imbalanced, and model predictions do not constitute properly calibrated probabilities (Fawcett 2006). The most appropriate method should be determined by the risk attitude of the investigator and the objectives of the study.

Accurately specifying conservation problems and objectives

Under a decision-theoretic approach, the first step is to translate broad conservation goals into measurable objectives, to inform decisions and the allocation of conservation resources (Nicholson & Possingham 2006). Defining clear conservation objectives is recognized as a constraint in evaluating conservation success at local (Kapos *et al.* 2008) and global scales (Mace *et al.* 2010). International biodiversity targets are typically expressed in vague terms, limiting our ability to measure the effectiveness of conservation actions globally. For

example, Target 12 of the Aichi Targets states that: “by 2020 the extinction of known threatened species has been prevented and their conservation status, particularly of those most in decline, has been improved and sustained” (Convention on Biological Diversity 2010), revealing no clear quantitative goal in extinction risk reduction or the assessment of species of unknown conservation status. Determining clear objectives for extinction risk reduction is therefore of considerable importance to assessing the effectiveness of conservation actions, as the IUCN Red List promotes conservation efforts at multiple spatial scales (Sodhi *et al.* 2011).

Required levels of precision and risk attitudes towards over or under-estimation of trends also need to be specified to monitor progress towards conservation targets. Incorporating objectives into model design is difficult if the economic or conservation costs of erroneous decisions are unknown (Sahlin *et al.* 2011), hence models are often evaluated based on predictive accuracy alone (Fielding & Bell 1997). In Chapter 4, I selected the best models of risk based on equal importance of sensitivity and specificity (Youden 1950), reflecting a precautionary attitude to red listing when the prevalence of threatened species in a group is low. This is in keeping with general IUCN guidelines (IUCN 2001), although precise misclassifications costs are not currently stated for the application of IUCN criteria, or for measurement of progress towards international biodiversity targets.

Chapter 5 illustrates the potential financial benefits of linking predictive modelling and decision theory, and focusing on cost-effectiveness rather than predictive performance alone. However, the desired precision (or variance) in estimates of the proportion of threatened species among taxonomic groups was not known, so necessary budgets for the re-assessment of Data Deficient species could not be computed. Results from Chapter 5 also indicate that estimates of variance based on a hypergeometric distribution may be more appropriate for small populations of Data Deficient species and relatively large budgets. As a consequence, relatively small changes in the specification of conservation problems may lead to large changes in methodologies to address data gaps. Planning cost-effective data collection to maximize the value of information (Dakins 1999; Yokota & Thompson 2004) will therefore require the precise specification of conservation problems and objectives.

When indicators of biodiversity change are designed to address multiple conservation objectives, different sampling strategies may be necessary to resolve data gaps and inform each objective. The overarching goal of the IUCN Red List is “to provide information and analyses on the status, trends and threats to species in order to inform and catalyze action for biodiversity conservation” (IUCN 2013b). This includes two practical objectives:

quantifying global patterns and trends in extinction risk globally, and pinpointing individual species at high risk of extinction (IUCN 2013b). The first objective requires the identification of both non-threatened and threatened species to accurately estimate risk levels, as achieved in this thesis. Predictive models of extinction risk could also address the second objective, and identify high-risk Data Deficient species for preferential re-assessment to data-sufficient categories. Observed or predicted species extinction risk is only part of the information required for efficient resource allocation (Possingham *et al.* 2002), and frameworks exist for species prioritization according to evolutionary distinctiveness (Isaac *et al.* 2007), functional distinctiveness (Petchey & Gaston 2006), project costs (Joseph *et al.* 2009), and likelihood of project success (Marsh *et al.* 2007; Joseph *et al.* 2009). Such frameworks could be applied to the prioritization of Data Deficient species for field surveys and re-assessment by the IUCN.

Developing a framework to resolve the effects of data gaps in biodiversity patterns

Resolving the effects of data gaps in estimating biodiversity patterns can be achieved within a simple framework (Figure 6.1), here applied to estimating patterns and trends in extinction risk globally. The framework aims to identify and resolve biases in a biodiversity pattern in light of scientific or conservation objectives. Following the identification of systematic biases in data availability (Chapter 2), the effects of data gaps on the pattern can be investigated by sensitivity analyses (Chapter 3). Data gaps are filled in through predictive modelling, which requires the specification of an appropriate model and the identification of predictor variables (Chapter 4). The effects of predictions on the biodiversity pattern are assessed through sensitivity analyses (Chapter 5). Finally, the sensitivity analyses or predictive model can be used to identify necessary data collection through decision theory or value of information theory, and inform the main objective (Chapter 5). Decision theory or value of information theory can also identify when it is most advantageous to collect surrogate data to inform the model, rather than directly collect data underlying the pattern (Chapter 5).

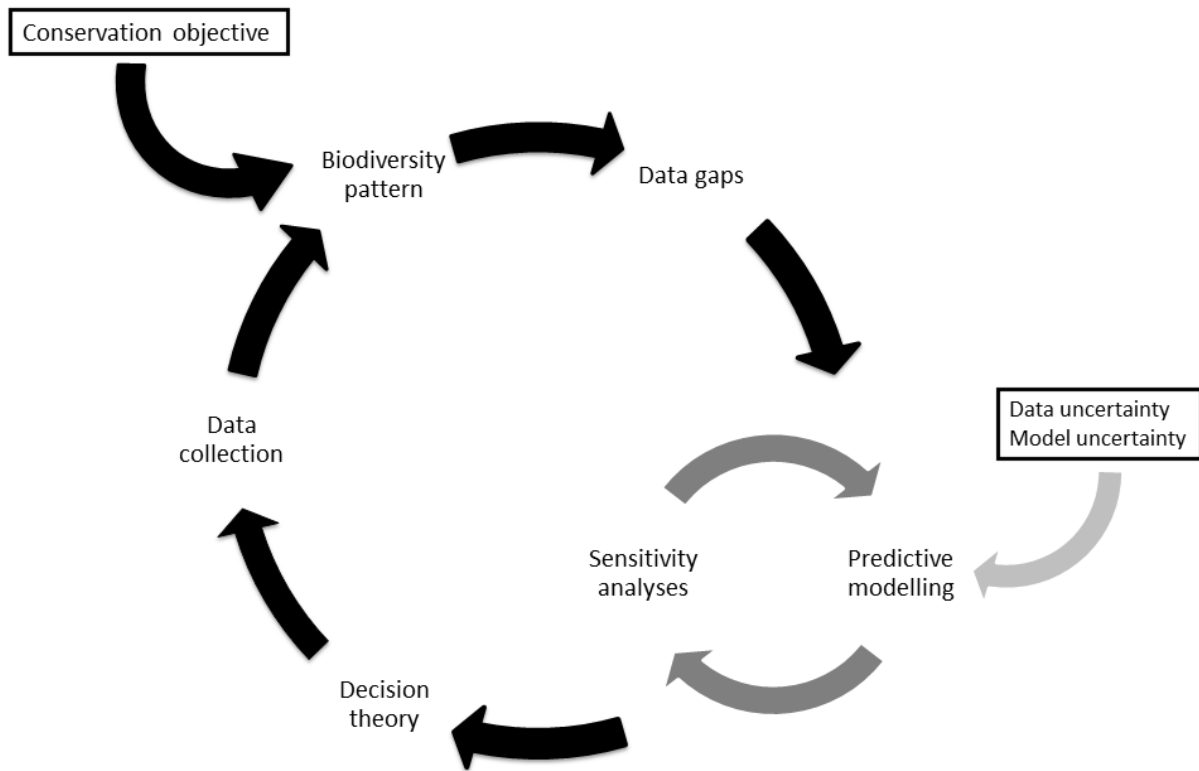


Figure 6.1 Framework to identify and resolve the effects of data gaps in estimating biodiversity patterns of conservation interest.

Such a framework could be applied to other shortfalls in biodiversity knowledge, such as limitations in our knowledge of the discovery and description of species (the Linnean shortfall; Whittaker *et al.* 2005), limitations in our knowledge of the geographical distribution of species (the Wallacean shortfall; Lomolino 2004), and limitations in our knowledge of the abundance of species in space and time (the Prestonian shortfall; Cardoso *et al.* 2011). For example, to address limitations in our knowledge of the biological and functional attributes of species (the Hutchinsonian shortfall; Mokany & Ferrier 2011), missing at random tests can be conducted to relate information deficiencies to variables of interest and identify data gaps (Nakagawa & Freckleton 2008). Modelling techniques such as phylogenetic imputation could then predict likely trait values (Bruggeman *et al.* 2009), and used to explore the sensitivity of biodiversity patterns and conservation objectives to predictions (González-Suárez *et al.* 2012). For some biodiversity patterns, predictive modelling may be conducted before the identification of data gaps and sensitivity analyses, as potential values for the biodiversity pattern are unknown. For example, to identify poorly-

sampled localities measures of inventory completeness need to be computed from predictive models of species richness (European Distributed Institute of Taxonomy 2007; Soberón *et al.* 2007).

The framework provides an extension of the indicator-policy framework (Nicholson *et al.* 2012) when applied to a biodiversity indicator, such as those developed to monitor progress towards international conservation targets (Jones *et al.* 2011). The proposed framework can identify when data gaps in the indicator may lead to erroneous conservation decisions, and break down the indicator-policy cycle.

Recommendations to IUCN for the application of the Data Deficient category

In light of the information gathered in this thesis on Data Deficient mammals, amphibians, reptiles, freshwater crabs, crayfish and odonates, I provide recommendations for the application of the Data Deficient category in IUCN Red List assessments. Under current IUCN regulations Data Deficient species can be assigned to two justification tags: uncertain provenance and uncertain taxonomy (IUCN 2012). In this thesis, I assigned Data Deficient species to eight justification tags: new species, taxonomic uncertainty, type series, few records, old records, unknown record provenance, unknown population status or distribution, and unknown threats (Chapter 2). I believe eight tags are necessary to quantify the wide range of knowledge deficiencies found in Data Deficient species. In some taxonomic groups, very few Data Deficient species could be assigned to current IUCN tags. For example, 1% and 2% of Data Deficient freshwater crabs could respectively be assigned to the unknown provenance and taxonomic uncertainty tags (Chapter 2). In addition, the two current IUCN tags cannot be used to efficiently prioritize Data Deficient species for further research, field surveys and re-assessment to data-sufficient categories. Actions required to re-assess species known from type specimens collected more than a hundred years ago will differ from those required to re-assess species for which threat data are lacking, two cases that cannot be distinguished by current IUCN tags.

Data Deficient tags must not only quantify species' knowledge deficiencies, but be linked to research actions to remove Data Deficient species from the category in future assessments. Assessors can use the IUCN classification scheme to indicate which research actions are needed on a species (Table 6.1). However, use of the classification scheme is no longer Recommended Supporting Information as of September 2012 (IUCN 2013c). I argue that the Research Needed classification scheme should become Required Supporting Information for

Data Deficient species assessments. Indeed, there is little conservation utility in noting that species are too poorly-known to assess extinction risk, yet not indicate which actions would resolve data deficiency. Research Needed actions are closely linked to the proposed Data Deficient tags and should be consistent within species assessments.

If no plausible research actions can be selected, I suggest that the inclusion of the species in the Red List be reconsidered. For Data Deficient species of doubtful taxonomic status, *nomen dubiums*, and species for which type series no longer exist, determination of conservation status would require re-description under a new scientific name or recovery of lost material. In regards to undescribed species, the IUCN recommends that “there must be a clear conservation benefit to justify the inclusion of such listings” (IUCN Standards and Petitions Subcommittee 2013). I recommend a similarly cautious attitude to including species of doubtful taxonomic status in the Data Deficient category of the Red List.

Table 6.1 IUCN classification of Research Needed actions (IUCN 2013b).

1 Research
1.1 Taxonomy
1.2 Population size, distribution & trends
1.3 Life history & ecology
1.4 Harvest, use & livelihoods
1.5 Threats
1.6 Actions
2 Conservation Planning
2.1 Species Action/Recovery Plan
2.2 Area-based Management Plan
2.3 Harvest & Trade Management Plan
3 Monitoring
3.1 Population trends
3.2 Harvest level trends
3.3 Trade trends
3.4 Habitat trends
4 Other

I recommend that the date of last record and details on past successful and unsuccessful surveys be included as Recommended Supporting Information in Data Deficient assessments. Date last recorded and details of surveys are already included as Required Supporting Information for Extinct, Extinct in the Wild, Critically Endangered (Possibly Extinct) and Critically Endangered (Possibly Extinct in the Wild) categories to justify assessments and

allow basic analyses (IUCN Standards and Petitions Subcommittee 2013). Such information would also be useful for Data Deficient species, as the date of last record is important in quantifying knowledge deficiency on a species and estimating the likelihood of observing a species (Sousa-Baena *et al.* 2013). Information on both successful and unsuccessful surveys can inform estimates of species detectability (Wintle *et al.* 2005), estimates of decreases in population or range size, and ultimately re-assessment to data-sufficient categories (Good *et al.* 2006). Information on last record and surveys may already be included in some Data Deficient species assessments, but inconsistently or with considerable semantic uncertainty.

Semantic uncertainty is a major source of uncertainty on the IUCN Red List (IUCN Standards and Petitions Subcommittee 2013), and arises from vagueness in the definition of terms in the criteria and lack of consistency in their usage. Semantic uncertainty is prevalent in Data Deficient species assessments and reduces their conservation utility. For example, the freshwater crab *Parathelphusa sarawakensis* from Malaysia is “listed as Data Deficient as very little is known about this species”, with no additional information on the type of information lacking. Assessments that are particularly vague or do not mention type series or old records may over-estimate information availability. As a result, the number of species assigned to the Data Deficient category due to severe forms of uncertainty may have been under-estimated (Chapter 2). Particular care should be taken in distinguishing type series and type localities, which represent different likelihoods of the occurrence of a species at a particular site. Indeed, a species observed from a type locality may have been observed once in the distant past, or multiple times relatively recently.

Butchart & Bird (2010) hypothesized the Data Deficient category to be the most misunderstood and controversial category on the Red List, and the most heterogeneous among taxonomic groups. My research confirms this statement: I find that heterogeneity results from both genuine differences in information availability and differences in reporting among groups. I believe that the application of Data Deficient tags and Research Needed classification could resolve most issues relating to vague and uninformative assessment rationales, without requiring large time commitment from assessors. Consistent tagging of Data Deficient species could minimize differences in reporting among taxonomic groups, and highlight genuine differences in information availability and assessor risk attitude among groups. I note that many Data Deficient species tagged under unknown population status and unknown threats in relatively well-known groups (such as mammals and crayfish) could be assigned to data-sufficient categories if homogeneous risk attitudes were implemented among taxonomic groups.

Concluding remarks

In this thesis, I identified and resolved the effects of Data Deficient species on the estimation of global patterns and levels of extinction risk. I showed that conservation objectives can be cost-effectively achieved by linking predictive macroecological models with decision theory. I believe decision theory could be used more widely in ecology to inform the estimation of biodiversity patterns.

References

- Adamowicz, S.J. & Purvis, A., 2006. Macroevolution and extinction risk patterns in freshwater crayfish. *Freshwater Crayfish*, 15, pp.1–23.
- Akcakaya, H.R. et al., 2000. Making Consistent IUCN Classifications under Uncertainty. *Conservation Biology*, 14(4), pp.1001–1013.
- Anselin, L., 1988. *Spatial econometrics: methods and models*, Dordrecht, Netherlands: Springer. 284 pp.
- Araújo, M.B. & Guisan, A., 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), pp.1677–1688.
- Arkive, 2013. Arkive. Available at: www.arkive.org. Accessed November 18, 2013.
- Baillie, J.E.M. et al., 2008. Toward monitoring global biodiversity. *Conservation Letters*, 1(1), pp.18–26.
- Bajomi, B. et al., 2010. Bias and dispersal in the animal reintroduction literature. *Oryx*, 44(03), pp.358–365.
- Baker, S.G., 1991. Evaluating a new test using a reference test with estimate sensitivity and specificity. *Communications in Statistics - Theory and Methods*, 20, pp.2739–2752.
- Balmford, A. et al., 2005. The 2010 challenge: data availability, information needs and extraterrestrial insights. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1454), pp.221–8.
- Balmford, A. & Gaston, K.J., 1999. Why biodiversity surveys are good value. *Nature*, 398(6724), pp.204–205.
- Barbosa, A.M. et al., 2010a. Is the human population a large-scale indicator of the species richness of ground beetles? *Animal Conservation*, 13(5), pp.432–441.
- Barbosa, A.M. et al., 2010b. Positive regional species-people correlations: a sampling artefact or a key issue for sustainable development? *Animal Conservation*, 13(5), pp.446–447.
- Bart, J. & Earnst, S., 2002. Double sampling to estimate density and population trends in birds. *The American Ornithologists' Union*, 119(1), pp.36–45.
- Benjamini, Y & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society Series B*, 57, pp.289–300
- Bennett, P.M. & Owens, I.P.F., 1997. Variation in extinction risk among birds: chance or evolutionary predisposition? *Proceedings of the Royal Society B: Biological Sciences*, 264(1380), pp.401–408.
- Bielby, J. et al., 2010. Modelling extinction risk in multispecies data sets: phylogenetically independent contrasts versus decision trees. *Biodiversity and Conservation*, 19(1), pp.113–127.
- Bielby, J. et al., 2008. Predicting susceptibility to future declines in the world's frogs. *Conservation Letters*, 1, pp.82–90.
- Bielby, J., Cunningham, A.A. & Purvis, A., 2006. Taxonomic selectivity in amphibians: ignorance, geography or biology? *Animal Conservation*, 9(2), pp.135–143.

- Bini, L.M. et al., 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity and Distributions*, 12(5), pp.475–482.
- Bivand, R. et al., 2014. Spdep package in R. Accessed at: <http://cran.r-project.org/web/packages/spdep/index.html> Accessed on the 7th of July 2013.
- Blackburn, T.M. & Gaston, K.J., 1994. Are Newly Described Bird Species Small-Bodied? *Biodiversity Letters*, 2(1), pp.16–20.
- Blackwell, M., 2011. The Fungi: 1, 2, 3. . .5.1 million species? *American Journal of Botany*, 98, pp.426–438.
- Bland, L.M. et al., 2012. Data uncertainty and the selectivity of extinction risk in freshwater invertebrates. *Diversity and Distributions*, 18(12), pp.1211–1220.
- Böhm, M. et al., 2013. The conservation status of the world's reptiles. *Biological Conservation*, 157, pp.372–385.
- Boitani, L. et al., 2011. What spatial data do we need to develop global mammal conservation strategies? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), pp.2623–32.
- Bonnet, X., Shine, R. & Lourdaï, O., 2002. Taxonomic chauvinism. *Trends in Ecology & Evolution*, 17(1), pp.2000–2002.
- Boyer, A.G., 2008. Extinction patterns in the avifauna of the Hawaiian islands. *Diversity and Distributions*, 14(3), pp.509–517.
- Breiman, L. et al., 1984. *Classification and regression trees*, Belmont, CA: Wadsworth International Group. 368 pp.
- Breiman, L., 2001. Random Forests. *Machine Learning*, 45, pp.5–32.
- Brito, D., 2010. Overcoming the Linnean shortfall: Data deficiency and biological survey priorities. *Basic and Applied Ecology*, 11(8), pp.709–713.
- Brodie, J.F., 2009. Is research effort allocated efficiently for conservation? Felidae as a global case study. *Biodiversity and Conservation*, 18(11), pp.2927–2939.
- Brooks, D. et al. 1993. Historical ecology: examining phylogenetics components of community evolution. In *Species diversity in ecological communities: historical and geographical perspectives*, Chicago, IL. University of Chicago Press. pp.267–280.
- Brooks, T.M. et al., 2006. Global biodiversity conservation priorities. *Science*, 313(5783), pp.58–61.
- Brown, J.H. & Lomolino, M. V, 1998. *Biogeography*, Sunderland, MA: Sinauer Press. 624 pp.
- Bruggeman, J., Heringa, J. & Brandt, B.W., 2009. PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, 37, pp.179–184.
- Brummitt, N., Bachman, S.P. & Moat, J., 2008. Applications of the IUCN Red List: towards a global barometer for plant diversity. *Endangered Species Research*, 6, pp.127–135.
- Butchart, S.H.M. et al., 2004. Measuring global trends in the status of biodiversity: red list indices for birds. *PLoS Biology*, 2(12), p.e383.

- Butchart, S.H.M. et al., 2005. Using Red List Indices to measure progress towards the 2010 target and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1454), pp.255–268.
- Butchart, S.H.M. et al., 2010. Global biodiversity: indicators of recent declines. *Science*, 328(5982), pp.1164–1168.
- Butchart, S.H.M. & Bird, J.P., 2010. Data Deficient birds on the IUCN Red List: What don't we know and why does it matter? *Biological Conservation*, 143, pp.239–247.
- Cantarello, E. et al., 2010. A multi-scale study of Orthoptera species richness and human population size controlling for sampling effort. *Die Naturwissenschaften*, 97(3), pp.265–71.
- Cardillo, M. et al., 2004. Human population density and extinction risk in the world's carnivores. *PLoS Biology*, 2(7), pp.909–914.
- Cardillo, M. et al., 2005. Multiple causes of high extinction risk in large mammal species. *Science*, 309, pp.1239–1241.
- Cardillo, M. et al., 2008. The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of the Royal Society B: Biological Sciences*, 275(1641), pp.1441–8.
- Cardillo, M. & Meijaard, E., 2012. Are comparative studies of extinction risk useful for conservation? *Trends in Ecology & Evolution*, 27(3), pp.167–171.
- Cardoso, P. et al., 2011. The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, 144(11), pp.2647–2655.
- Carwardine, J. et al., 2008. Cost-effective priorities for global mammal conservation. *Proceedings of the National Academy of Sciences*, 105(32), pp.11446–50.
- Ceballos, G. & Ehrlich, P.R., 2009. Discoveries of new mammal species and their implications for conservation and ecosystem services. *Proceedings of the National Academy of Sciences*, 106(10), pp.3841–3846.
- Chiew, F. & McMahon, T. 2002. Modelling the impacts of climate change on Australian streamflow. *Hydrological processes*. 16(6), pp.1235–1245
- CIESIN, 2002. Country-level Population and Downscaled Projections based on the B2 Scenario (1990), Center for International Earth Science Information Network (CIESIN). Available at: <http://www.ciesin.columbia.edu/datasets/downscaled>. Accessed on 10th February 2011.
- CIESIN, 2005a. Gridded Population of the World (2000), Version 3 (GPWv3), Center for International Earth Science Information Network (CIESIN). Available at: <http://sedac.ciesin.columbia.edu/gpw>. Accessed on 10th February 2011.
- CIESIN, 2005b. Last of the Wild Data Version 2 (LWP-2): Global Human Footprint dataset (HF), Center for International Earth Science Information Network (CIESIN). Available at: <http://sedac.ciesin.columbia.edu/data/collection/wildareas-v2>. Accessed on 10th February 2011.
- Clausnitzer, V. & Jödicke, R. 2004. Guardians of the watershed. *International Journal of Odonatology*. 7(2), pp.1–111
- Clausnitzer, V. et al., 2009. Odonata enter the biodiversity crisis debate: the first global assessment of an insect group. *Biological Conservation*, 142(8), pp.1864–1869.

- Collen, B. et al., 2008a. Monitoring change in vertebrate abundance: the Living Planet Index. *Conservation Biology*, 23(2), pp.317–27.
- Collen, B. et al., 2008b. The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, 1(2), pp.75–88.
- Collen, B. et al., 2009. Broadening the coverage of biodiversity assessments. In *Wildlife in a changing world. An analysis of the 2008 IUCN Red List of Threatened Species*. Gland, Switzerland: IUCN, pp. 67–75.
- Collen, B. et al., 2011. Investing in evolutionary history: implementing a phylogenetic approach for mammal conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), pp.2611–2622.
- Collen, B. et al., 2014. Global patterns of freshwater species diversity, threat and endemism. *Global Ecology and Biogeography*, 23(1), pp.40–51.
- Collen, B. & Bailie, J.M., 2010. The barometer of life: sampling. *Science*, 329, p.140.
- Collen, B., Purvis, A. & Gittleman, J.L., 2004. Biological correlates of description date in carnivores and primates. *Global Ecology and Biogeography*, 13, pp.459–467.
- Convention on Biological Diversity, 2010. TARGET 12 - Technical Rationale. In *COP10 Decisions Tenth meeting of the Conference of the Parties to the Convention on Biological Diversity*. Nagoya, Japan: CBD.
- Cooper, N. et al., 2008. Macroecology and extinction risk correlates of frogs. *Global Ecology and Biogeography*, 17(2), pp.211–221.
- Cressie, N.A.C., 1993. *Statistics for spatial data*, USA: Wiley. 900 pp.
- Critical Ecosystems Partnership Fund, 2013. Critical Ecosystems Partnership Fund Report. Available at:<http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/ENVIRONMENT/EXTBIODIVERSITY/0,,contentMDK:20473187~menuPK:1170323~pagePK:148956~piPK:216618~theSitePK:400953,00.html>. Accessed on 11th November 2013.
- Cumberlidge, N. et al., 2009. Freshwater crabs and the biodiversity crisis: importance, threats, status, and conservation challenges. *Biological Conservation*, 142(8), pp.1665–1673.
- Cutler, R.D. et al., 2007. Random forests for classification in ecology. *Ecology*, 88(11), pp.2783–92.
- Dakins, M.E., 1999. The Value of the Value of Information. *Human and Ecological Risk Assessment*, 5(2), pp.37–41.
- Davidson, A.D. et al., 2009. Multiple ecological pathways to extinction in mammals. *Proceedings of the National Academy of Sciences*, 106(26), pp.10702–10705.
- Davidson, A.D. et al., 2012. Drivers and hotspots of extinction risk in marine mammals. *Proceedings of the National Academy of Sciences*, 109(9), pp.3395–400.
- Davies, R. G. et al. 2006 Human impacts and the global distribution of extinction risk. *Proceedings of the Royal Society Series B: Biological Sciences*, 273, 2127–2133.
- De'ath, G., 2007. Boosted Trees for Ecological Modeling and Prediction. *Ecology*, 88(1), pp.243–251.

- De'ath, G. & Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), pp.3178–3192.
- Dennis, R.L.H. & Thomas, C.D., 2000. Bias in Butterfly Distribution Maps: The Influence of Hot Spots and Recorder's Home Range. *Journal of Insect Conservation*, 4(2), pp.73–77.
- Diniz-Filho, J.A.F. et al., 2005. Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. *Global Ecology and Biogeography*, 14(5), pp.469–477.
- Diniz-Filho, J.A.F. et al., 2010. Defying the curse of ignorance: perspectives in insect macroecology and conservation biogeography. *Insect Conservation and Diversity*. 3(3), pp.172–179
- Dobson, A., 2005. Monitoring global rates of biodiversity change: challenges that arise in meeting the Convention on Biological Diversity (CBD) 2010 goals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1454), pp.229–41.
- Drake, J.M., Randin, C. & Guisan, A., 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, 43(3), pp.424–432.
- Duda, R.O., Hart, P.E. & Stork, D.G., 2001. *Pattern Classification*, USA: Wiley. 654 pp.
- Elith, J. & Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), pp.66–77.
- Elith, J., Leathwick, J.R. & Hastie, T., 2008. A working guide to boosted regression trees. *The Journal of Animal Ecology*, 77(4), pp.802–13.
- European Distributed Institute of Taxonomy, 2007. *Do we need to estimate inventory completeness? Utility and drawbacks of modelling techniques for biodiversity databases mining*, 24 pp. Available at: http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=1&cad=rja&uact=8&ved=0CC8QFjAA&url=http%3A%2F%2Fcybertaxonomy.eu%2Fblog%2Ffiles_edit_wp5%2F2007-07-26_D5.35_%26_D5.38.doc&ei=tqchU8juCoj17Aanl4BQ&usg=AFQjCNHvGqhOJL-KU1XNAR3WKf7pArxrTQ&sig2=39ggA14OEdWQNQtosvf8Iw&bvm=bv.62922401,d.ZGU. Accessed on 11th November 2013.
- Failing, L. & Gregory, R., 2003. Ten common mistakes in designing biodiversity indicators for forest policy. *Journal of Environmental Management*, 68(2), pp.121–132.
- Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874.
- Ficetola, G.F. et al., 2014. An evaluation of the robustness of global amphibian range maps. *Journal of Biogeography*, 41(2), pp.211–221.
- Ficetola, G.F. et al., 2013. Estimating patterns of reptile biodiversity in remote regions. *Journal of Biogeography*, 40(6), pp.1202–1211.
- Fielding, A.H. & Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/ absence models. *Environmental Conservation*, 24, 38–49. *Environmental Conservation*, 24, pp.38–49.
- Fisher, D.O. & Blomberg, S.P., 2011. Correlates of rediscovery and the detectability of extinction in mammals. *Proceedings of the Royal Society B: Biological Sciences*, 278(1708), pp.1090–7.
- Fisher, D.O. & Owens, I.P.F., 2004. The comparative method in conservation biology. *Trends in Ecology & Evolution*, 19(7), pp.391–398.

- Flach, P., Hernandez-Orallo, J. & Ferri, C., 2011. A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance. *Proceedings of the 28th International Conference of Machine Learning*.
- Freckleton, R.P., Cooper, N. & Jetz, W., 2011. Comparative methods as a statistical fix: the dangers of ignoring an evolutionary model. *The American Naturalist*, 178(1), pp.E10–7.
- Freund, Y. & Schapire, R.E., 1996. Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning*, pp. 148–156.
- Fritz, S.A. & Purvis, A., 2010. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24(4), pp.1042–51.
- 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(6), pp.538–49.
- Gardner, T.A. et al., 2008. The cost-effectiveness of biodiversity surveys in tropical forests. *Ecology Letters*, 11(2), pp.139–50.
- Gaston, K.J., Blackburn, T.M. & Loder, N., 1995. Which species are described first?: the case of North American butterflies. *Biodiversity and Conservation*, 4, pp.119–127.
- Gaston, K.J. & May, R.M., 1992. Taxonomy of taxonomists. *Nature*, 356, pp.281–282.
- Giam, X. et al., 2012. Reservoirs of richness: least disturbed tropical forests are centres of undescribed species diversity. *Proceedings of the Royal Society B: Biological Sciences*, 279(1726), pp.67–76.
- Gilbert, R.O., 1987. *Statistical methods for environmental pollution monitoring*, New York, NY: Van Nostrand Reinhold. 320 pp.
- Global Biodiversity Information Facility, 2013. Global Biodiversity Information Facility. Copenhagen, Denmark. Available at: www.gbif.org. Accessed on 11th September 2013.
- González-Suárez, M., Lucas, P.M. & Revilla, E., 2012. Biases in comparative analyses of extinction risk: mind the gap. *The Journal of Animal Ecology*, 81(6), pp.1211–22.
- Good, T.C., Zjhra, M.L. & Kremen, C., 2006. Addressing Data Deficiency in classifying extinction risk: a case study of a radiation of Bignoniaceae from Madagascar. *Conservation Biology*, 20(4), pp.1099–1110.
- Grenyer, R. et al., 2006. Global distribution and conservation of rare and threatened vertebrates. *Nature*, 444(7115), pp.93–96.
- Griffiths, H.J., 2010. Antarctic marine biodiversity--what do we know about the distribution of life in the Southern Ocean? *PloS One*, 5(8), p.e11683.
- Hand, D.J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), pp.103–123.
- Hand, D.J., 2012. Assessing the Performance of Classification Methods. *International Statistical Review*, 80(3), pp.400–414.
- Hand, D.J. & Anagnostopoulos, C., 2012. A better Beta for the H measure of classification performance. *arXiv preprint arXiv:1202.2564*

- Harper, M.J. et al., 2004. Overcoming bias in ground-based surveys of hollow-bearing trees using double-sampling. *Forest Ecology and Management*, 190(2-3), pp.291–300.
- Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning*, NY, USA: Springer. 746 pp.
- Hijmans, S.E. et al., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, pp.1965–1978.
- Hilton-taylor, C. et al., 2000. Assessment mismatches must be sorted out : they leave species at risk. *Nature*, 404, pp.5–6.
- Hilton-Taylor, C. et al., 2009. State of the world's species. In *Wildlife in a changing world. An analysis of the 2008 IUCN Red List of Threatened Species*. Gland, Switzerland: IUCN, pp. 15–41.
- Ho, L.S.T. & Ane, C., 2013. Package “phylolm”, Phylogenetic linear regression. Available at: <http://www.stat.wisc.edu/~lamho/phylolm/>. Accessed on 5th August 2013.
- Hoffmann, M. et al., 2010. The Impact of Conservation on the Status of the World' s Vertebrates. *Science*, 330, pp.1503–1509.
- Hortal, J., 2008. Uncertainty and the measurement of terrestrial biodiversity gradients. *Journal of Biogeography*, 35(8), pp.1335–1336.
- Hughes, L., 2003. Climate change and Australia: Trends, projections and impacts. *Austral Ecology*, 28, pp.423-443
- Imhoff, M.L. et al., 2004. Global patterns in human consumption of net primary production. *Nature*, 429, pp.870–873.
- Isaac, N.J.B. et al., 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. *PloS One*, 2(3), p.e296.
- Isaac, N.J.B., Mallet, J. & Mace, G.M., 2004. Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology & Evolution*, 19(9), pp.464–9.
- IUCN, 2001. IUCN Red List Categories and Criteria: version 3.1. Species Survival Commission, 30 pp.
- IUCN, 2008. 2008 IUCN Red List of Threatened Species. Available at: www.iucnredlist.org. Accessed on the 25th of July 2009.
- IUCN, 2010. 2010 IUCN Red List of Threatened Species. Available at: www.iucnredlist.org. Accessed on the 10th of October 2010.
- IUCN, 2012. *Rules of Procedure IUCN Red List Assessment Process 2015–2016. Version 2.0.*, Available at: http://www.iucnredlist.org/documents/Rules_of_Procedure_for_Red_List_2013-2016.pdf. Accessed on the 7th of September 2013.
- IUCN, 2013a. 2013 IUCN Red List of Threatened Species. Available at: www.iucnredlist.org. Accessed on the 8th of December 2013.
- IUCN, 2013b. *Documentation standards and consistency checks for IUCN Red List assessments and species accounts. Version 2.0.* Available at: <http://www.iucnredlist.org/technical-documents/red-list-training/red-list-guidance-docs>. Accessed on the 7th of September 2013.

- IUCN, 2013c. Red List Overview. *IUCN Red List of Threatened Species*. Available at: <http://www.iucnredlist.org/about/red-list-overview>. Accessed on the 12th of April 2013.
- IUCN Standards and Petitions Subcommittee, 2013. *Guidelines for Using the IUCN Red List Categories and Criteria. Version 10.1.*, Available at: <http://www.iucnredlist.org/documents/RedListGuidelines.pdf>. Accessed on the 8th of December 2013.
- Ives, A.R. & Garland, T., 2010. Phylogenetic logistic regression for binary dependent variables. *Systematic biology*, 59(1), pp.9–26.
- Jackson, R. et al., 2001. Water in a changing world, *Ecological Applications*, 11(4), pp.1027–1045.
- Johnson, C.N., Delean, S. & Balmford, A., 2002. Phylogeny and the selectivity of extinction in Australian marsupials. *Animal Conservation*, 5(2), pp.135–142.
- Jones, J.P.G. et al., 2011. The Why, What, and How of Global Biodiversity Indicators Beyond the 2010 Target. *Conservation Biology*, 25(3), pp.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(9), pp.2648–2648.
- Jones, K.E., Purvis, A. & Gittleman, J.L., 2003. Biological correlates of extinction risk in bats. *The American Naturalist*, 161(4), pp.601–14.
- Jones, K.E. & Safi, K., 2011. Ecology and evolution of mammalian biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1577), pp.2451–2461.
- Jones, M.J., Fielding, A. & Sullivan, M., 2006. Analysing extinction risk in parrots using decision trees. *Biodiversity and Conservation*, 15(6), pp.1993–2007.
- Joppa, L.N. et al., 2011. Biodiversity hotspots house most undiscovered plant species. *Proceedings of the National Academy of Sciences*, 108(32), pp.13171–6.
- Joseph, L.N., Maloney, R.F. & Possingham, H.P., 2009. Optimal allocation of resources among threatened species: a project prioritization protocol. *Conservation Biology*, 23(2), pp.328–38.
- Kalkman, V. et al., 2008. Global diversity of dragonflies (Odonata) in freshwater. *Hydrobiologia*, 595(1), pp.351–363.
- Kampichler, C. et al., 2010. Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Informatics*, 5(6), pp.441–450.
- Kapos, V. et al., 2008. Calibrating conservation: new tools for measuring success. *Conservation Letters*, 1(4), pp.155–164.
- Kerr, J. & Currie, D. 1995. Effects of human activity on global extinction risk, *Conservation Biology*, 9(6), pp.1528–1538.
- Knight, A.T. et al., 2010. Barometer of life: more action, not more data. *Science*, 329(5988), p.141.
- Koh, L.P., Sodhi, N.S. & Brook, B.W., 2004. Ecological correlates of extinction risk in tropical butterflies. *Conservation Biology*, 18(6), pp.1571–1578.
- Köhler, J. et al., 2005. New Amphibians and Global Conservation: A Boost in Species Discoveries in a Highly Endangered Vertebrate Group. *BioScience*, 55(8), p.693.

- Kuhn, M., 2008. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), pp.1–26.
- Ladle, R.J. et al., 2011. The causes and biogeographical significance of species' rediscovery. *Frontiers of Biogeography*, 3(3), pp.111–119.
- Larson, E.R. & Olden, J.D., 2010. Latent extinction and invasion risk of crayfishes in the southeastern United States. *Conservation Biology*, 24(4), pp.1099–1110.
- Lawson, C.R. et al., 2014. Prevalence, thresholds and the performance of presence-absence models R. *Methods in Ecology and Evolution*, 5(1), pp.54–64.
- Leader-Williams, N., Adams, W.M. & Smith, R.J., 2010. Deciding what to save: trade-offs in conservation. In Nigel Leader-Williams & William M. Adams, eds. *Trade-Offs in Conservation: Deciding What to Save*. Chichester, UK: Wiley-Blackwell, pp.3–14.
- Lee, T.M. & Jetz, W., 2011. Unravelling the structure of species extinction risk for predictive conservation science. *Proceedings of the Royal Society B: Biological Sciences*, 278, pp.1329–1338.
- Legendre, L. & Legendre, P., 1998. *Numerical Ecology* 2nd Edition, Amsterdam, Netherlands: Elsevier Science. 1006 pp.
- Lek, S. & Gue, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, 120, pp.65–73.
- Liu, C. et al., 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 3, pp.385–393.
- Lobo, J.M. et al., 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? *Diversity and Distributions*, 13(6), pp.772–780.
- Lomolino, M. V, 2004. Conservation biogeography. In M. V Lomolino & L. R. Heaney, eds. *Frontiers of Biogeography*. Sunderland, MA: Sinauer Associates, pp. 293–296.
- Luck, G.W. et al., 2010. What drives the positive correlation between human population density and bird species richness in Australia? *Global Ecology and Biogeography*, 19(5), pp.673–683.
- Mace, G.M., 2004. The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1444), pp.711–9.
- Mace, G.M. et al., 2008. Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species. *Conservation Biology*, 22(6), pp.1424–1442.
- Mace, G.M. et al., 2010. Biodiversity targets after 2010. *Current Opinion in Environmental Sustainability*, 2(1-2), pp.3–8.
- Mace, G.M. & Baillie, J.E.M., 2007. The 2010 biodiversity indicators: challenges for science and policy. *Conservation Biology*, 21(6), pp.1406–1413.
- Malmqvist, B. & Rundle, S. 2002. Threats to the running water ecosystems of the world, *Environmental Conservation*, 29(2), pp.134–153.
- Marsh, H. et al., 2007. Optimizing allocation of management resources for wildlife. *Conservation Biology*, 21(2), pp.387–99.

- Martín-López, B., González, J. a. & Montes, C., 2011. The pitfall-trap of species conservation priority setting. *Biodiversity and Conservation*, 20(3), pp.663–682.
- May, J.A. & Clark, R.M., 2002. Taxonomic Bias in Conservation Research. *Science*, 297(5579), pp.191–192.
- May, R. M., Lawton, J. H. & Stork, N. E. 1995 Assessing extinction rates. In J. H. Lawton & R. M. May *Extinction rates*, pp. 1–24. Oxford, UK: Oxford University Press.
- May, R.M., 2011. Why worry about how many species and their loss? *PLoS Biology*, 9(8), p.e1001130.
- McCarthy, D.P. et al., 2012. Financial costs of meeting global biodiversity conservation targets: current spending and unmet needs. *Science*, 338(6109), pp.946–9.
- McDonald-Madden, E. et al., 2010. Monitoring does not always count. *Trends in Ecology & Evolution*, 25(10), pp.547–50.
- McKinney, M.L., 1997. Extinction vulnerability and selectivity: combining ecological and paleontological views, *Annual Review of Ecology and Systematics*, 28(1), pp.495-516.
- McKinney, M.L., 2010. Shedding some light on people and biodiversity. *Animal Conservation*, 13(5), pp.444–445.
- Metrick, A. & Weitzman, M.L., 1996. Patterns of behavior in endangered species preservation. *Land and Economics*, 72(1), pp.1–16.
- Millenium Ecosystem Assessment, 2005. Ecosystems and human well-being. Synthesis. Vol. 5 Washington D.C.: Island Press. 53 pp.
- Moerman, D.E. & Estabrook, G.F., 2006. The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*, 33, pp.1969–1974.
- Mohamed bin Zayed Species Conservation Fund, 2013. The Mohamed bin Zayed Species Conservation Fund. Available at: <http://www.speciesconservation.org/>. Accessed on the 13th of October 2012.
- Mokany, K. & Ferrier, S., 2011. Predicting impacts of climate change on biodiversity: a role for semi-mechanistic community-level modelling. *Diversity and Distributions*, 17(2), pp.374–380.
- Mora, C. et al., 2011. How many species are there on Earth and in the ocean? G. M. Mace, ed. *PLoS Biology*, 9(8), p.e1001127.
- Morais, A.R. et al., 2013. Unraveling the conservation status of Data Deficient species. *Biological Conservation*, 166, pp.98–102.
- Morrow, E.H. & Pitcher, T.E., 2003. Sexual selection and the risk of extinction in birds. *Proceedings of the Royal Society B: Biological Sciences*, 270(1526), pp.1793–9.
- Murray, K.A. et al., 2014. Threat to the point: improving the value of comparative extinction risk analysis for conservation action. *Global Change Biology*. 20(2), pp.483-494.
- Myers, N. et al., 2000. Biodiversity hotspots for conservation priorities. *Nature*, 403(6772), pp.853–8.
- Nakagawa, S. & Freckleton, R.P., 2008. Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution*, 23(11), pp.592–6.

- Nakagawa, S. & Schielzeth, H., 2013. A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), pp.133–142.
- Nee, S. & May, R. 1997. Extinction and the loss of evolutionary history, *Science*, 278(5338), pp.692–694.
- Nelson, A., 2008. *Estimated travel time to the nearest city of 50,000 or more people in year 2006*, Accessed at: <http://bioval.jrc.ec.europa.eu/products/gam/download.htm>. Accessed on the 10th of October 2011.
- Nicholson, E. & Possingham, H.P., 2006. Objectives for Multiple-Species Conservation Planning. *Conservation Biology*, 20(3), pp.871–881.
- Nicholson, E. et al. 2012 Making Robust Policy Decisions Using Global Biodiversity Indicators. *PLoS One*, 7, e41128.
- Olden, J.D., Lawler, J.J. & Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, 83(2), pp.171–93.
- Olson, D. et al., 2001. Terrestrial ecoregions of the world: a new map of life on earth, *BioScience*, 51(11), pp.933–938.
- Orme, D.C. et al., 2012. Package “caper”, Comparative analyses of phylogenetics and evolution in R. Available at: <http://cran.r-project.org/web/packages/caper/caper.pdf>. Accessed on the 20th of October 2010.
- Owens, I. & Bennett, P.M., 2000. Ecological basis of extinction risk in birds: habitat loss versus human persecution and introduced predators. *Proceedings of the National Academy of Sciences*, 97(22), pp.12144–12148.
- Ozesmi, S., Tan, C. & Ozesmi, U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecological Modelling*, 195(1-2), pp.83–93.
- Pandit, M.K., Pockock, M.J.O. & Kunin, W.E., 2011. Ploidy influences rarity and invasiveness in plants. *Journal of Ecology*, 99(5), pp.1108–1115.
- Patterson, B.D., 1994. Accumulating Knowledge on the Dimensions of Biodiversity: Systematic Perspectives on Neotropical Mammals. *Biology Letters*, 2(3), pp.79–86.
- Pautasso, M. & McKinney, M.L., 2007. The botanist effect revisited: plant species richness, county area, and human population size in the United States. *Conservation Biology*, 21(5), pp.1333–40.
- Pearce, J. & Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133(3), pp.225–245.
- People’s Trust for Endangered Species, 2013. People’s Trust for Endangered Species. Available at: <http://www.ptes.org/>. Accessed on the 13th of October 2013.
- Perkins, N.J. & Schisterman, E.F., 2006. The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, 163(7), pp.670–5.
- Petchey, O.L. & Gaston, K.J., 2006. Functional diversity: back to basics and looking forward. *Ecology Letters*, 9(6), pp.741–58.

- Peterson, A.T., Slade, N.A. & History, N., 1998. Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. *Diversity and Distributions*, 4(3), pp.95–105.
- Poduri S. R. S. Rao, 2005. Double Sampling. In P. Armitage & T. Colton, eds. *Encyclopedia of Biostatistics*. Chichester, UK: John Wiley & Sons, Ltd, 718 pp.
- Possingham, H.P. et al., 2002. Limits to the use of threatened species. *Trends in Ecology & Evolution*, 17(11), pp.503–507.
- Prasad, A.M., Iverson, L.R. & Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2), pp.181–199.
- Purvis, A. et al., 2005. Correlates of extinction risk: phylogeny, biology, threat and scale. In A. Purvis, J. L. Gittleman, & T. M. Brooks, eds. *Phylogeny and Conservation*, Cambridge, UK: Cambridge University Press. pp.295–316.
- Purvis, A., 2008. Phylogenetic approaches to the study of extinction. *Annual Review of Ecology, Evolution and Systematics*, 39, pp.301–319.
- Purvis, A., et al., 2000a. Predicting extinction risk in declining species. *Proceedings of the Royal Society B: Biological Sciences*, 267, pp.1947–1952.
- Purvis, A., Jones, K.E. & Mace, G.M., 2000b. Extinction. *BioEssays*, 22(12), pp.1123–1133.
- R Development Core Team, 2010. R 2.12.0: A Language and Environment for Statistical Computing. Available at: <http://www.r-project.org>. Accessed on 10th of October 2010.
- Rayner, L., Ellis, M. & Taylor, J.E., 2011. Double sampling to assess the accuracy of ground-based surveys of tree hollows in eucalypt woodlands. *Austral Ecology*, 36(3), pp.252–260.
- Recknagel, F., 2001. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146, pp.303–310.
- Reddy, S. & Davalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, 30(11), pp.1719–1727.
- Reeder, D.A.M., Helgen, K.M. & Wilson, D.E., 2007. Global trends and biases in new mammal species discoveries. *Occasional Papers of the Museum of Texas Tech University*, 269, pp.1–35.
- Revenga, C. et al., 2005. Prospects for monitoring freshwater ecosystems towards the 2010 targets. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1454), pp.397–413.
- Rodrigues, A.S.L. et al., 2006. The value of the IUCN Red List for conservation. *Trends in Ecology & Evolution*, 21(2), pp.71–76.
- Rondinini, C. et al., 2011a. Reconciling global mammal prioritization schemes into a strategy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), pp.2722–8.
- Rondinini, C. et al., 2011b. Global habitat suitability models of terrestrial mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), pp.2633–41.
- Rondinini, C. et al., 2006. Trade offs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, 9(10), pp.1136–45.
- Rondinini, C. et al., 2013. Update or outdate: long-term viability of the IUCN Red List. *Conservation Letters*, Early View.

- Russell, G. et al., 1998. Present and future taxonomic selectivity in bird and mammal extinctions. *Conservation Biology*, 12(6), pp.1365-1376.
- Safi, K. & Pettorelli, N., 2010. Phylogenetic, spatial and environmental components of extinction risk in carnivores. *Global Ecology and Biogeography*, 19, pp.352-362.
- Sahlin, U. et al., 2011. A benefit analysis of screening for invasive species - base-rate uncertainty and the value of information. *Methods in Ecology and Evolution*, 2, pp.500-508.
- Samways, M. & Böhm, M., 2010. Invertebrata. Are vertebrates representative of animal biodiversity as a whole? In J. E. M. Bailie et al., eds. *Evolution lost: status and trends of the world's vertebrates*. London, UK: Zoological Society of London, pp. 55-61.
- Sastre, P. & Lobo, J.M., 2009. Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, 142(2), pp.462-467.
- Scheffers, B.R. et al., 2011. The world's rediscovered species: back from the brink? *PloS One*, 6(7), p.e22531.
- Scheffers, B.R. et al., 2012. What we know and don't know about Earth's missing biodiversity. *Trends in Ecology & Evolution*, 27(9), pp.501-10.
- Schipper, J. et al., 2008. The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science*, 322(5899), pp.225-30.
- Schmidt, K., 1932. Stomach contents of some american coral snakes, with the description of a new species of *Geophis*. *Copeia*, 1, pp.6-9.
- Soberón, J. et al., 2007. Assessing completeness of biodiversity databases at different spatial scales. *Ecography*, 30(1), pp.152-160.
- Sodhi, N.S. et al., 2008. Measuring the Meltdown: Drivers of Global Amphibian Extinction and Decline. *PLoS One*, 3(2), p.e1636.
- Sodhi, N.S. et al., 2011. Conservation successes at micro-, meso- and macroscales. *Trends in Ecology & Evolution*, 26(11), pp.585-94.
- Sousa-Baena, M.S., Garcia, L.C. & Peterson, T.A., 2013. Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. *Biological Conservation*. Early View.
- Strayer, D. 2006. Challenges for freshwater invertebrate conservation, *Journal of the North American Benthological Society*, 25(2), pp.271-287.
- Strayer, D.L. & Dudgeon, D., 2010. Freshwater biodiversity conservation: recent progress and future challenges. *Journal of the North American Benthological Society*, 29(1), pp.344-358.
- Stuart, S.N. et al., 2004. Status and trends of amphibian declines and extinctions worldwide. *Science*, 306(5702), pp.1783-1786.
- Stuart, S.N. et al., 2010. The barometer of life. *Science*, 328(5975), p.177.
- Sullivan, M.S. et al., 2000. Comparative analyses of correlates of Red data book status: a case study using European hoverflies (Diptera: Syrphidae). *Animal Conservation*, 3, pp.91-95.
- Tenenbein, A., 1970. A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, 65(331), pp.1350-1361.

- Tenenbein, A., 1971. A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications: Sample size Determination. *International Biometric Society*, 27(4), pp.935–944.
- Tenenbein, A., 1972. A Double Sampling Scheme for Estimating from Misclassified Multinomial Data with Applications to Sampling Inspection. *Technometrics*, 14(1), pp.187–202.
- The Royal Society, 2003. *Measuring biodiversity for conservation*, London, UK. 57 pp.
- Tobler, M. et al., 2007. Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodiversity and Conservation*, 16(3), pp.659–677.
- Trimble, M.J. & Van Aarde, R.J., 2010. Species inequality in scientific study. *Conservation Biology*, 24(3), pp.886–90.
- Trindade-Filho, J. et al., 2012. How does the inclusion of Data Deficient species change conservation priorities for amphibians in the Atlantic Forest? *Biodiversity and Conservation*, 21(10), pp.2709–2718.
- Vale, M.M. & Jenkins, C.N., 2012. Across-taxa incongruence in patterns of collecting bias. *Journal of Biogeography*, 39(9), pp.1744–1748.
- Vié, J.-C. et al., 2009. The IUCN Red List: a key conservation tool. In J.-C. Vié, C. Hilton-Taylor, & S. N. Stuart, eds. *Wildlife in a changing world. An analysis of the 2008 IUCN Red List of Threatened Species*. Gland, Switzerland: IUCN, pp. 1–13.
- Vorosmarty, C.J. et al., 2010. Global threats to human water security and river biodiversity. *Nature*, 467(7315), pp.555–561.
- Waller, L.A. & Gotway, C.A., 2004. *Applied spatial statistics for public health data*, Hoboken, NJ: Wiley. 520 pp.
- Webb, A., 2002. *Statistical Pattern Recognition*, Chichester, UK: Wiley. 496 pp.
- Whittaker, R.J. et al., 2005. Conservation Biogeography: assessment and prospect. *Diversity and Distributions*, 11(1), pp.3–23.
- Wilson, D.E. & Reeder, D.M., 2005. *Mammal species of the world. A taxonomic and geographic reference*, Baltimore, MD: Johns Hopkins University Press. 2000 pp.
- Wilson, K.A. et al., 2011. Prioritizing conservation investments for mammal species globally. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), pp.2670–80.
- Wilson, K.A. et al., 2006. Prioritizing global conservation efforts. *Nature*, 440(7082), pp.337–40.
- Wintle, B.A. et al., 2005. Estimating and dealing with detectability in occupancy surveys for forest owls and arboreal marsupials, *Journal of Wildlife Management*, 69(3), pp.915–917.
- World Association of Zoos and Aquaria, 2013. World Association of Zoos and Aquaria. Available at: <http://www.waza.org/en/site/home/>. Accessed on the 11th of November 2013.
- Yeo, D. et al., 2008. Global diversity of crabs (Crustacea: Decapoda: Brachyura) in freshwater. *Hydrobiologia*, 595(1), pp.275–286.
- Yesson, C. et al., 2007. How Global Is the Global Biodiversity Information Facility?, *PLoS One*, 2(11), pp.e1124.

Yokota, F. & Thompson, K.M., 2004. Value of information analysis in environmental health risk management decisions: past, present, and future. *Risk analysis :an official publication of the Society for Risk Analysis*, 24(3), pp.635–50.

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

Zhou, X.H., McClish, D.K. & Obuchowski, N.A., 2002. *Statistical Methods in Diagnostic Accuracy*, New York, NY: Wiley. 437 pp.

Zweifel, R.G., 2000. Partition of the Australopapuan microhylid frog genus *Sphenophryne* with descriptions of new species. *Bulletin of the American Museum of Natural History*, pp.1–130.

Appendix I

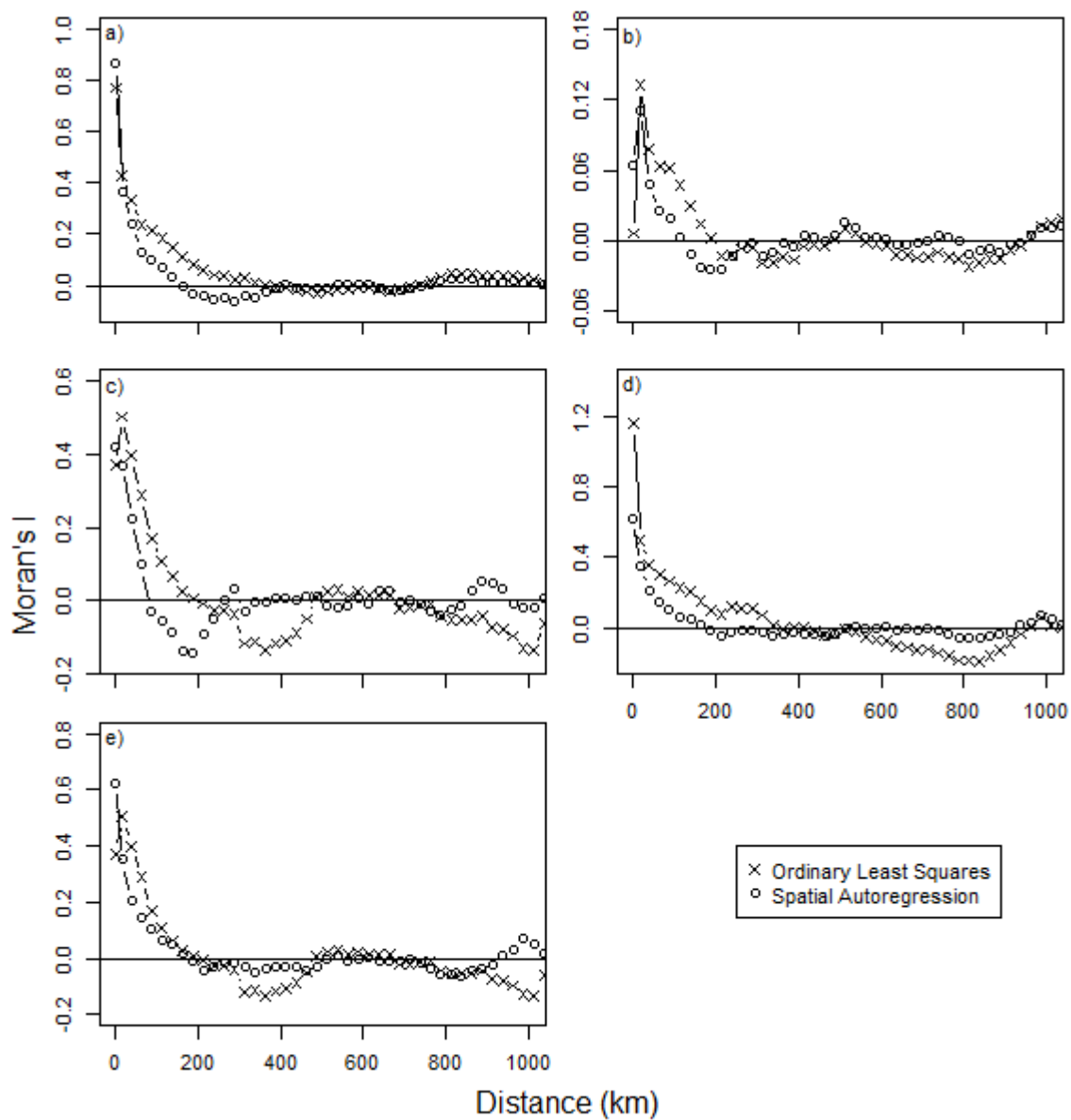


Figure S2.1 Correlogrammes of residual prevalence of data deficiency in mammals (a), amphibians (b), reptiles (c), freshwater crabs (d) and crayfish (e) for Ordinary Least Square regressions and Spatial Autoregressions.

Table S2.1 Matrix of spatial congruence in Data Deficient species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish. The comparison is presented for the richest 2.5% of 4,252 cells. Numerical values indicate, for each column, the proportion of hotspot cells encompassed by the hotspot cells of the row. A value of 1 indicates perfect congruence among groups.

	Mammals	Amphibians	Reptiles	Freshwater crabs	Crayfish
Mammals		0.35	0.07	0.08	0.02
Amphibians	0.21		0.06	0.10	0.02
Reptiles	0.05	0.08		0.16	0.01
Freshwater crabs	0.06	0.12	0.15		0.03
Crayfish	0.02	0.04	0.01	0.04	

Table S2.2 Matrix of spatial congruence in Data Deficient species richness in mammals, amphibians, reptiles, freshwater crabs, and crayfish. The comparison is presented for the richest 10% of 4,252 cells. Numerical values indicate, for each column, the proportion of hotspot cells encompassed by the hotspot cells of the row. A value of 1 indicates perfect congruence among groups.

	Mammals	Amphibians	Reptiles	Freshwater crabs	Crayfish
Mammals		0.37	0.11	0.30	0.03
Amphibians	0.52		0.28	0.41	0.08
Reptiles	0.19	0.34		0.37	0.05
Freshwater crabs	0.33	0.32	0.24		0.02
Crayfish	0.03	0.06	0.03	0.02	

Table S2.3 Ordinary least squares regressions of the prevalence of data deficiency in (a) mammals, (b) amphibians, (c) reptiles, (d) freshwater crabs, and (e) crayfish. HPD: human population density. S.E.: standard error. *: $p < 0.01$, **: $p < 0.001$, *** $p < 0.0001$.

Parameter	Estimate	S.E.	t value
a) Mammals (residual d.f. = 7,538)			
Intercept	7.48	0.067	110.4***
Species richness	-2.15	0.025	-84.12***
HPD	0.058	0.004	15.13***
Remoteness	-0.052	0.009	-5.49***
Species richness ²	0.169	0.003	58.6***
Species richness x remoteness	0.081	0.003	24.8***
b) Amphibians (residual d.f. = 6,017)			
Intercept	6.89	0.146	46.93***
Species richness	-1.588	0.038	-41.03***
HPD	-0.038	0.008	-4.43***
Remoteness	0.098	0.043	2.29
Species richness ²	0.094	0.003	28.01***
Remoteness ²	-0.011	0.004	-3.13*
Species richness x HPD	0.027	0.003	7.71***
c) Reptiles (residual d.f. = 6,230)			
Intercept	7.04	0.061	114.6***
Species richness	-1.37	0.038	-35.86***
HPD	-0.051	0.006	-8.839***
Accessibility	-0.0059	0.0087	-0.686
Species richness ²	0.033	0.005	6.29***
Species richness x HPD	0.0657	0.003	22.67***
Species richness x remoteness	0.045	0.005	8.58***
d) Freshwater crabs (residual d.f. = 3,274)			
Intercept	7.19	0.133	54.2***
Species richness	-0.65	0.091	-7.08***
HPD	-0.24	0.021	-12.03***
Remoteness	0.018	0.018	0.984
Species richness ²	0.127	0.009	12.86***
HPD ²	0.031	0.003	9.35***
Species richness x HPD	0.052	0.007	6.83***
Species richness x remoteness	-0.054	0.013	-4.14***
e) Crayfish (residual d.f. = 1,885)			
Intercept	6.03	0.246	24.56 ***
Species richness	-0.782	0.084	-9.21***
HPD	-0.023	0.0198	-1.16
Remoteness	0.284	0.08	3.54**
Species richness ²	0.112	0.008	12.62***
HPD ²	0.014	0.004	3.67**
Remoteness ²	-0.019	0.006	-2.95*
Species richness x remoteness	-0.06	0.012	-4.85***

Table S2.4 Results of the Lagrange multiplier tests for the prevalence of data deficiency in mammals, amphibians, reptiles, freshwater crabs, and crayfish. All tests have one degree of freedom.

	Spatial error		Spatial lag	
	Lagrange multiplier	p-value	Lagrange multiplier	p-value
Mammals	12,662	<0.0001	126	<0.0001
Amphibians	2,924	<0.0001	3.4	0.06
Reptiles	9,340	<0.0001	128	<0.0001
Freshwater crabs	16,566	<0.0001	116	<0.0001
Crayfish	3,213	<0.0001	5	0.02

Table S2.5 Quartile and median values of human population density and remoteness in the spatial models of prevalence of data deficiency in mammals, amphibians, reptiles, freshwater crabs, and crayfish.

	Human population density			Remoteness		
	First quartile	Median	Third quartile	First quartile	Median	Third quartile
Mammals	0	2.22	3.91	5.73	6.56	7.14
Amphibians	1.66	2.5	3.57	5.58	6.45	6.98
Reptiles	1.1	2.74	4.19	5.58	6.37	6.93
Freshwater crabs	2	2.77	4.15	5.64	6.52	6.99
Crayfish	1.54	2.55	3.82	5.14	5.83	6.72

Table S2.6 Single predictor generalized mixed models of Data Deficient status in mammals, with nested taxonomic levels. HPD: human population density. S.E.: standard error. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Predictor	Number of data-sufficient species	Number of Data Deficient species	Estimate	S.E.	z score	Variance due to order; family; genus
Range size	4,244	663	-0.35	0.016	-21.11***	0.442;0.569;0.83
Body size	3,039	227	-0.21	0.058	-3.74***	0.043;0.537;1.183
Number of habitats	4,236	653	-1.39	0.104	-13.4***	0.255;0.443;0.607
HPD	3,917	538	-0.15	0.038	-3.81***	0.218;0.503;0.612
Remoteness	4,091	622	-0.05	0.057	-0.95	0.189;0.504;0.66

Table S2.7 Multiple predictor generalized mixed models of Data Deficient status in mammals (a) Model including body size: 2,809 data-sufficient and 184 Data Deficient species. Variance due to order: 0.0009; family: 0.89; genus: 4.29. AIC = 1,125.7, marginal $R^2 = 0.212$, conditional $R^2 = 0.692$. (b) Model excluding body size: 3,739 data-sufficient and 478 Data Deficient species. Variance due to order: 0.88; family: 0.86; genus: 0.69. AIC = 2,385.6, marginal $R^2 = 0.237$, conditional $R^2 = 0.562$. HPD: human population density. S.E.: standard error. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Predictor	Estimate	S.E.	z score
a) Including body size			
Intercept	6.36	1.379	4.61***
Range size	-0.67	0.106	-6.28***
Body size	-0.93	0.27	-3.48***
Number of habitats	-0.78	0.24	-3.24**
HPD	-0.19	0.08	-2.42*
Range size x body size	0.05	0.02	2.32*
b) Excluding body size			
Intercept	2.13	0.41	5.27***
Range size	-0.36	0.02	-17.2***
Number of habitats	-0.86	0.14	-6.41***
HPD	-0.22	0.04	-5.2***

Appendix II

Table S3.1 Global taxonomic selectivity in odonates. DD: Data Deficient. ns= non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. NA: expected value not calculated due the null number of threatened and non-threatened species.

Family	Data Deficiency			DD species excluded			DD species non-threatened			DD species threatened		
	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend	Observed	Expected	Trend
	***			***			*			***		
Aeschnidae	39	35	+	7	8.6	-	7	9.1	-	46	44.7	+
Amphipterygidae	2	0.6	+	0	NA	NA	0	0.2	-	2	0.8	+
Austropetaliidae	1	12.6	-	0	0.4	-	0	0.3	-	1	1.7	-
Calopterygidae	8	16.2	-	5	5.3	-	5	4.2	+	13	20.7	-
Chlorocyphidae	8	7.4	+	5	1.9	+	5	1.9	+	13	9.5	+
Chlorogomphidae	4	2.6	+	3	0.5	+	3	0.7	+	7	3.3	+
Chlorolestidae	0	0.3	-	0	0.1	-	0	0.1	-	0	0.4	-
Ceonagrionidae	85	88.4	-	17	23.7	-	17	23	-	102	112.9	-
Cordulegastridae	7	5.8	+	4	1.4	+	4	1.5	+	11	7.5	+
Corduliidae	24	22.7	+	5	5.8	-	5	5.9	-	29	28.9	+
Euphaeidae	8	7.1	+	4	1.8	+	4	1.9	+	12	9.1	+
Gomphidae	115	116.9	-	21	31	-	21	30.4	-	136	149.4	-
Hemiphlebiidae	0	0.3	-	1	0.1	+	1	0.1	+	1	0.4	+
Isostictidae	6	3.6	+	1	0.6	+	1	0.9	+	7	4.5	+
Lestidae	11	10.4	+	2	2.6	-	2	2.7	-	13	13.2	-
Lestoideidae	0	1.3	-	0	0.5	-	0	0.3	-	0	1.7	-
Libellulidae	56	93.6	-	21	29.3	-	21	24.3	-	77	119.6	-
Macromiidae	8	5.8	+	2	1.3	+	2	1.3	+	7	6.2	+
Megapodagrionidae	44	26.9	+	9	4.9	+	9	7.2	+	53	34.4	+

Perilestidae	1	1.9	-	0	0.6	-	0	0.5	-	1	2.5	-
Petaluridae	0	1	-	1	0.5	+	1	0.4	+	1	1.7	-
Platycnemididae	31	19.7	+	9	3.8	+	9	5.2	+	40	25.2	+
Platystictidae	26	14.9	+	9	2.5	+	9	4	+	35	19	+
Polythoridae	5	4.2	+	0	1	-	0	1.1	-	5	5.4	-
Protoneuridae	31	23.6	+	8	5.3	+	8	6.4	+	39	30.2	+
Pseudolestidae	5	1.9	+	0	0.1	-	0	0.5	-	5	2.5	+
Pseudostigmatidae	0	0.6	-	0	0.3	-	0	0.2	-	0	0.8	-
Synlestidae	2	1.3	+	0	0.3	-	0	0.3	-	2	1.7	+
Synthemistidae	3	3.2	-	1	0.9	+	1	0.8	+	4	4.1	-

Appendix III

Database

I collated a trait database for 4,461 terrestrial mammal species. I based my analysis on the taxonomy provided in ‘Mammal Species of the World 3’ (Wilson & Reeder 2005). I classified species as non-threatened (LC, NT), threatened (VU, EN, CR) and Data Deficient (DD) (IUCN 2008). I 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 a 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.

I selected the following life-history and ecological variables due to their high completeness (Jones *et al.* 2009b): body mass (68.4% complete), litter size (48.8% complete), habitat breadth (52% complete), trophic level (42% complete), and number of IUCN listed habitats (100% complete). Because some ML methods cannot cope with missing data, I phylogenetically imputed missing life-history and ecological variables using a global mammal phylogeny (Fritz *et al.* 2009) and the PhyloPARS method (Bruggeman *et al.* 2009).

PhyloPARS estimates missing values at the nodes of a phylogeny using a limited number of observations and a specified evolutionary model, and allows for correlated evolution of different traits. Species which were not present on the phylogeny were assigned the median trait value for their genus or family. I recorded the biogeographical realm of each species from the IUCN Red List assessments (IUCN 2008), as well as their geographical range size (IUCN 2010) and latitude of range centroid. I extracted habitat suitability information from (Rondinini *et al.* 2011b) and computed the proportion of each species’ range deemed ‘highly suitable’. For each species, I also derived External Threat Index (ETI) values following the method proposed by Cardillo *et al.* (2004). The ETI for a given species is the mean threat status of all species present in the focal species’ range, weighted by the overlap in range between the focal species and all other species. The ETI is therefore a proxy measure of the level of threat within a species’ distribution. Using species range maps, environmental and anthropogenic threat variables were derived from global grids. All data extractions were conducted in ArcGIS version 9.2. All geographical variables were 100% complete for each species. Trait distributions were similar for data-sufficient and Data Deficient species (Figure S4.1).

Machine Learning tools

I compared the ability of seven commonly used Machine Learning (ML) algorithms (classification trees, random forests, boosted trees, k-nearest neighbours, support vector machines, neural networks and decision stumps) to predict species' threat status. I briefly introduce each ML tool.

Classification Trees. Classification Trees (CT) were first introduced by Breiman (Breiman *et al.* 1984) and explain variation in a response variable by repeatedly splitting the data into more homogeneous groups, using combinations of explanatory variables. Each terminal leaf is characterized by the value of the response variable, the number of observations in the group and the corresponding threshold values of the explanatory variables that define it.

Predictions are made by sorting new species down the CT until a leaf is reached. CTs make no distributional assumptions about the explanatory or response variables (Prasad *et al.* 2006), can fit non-linear relationships and high-order interactions and can handle missing values in the explanatory variables (De'ath & Fabricius 2000). However, they can be sensitive to small changes in the underlying data and can only approximate linear functions (Prasad *et al.* 2006). CTs have been widely used in ecology (De'ath & Fabricius 2000), including threatened species classification (Boyer 2008; Bielby *et al.* 2010; Larson & Olden 2010). I optimized tree depth (number of splits) during model training.

Random Forests. Random Forests (RF) are an ensemble method related to classification trees: many classification trees are constructed and classes are predicted by a majority vote (Breiman 2001). For each tree, only a randomly chosen subset of the explanatory variables is used at each node, which reduces correlation between trees and improves the overall classification accuracy of the RF. RFs are generally robust to overfitting, outliers and noise. RFs give direct estimates of variable importance and can model complex interactions between explanatory variables (Cutler *et al.* 2007), but unlike CTs they do not provide a simple representation of the classifier in the form of a single tree. RFs have become increasingly popular in ecology due to their high predictive power (Prasad *et al.* 2006; Cutler *et al.* 2007). I grew 500 trees at each model iteration and optimized the number of variables chosen randomly at each node.

Boosted Classification Trees. Boosted Classification Trees (BT) are an ensemble method constructed with a boosting algorithm (Freund & Schapire 1996) which begin by constructing a single classification tree. Instances (in this case, species) are then weighted by whether the initial tree predicted their class correctly. If the model did not predict the class

correctly, the instance is given extra weight in the following classification tree. The process continues until a stopping criterion is reached, such as a predefined number of trees. The final model predicts the class membership of a new example with a weighted voting scheme, where the voting power of each tree is proportional to its accuracy. BTs can therefore modify a classification tree with low predictive accuracy ('weak classifier') into a 'strong classifier' by focusing on difficult cases (Freund & Schapire 1996). The final model can be understood as an additive regression model in which individual terms are classification trees. BTs often show high predictive accuracy (Elith *et al.* 2008) but are prone to overfitting. I optimized the number of trees grown, tree depth and learning rate.

K-Nearest Neighbours. The K-Nearest Neighbour (KNN) is a learning algorithm based on instances (Hastie *et al.* 2009). Given an instance (in this case, species) its k closest neighbours are found in the n -dimensional feature space, where n denotes the number of explanatory variables. The class label of the instance is determined using a majority vote of the neighbours. A number of distance metrics have been proposed, but the most commonly used is Euclidian distance. KNNs have low memory requirements, but are sensitive to irrelevant explanatory variables and can exhibit higher error rates than more advanced methods. For each dataset, I chose the best performing classifiers created with a range of k values.

Support Vector Machines. Support Vector Machines (SVM) rely on processing the data to represent the pattern in a high dimension, typically much higher than the original feature space (Hastie *et al.* 2009). Using a kernel function, a SVM constructs a separating hyperplane between the training instances of both classes in the new space. Training of the SVM allows the determination of the separating hyperplane with the largest margin between the two classes. The support vectors are the training samples that define the optimal separating hyperplane and are the most difficult cases to classify; they are the patterns most informative for the classification task. SVMs are highly accurate classifiers, which do not suffer from local optima and are less prone to over-fitting than other methods (Duda *et al.* 2001). However, the parameters of the model are difficult to interpret. I used a Radial Basis kernel function and optimized sigma (inverse kernel width) and c , the cost of constraint violation.

Neural Networks. Neural networks are non-linear mapping structures based on Rosenblatt's perceptron. The most popular neural network is the multi-layer feed-forward network trained by a back-propagation algorithm (Recknagel 2001). Neurons are arranged in successive layers, and information flows uni-directionally from the input layer (explanatory variables)

to the output layer (response variable) through the hidden layer(s). Each hidden neuron is connected to each input and output neuron, and the strength of the initial connections are determined at the start training. Predicted and observed classes are compared, and learning is achieved through the updating of weights at each connection using back-propagation. Neural Networks often show high predictive performance and have been used in a wide range of ecological studies, however they suffer from slow training, and have been criticized for being a “black box” method with a tendency to overfit the data (Ozesmi *et al.* 2006). I optimized the number of neurons in the hidden layer and the model weight decay.

Decision stumps. In order to assess the role of geographical range size in determining extinction risk, I computed a decision stump (DT) for each dataset. Decision stumps are CTs derived from a single explanatory variable, in this case, geographical range size. Decision stumps effectively identify a geographical range size threshold above and below which species are attributed to a threat level.

Training of Machine Learning tools

I pre-processed the predictor variables as described in the package *caret* (Kuhn 2008). Numeric predictors were transformed, centred and scaled to a mean of zero and standard deviation of one, a common procedure in ML data pre-processing. For each taxonomic dataset separately, I then removed variables with near-zero variance, as these predictors may acquire zero-variance when the data are split into cross-validation sub-samples. I also removed highly correlated predictors (correlation coefficient > 0.9) as these can bias model fitting procedures. I set aside all DD species to form a prediction set. I randomly partitioned non-DD species into a training set comprising 75% of species and a validation set comprising 25% of species, to assess the performance of different ML methods. For each ML tool in turn, I optimized tuning parameters using ten-fold cross-validation on the training set. During each iteration of the cross-validation, the algorithm was trained on nine tenth of the data and tested on the excluded tenth (test fold), creating a set of built classifiers. Classifier performance was estimated by comparing the predicted and observed threat level of the species in the test folds. For each combination of tuning parameters, I measured the area under the receiver operating characteristic curve (AUC). The ROC curve is a graphical plot of the sensitivity against the false positive rate (1- specificity) of a classifier. The sensitivity of the classifier is the proportion of threatened species correctly identified, while the specificity is the proportion of non-threatened species correctly identified. AUC provides a tool for model selection which is insensitive to class imbalance and does not require the specification of misclassification costs (Fawcett 2006). Values of AUC higher than 0.7

indicate a good fit of the classifier to the data, while values higher than 0.9 indicate an extremely good fit (Pearce & Ferrier 2000). I selected the optimal tuning parameters for each ML tool using AUC rather than overall accuracy. Given the large class imbalance in some of the datasets, accuracy would provide a skewed measure of classifier performance. For example, 22.1% of mammals are threatened in the global dataset, hence any classifier that would classify all species as non-threatened (i.e. make no decision) would achieve an accuracy of 77.9%. Optimal tuning parameters for each ML tool can be found in Table S4.2. ML tools were compared independently on the validation sets previously set aside, and the best ML tool for each dataset was selected using AUC. As predictions of threat were probabilistic, predicting the threat category of a species required the determination of a predicted probability of threat above which a species should be classified as threatened. I used Youden's index (Youden 1950), to identify the optimal cutoff point. The Youden index Y is defined as ($Y = \text{sensitivity} + \text{specificity} - 1$), and can be intuitively interpreted as the point on the ROC curve farthest from chance (Perkins & Schisterman 2006). This method effectively assigns equal importance to sensitivity and specificity. Using the optimal cutoff point, I predicted the binary threat status of species in the validation sets and computed additional performance metrics, including specificity and sensitivity. I computed performance metrics for order-level predictions from the global model using both globally and ordinally optimized cutoff points (Table S4.3).

Multiple classification performance measures are commonly used among different research fields, reflecting varying attitudes towards misclassification costs (Hand 2012). To investigate the role of performance measure on my results, I repeated all analyses by maximizing the H measure, a recently developed alternative to AUC which allows the specification of misclassification costs (Hand 2009 but see Flach *et al.* 2011). I selected the prior distribution of misclassification costs based on the Beta($\pi_1 + 1$; $\pi_0 + 1$) distribution (Hand & Anagnostopoulos 2012), where π_1 is the proportion of threatened species in the sample, and π_0 the proportion of non-threatened species in the sample. The distribution takes a balanced view of misclassification costs when faced with unbalanced datasets, setting the mode of the relative misclassification severity distribution at $c = \pi_1$. As with models trained with AUC, I found a significant difference in performance among tools (Friedman test, $\chi^2 = 17.8$, $p = 0.006$, $df = 6$). *Post hoc* symmetry tests showed that this difference was caused by the difference between highly predictive boosted trees vs. k-nearest neighbours ($p = 0.01$, $df = 1$), and boosted trees vs. decision stumps based on geographical range size alone ($p = 0.03$, $df = 1$). The best model for all mammals and rodents remained random forests, and the best model for bats remained boosted trees (Table S4.5). The best models in carnivores and primates were

boosted trees, in contrast to neural networks and support vector machines respectively for models trained on AUC (Table S4.5). Model predictions between the best global model trained on AUC and the best model trained with the H measure were highly consistent (Table S4.7).

Table S4.1 Sources for the terrestrial mammal database.

Variable	Unit	Source	Resolution
Taxonomy (Order, Family, Genus)		IUCN 2008	
Body Mass	Grams	Jones <i>et al.</i> 2009	
Litter Size		Jones <i>et al.</i> 2009	
Habitat Breadth		Jones <i>et al.</i> 2009	
Trophic Level		Jones <i>et al.</i> 2009	
Number of IUCN Habitats		IUCN 2008	
Biogeographical Realms		IUCN 2008	
Range Size	km ²	IUCN 2010	
Latitude of Range Centroid	Degrees latitude	IUCN 2010	
High Habitat Suitability	Percent of range size	IUCN 2010; Rondinini <i>et al.</i> 2011	
Mean Annual Temperature	Degrees (°C)	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean Temperature Seasonality	Standard deviation*100	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean Annual Precipitation	Millimetres	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean Precipitation Seasonality	Coefficient of variation	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean Annual Net Primary Productivity (1976-2000)	Grams per m ² per year	Imhoff <i>et al.</i> 2004	0.25 degrees
Minimum Elevation	Meters	Hijmans <i>et al.</i> 2005	30 arc seconds
Elevation Range	Meters	Hijmans <i>et al.</i> 2005	30 arc seconds
External Threat Index		IUCN 2008; IUCN 2010	
Mean Human Population Density (2000)	People per unit area	CIESIN 2005a	2.5 arc minutes
Minimum Human Population Density (2000)	People per unit area	CIESIN 2005a	2.5 arc minutes
Mean Human Footprint	Human Influence Index normalized per region and biome	CIESIN 2005b	30 arc seconds
Mean Human Appropriation of Net Primary Productivity	Percent of NPP	Imhoff <i>et al.</i> 2004	0.25 degrees
Mean GDP (1990)	Dollars per person per year	CIESIN 2002	0.25 degrees

Table S4.2 Optimal tuning parameters for models trained with AUC among datasets. CT: Classification Tree, RF: Random Forests, BT: Boosted Trees, KNN: K-Nearest Neighbours, SVM: Support Vector Machines, NNET: Neural Networks. AUC: area under the receiver operator characteristic curve.

	CT	RF	BT			KNN	SVM		NNET	
	Tree depth	Number of variables randomly sampled	Number of trees	Tree depth	Learning rate	Number of neighbours	Sigma inverse kernel width	Cost of constraint violation	Number of units in the hidden layer	Weight decay
Global	0	6	212	11	0.1	29	0.0244	1	1	0.1
Bats	0.00931	9	96	13	0.1	27	0.0261	0.5	15	0.0422
Carnivores	0.0329	30	127	7	0.1	15	0.0227	1	1	0.00133
Primates	0.0106	8	206	6	0.1	9	0.0296	2	13	0
Rodents	0.0325	2	53	7	0.1	29	0.03	1	17	0.1

Table S4.3 Measures of model performance among validation sets for models trained on AUC. AUC: area under the receiver operator characteristic curve.

	Cutoff	Sensitivity	Specificity	Accuracy	H	AUC	Youden
Dataset predictions							
Global	0.282	0.935	0.887	0.898	0.785	0.944	0.754
Bats	0.067	0.914	0.842	0.854	0.597	0.897	0.756
Carnivores	0.808	0.900	0.917	0.913	0.759	0.961	0.817
Primates	0.547	0.861	0.727	0.803	0.499	0.873	0.588
Rodents	0.24	0.843	0.933	0.918	0.728	0.951	0.790
Order-level predictions from the global model (globally optimized cutoff point)							
Bats	0.282	0.821	0.937	0.916	0.74	0.956	0.758
Carnivores	0.282	0.778	0.939	0.905	0.773	0.969	0.717
Primates	0.282	1	0.743	0.888	0.732	0.955	0.743
Rodents	0.282	0.908	0.898	0.899	0.795	0.969	0.806
Order-level predictions from the global model (ordinally optimized cutoff point)							
Bats	0.192	0.897	0.856	0.864	0.74	0.956	0.779
Carnivores	0.162	0.888	0.849	0.857	0.773	0.969	0.85
Primates	0.472	0.844	0.886	0.863	0.732	0.955	0.752
Rodents	0.566	0.846	0.973	0.853	0.795	0.969	0.834

Table S4.4 Optimal tuning parameters for models trained with the H measure among datasets. CT: Classification Tree, RF: Random Forests, BT: Boosted Trees, KNN: K-Nearest Neighbours, SVM: Support Vector Machines, NNET: Neural Networks.

	CT	RF	BT			KNN	SVM		NNET	
	Tree depth	Number of variables randomly sampled	Number of trees	Tree depth	Learning rate	Number of neighbours	Sigma inverse kernel width	Cost of constraint violation	Number of units in the hidden layer	Weight decay
Global	0.0025	3	50	12	0.1	29	0.0245	2	1	0.1
Bats	0.0124	10	175	9	0.05	53	0.0246	0.25	15	0.0422
Carnivores	0.011	25	239	7	0.01	25	0.0224	0.25	1	0.00133
Primates	0.0106	23	89	11	0.05	9	0.0258	2	13	0
Rodents	0.0135	5	118	5	0.05	35	0.0295	1	17	0.1

Table S4.5 H measure for each combination of tool and dataset on the validation sets, for models trained on H measure. CT: Classification Tree, RF: Random Forests, BT: Boosted Trees, KNN: K-Nearest Neighbours, SVM: Support Vector Machine, NNET: Neural Networks, DS: Decision Stump.

	CT	RF	BT	KNN	SVM	NNET	DS
Global	0.518	0.642	0.611	0.519	0.619	0.557	0.389
Bats	0.582	0.555	0.61	0.479	0.523	0.509	0.382
Carnivores	0.505	0.726	0.789	0.404	0.663	0.759	0.372
Primates	0.356	0.497	0.560	0.334	0.525	0.331	0.272
Rodents	0.62	0.727	0.721	0.591	0.71	0.702	0.623

Table S4.6 Measures of model performance among validation sets for models trained on H measure. AUC: area under the receiver operator characteristic curve.

	Cutoff	Sensitivity	Specificity	Accuracy	H	AUC	Youden
Global	0.221	0.867	0.834	0.906	0.642	0.93	0.702
Bats	0.17	0.914	0.83	0.889	0.61	0.889	0.744
Carnivores	0.232	0.9	0.86	0.935	0.789	0.961	0.761
Primates	0.566	0.767	0.818	0.829	0.56	0.886	0.586
Rodents	0.17	0.86	0.887	0.935	0.727	0.948	0.773

Table S4.7 Differences in model predictions on the validation set between the best global model trained on AUC, and the best global model trained on H measure. Predictions for the model trained on AUC are taken as ‘true’ classes in the confusion matrix and following classification performance measures. Accuracy: 0.868. Sensitivity: 0.923. Specificity: 0.852. The validation set contains 991 species. AUC: area under the receiver operator characteristic curve.

		AUC predictions	
		Non-threatened	Threatened
H measure predictions	Non-threatened	656	17
	Threatened	114	204

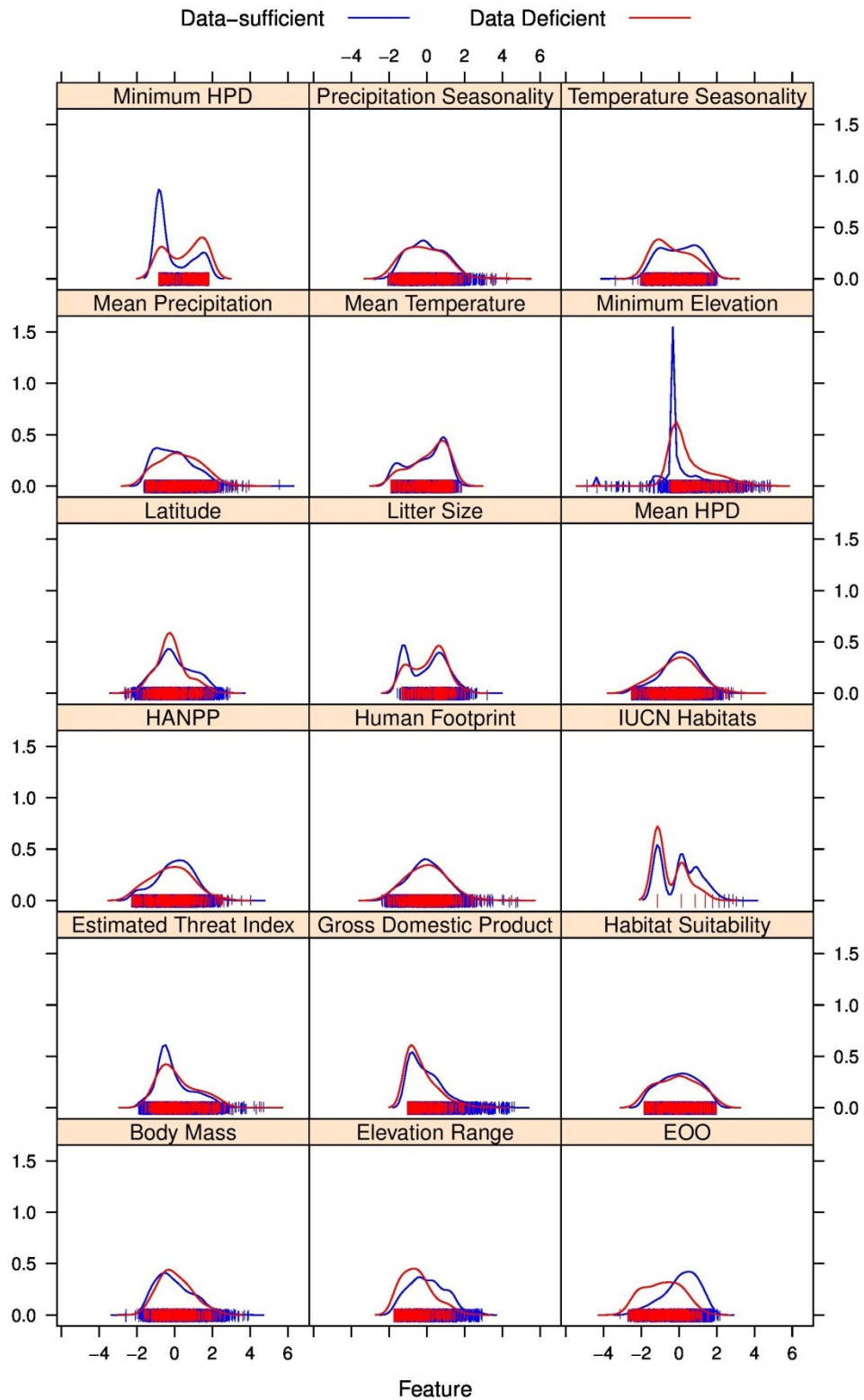


Figure S4.1 Overlay density plot of numeric explanatory variables in Data Deficient (n=493) and data-sufficient species (n=3,967). HPD: Human population density. HANPP: Human appropriation of net primary productivity.

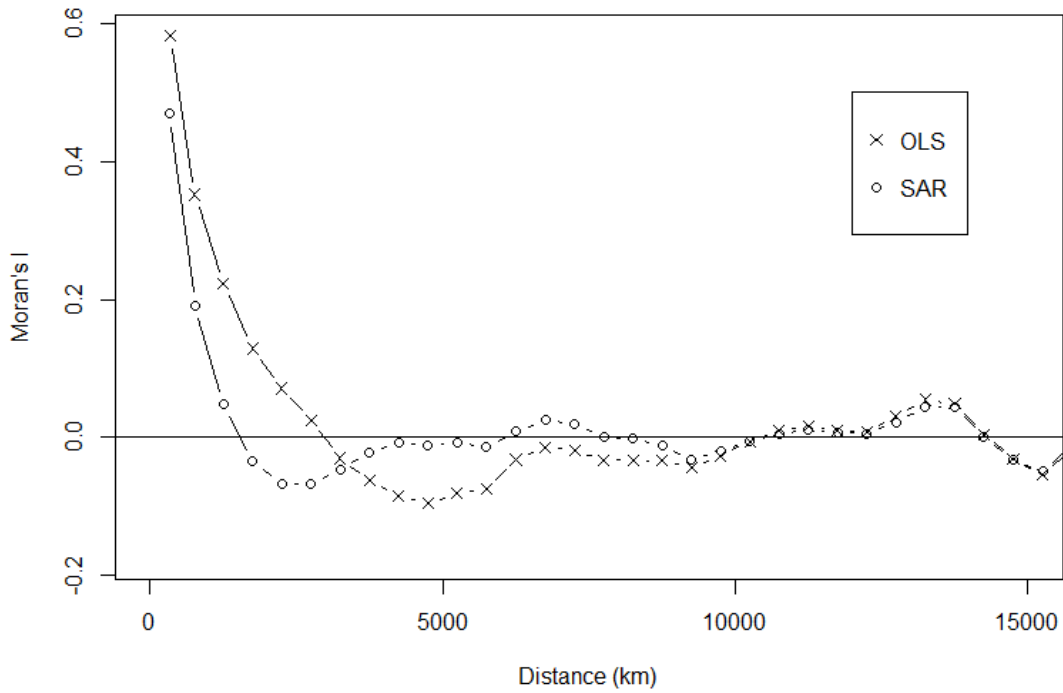


Figure S4.2 Correlogram of residual extinction risk in assemblages in the validation set. OLS: Ordinary Least Squares model. SAR: Spatial Autoregressive model.

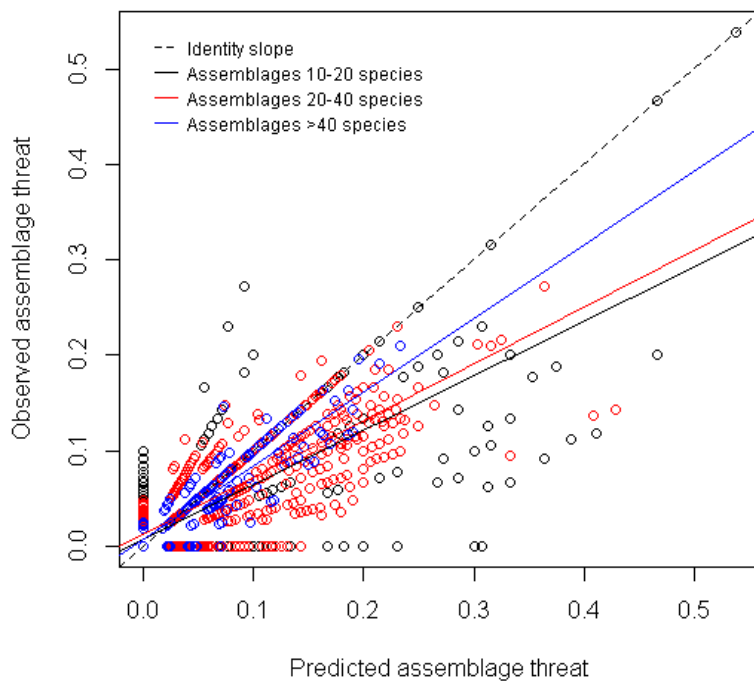


Figure S4.3 Observed and predicted assemblage threat across assemblage sizes in the global validation set (n=4,505).

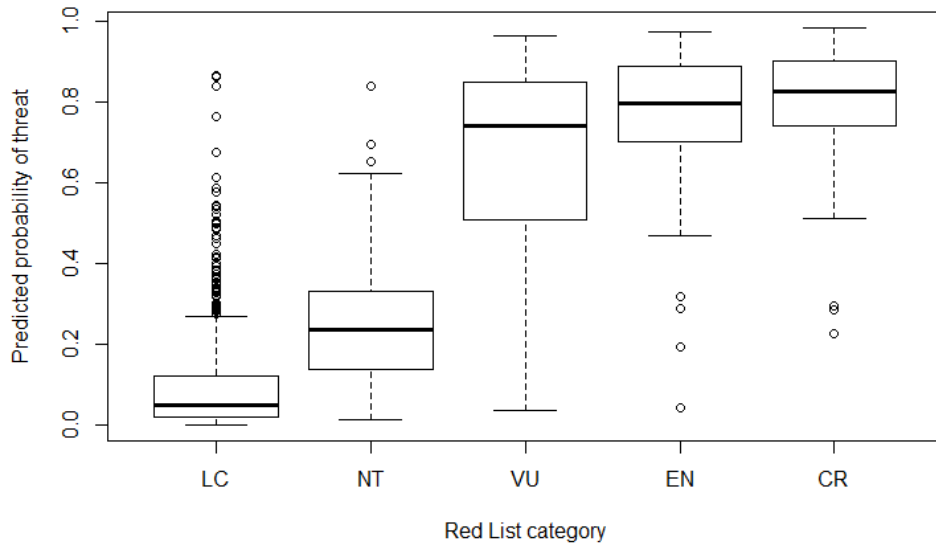


Figure S4.4 Predicted probability of threat from the global model against Red List category in the global validation set (n=991).

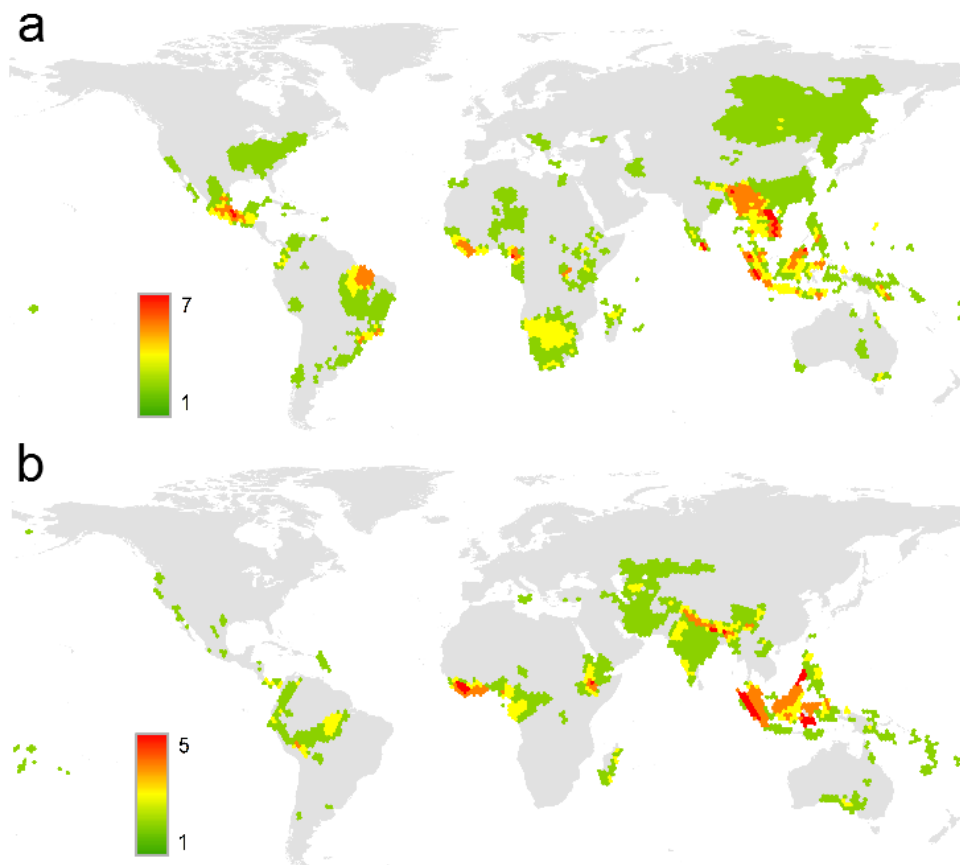


Figure S4.5 Number of false negative (a) and false positive (b) species classification in the global validation set (n=991). False negatives are threatened species incorrectly predicted as non-threatened. False positives are non-threatened species incorrectly predicted as threatened.

Appendix IV

Databases

I collated species trait data on four taxonomic groups: terrestrial mammals (hereafter, mammals), amphibians, reptiles and crayfish. All data were 100 % complete for species included in the modelling. For mammals I used the trait database in Chapter 4 (Table S4.1). I collected information for 478 amphibian species from Bielby *et al.* (2008) and Cooper *et al.* (2008) (Table S5.1a).

A trait database on 1,500 reptile species assessed by the Sampled Red List (Böhm *et al.* 2013) was compiled in collaboration with the National Autonomous University of Mexico, Stony Brook University, Nature Serve and the Institute of Zoology (Table S5.1b). Data were collected from species descriptions, field guides, museum specimens and published life-history studies (references available upon request), and supplemented with data obtained from species experts during the IUCN Red List assessment process. I selected the following life-history and ecological variables: maximum body size (snout-vent length), reproductive mode, habitat mode, trophic level, continental presence and number of IUCN-listed habitats. I recorded the geographical range size and latitude of range centroid for each species (Böhm *et al.* 2013). I extracted mean values in the species' range of 12 environmental and anthropogenic threat variables (Table S5.1b) in ArcGIS 9.2.

I collected information on the 586 freshwater crayfish species recently assessed by the IUCN (IUCN 2010). I collected maximum body size from species descriptions, field guides and museum specimens (references available upon request). I used maximum body size as mean body size is generally not available for crayfish species. I found three measures of crayfish body size: occipital carapace length (OCL), carapace length (CL), and body length (BL). I used CL as my preferred measure of body size as it was available for most species (397 species). For species for which maximum CL was missing, I preferentially transformed OCL values, as crayfish BL is a more variable measure of crayfish body size than OCL. I corrected OCL into CL for 83 species and BL into CL for 106 species. I developed correction factors between OCL and CL, and BL and CL from a database of morphological measurements of 1,743 specimens. These measurements were obtained from species descriptions, museum plates, museum specimens and field specimens (references available upon request). I used species-specific correction factors when available, if not I used genus-specific correction factors (29 species). I had no maximum body size information for four species. I followed

Adamowicz & Purvis (2006) and assigned species to four habitat types: 1) streams, 2) temporary or standing waters, 3) burrows, and 4) caves. I used data from Adamowicz & Purvis (2006), IUCN assessments (IUCN 2010), field guides and species descriptions. As some species display habitat flexibility, I derived a measure of habitat specialisation from the number of IUCN-listed habitats occupied by each species (IUCN 2010). I recorded the geographical range size and latitude of range centroid for each species (IUCN 2010). I extracted values in the species' range of 12 environmental and anthropogenic threat variables (Table S5.1c) in ArcGIS 9.2.

Table S5.1 Species trait data included in the models of extinction risk in reptiles, amphibians and crayfish.

a) Amphibians

Variables	Unit	Source	Resolution
Taxonomy (Family, Genus)		Bielby <i>et al.</i> 2008	
Body size	Maximum snout-vent length (millimetres)	Bielby <i>et al.</i> 2008	
Aquatic life stage	Yes, No	Bielby <i>et al.</i> 2008	
Habitat breadth		Cooper <i>et al.</i> 2008	
Tropical distribution	Yes, No	Cooper <i>et al.</i> 2008	
Range size	km ²	Bielby <i>et al.</i> 2008	
Latitude of range centroid	Degrees latitude	Cooper <i>et al.</i> 2008	
Median isothermality	(Mean diurnal range/temperature annual range)*100	Bielby <i>et al.</i> 2008	30 arc second
Median maximum temperature of the warmest month	Degrees (°C)	Bielby <i>et al.</i> 2008	30 arc second
Median precipitation of the driest quarter	Millimetres	Bielby <i>et al.</i> 2008	30 arc second
Median precipitation seasonality	Coefficient of variation	Bielby <i>et al.</i> 2008	30 arc second
Median annual actual evapotranspiration	Millimetres	Bielby <i>et al.</i> 2008	30 arc second
Median annual net primary productivity (1976-2000)	Grams per m ² per year	Bielby <i>et al.</i> 2008	30 arc second
Median altitude	Meters	Bielby <i>et al.</i> 2008	30 arc second
Median human population density (2000)	People per km ²	Bielby <i>et al.</i> 2008	30 arc second

b) Reptiles

Variables	Unit	Source	Resolution
Taxonomy (Family, Genus)		Böhm <i>et al.</i> 2013	
Body size	Maximum snout-vent length (mm)	Species descriptions, museum specimens, literature, experts	
Reproductive mode	Oviparous, ovoviparous or viviparous	Species descriptions, museum specimens, literature, experts	
Trophic level	Carnivorous, Invertebrates, Herbivorous or Omnivorous	Species descriptions, museum specimens, literature, experts	
Habitat mode	Aquatic, Arboreal, Terrestrial, Saxicolous, Fossorial, Semi-aquatic, Semi-arboreal, Semi-fossorial, Semi-saxicolous	Species descriptions, museum specimens, literature, experts	
Number of IUCN habitats		Böhm <i>et al.</i> 2013	
Continent	Continent, Island, Both	Böhm <i>et al.</i> 2013	
Range size	km ²	Böhm <i>et al.</i> 2013	
Latitude of range centroid	Degrees latitude	Böhm <i>et al.</i> 2013	
Mean annual temperature	Degrees (°C)	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Mean temperature seasonality	Standard deviation*100	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Mean annual precipitation	Millimetres	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Mean precipitation seasonality	Coefficient of variation	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Mean annual net primary productivity (1976-2000)	Grams per m ² per year	Imhoff <i>et al.</i> 2004	0.25 degrees
Minimum elevation	Meters	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Elevation range	Meters	Hijmans <i>et al.</i> 2005	2.5 arc minutes
Mean human population density (2000)	People per km ²	CIESIN 2005a	2.5 arc minutes
Minimum human population density (2000)	People per km ²	CIESIN 2005a	2.5 arc minutes
Mean human footprint	Human Influence Index normalized per region and biome	CIESIN 2005b	2.5 arc minutes
Mean human appropriation of net primary productivity	Percent of NPP	Imhoff <i>et al.</i> 2004	0.25 degrees
Mean GDP (1990)	Dollars per person per year	CIESIN 2002	0.25 degrees

c) Crayfish

Variables	Unit	Source	Resolution
Taxonomy (Family, Genus)		IUCN 2010	
Body size	Maximum carapace length (mm)	Species descriptions and museum specimens	
Habitat type	Stream, Temporary or standing water, Burrow, Cave	Adamowicz & Purvis (2006), species descriptions and field guides	
Number of IUCN habitats		IUCN 2010	
Range size	km ²	IUCN 2010	
Latitude of range centroid	Degrees latitude	IUCN 2010	
Mean annual temperature	Degrees (°C)	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean temperature seasonality	Standard deviation*100	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean annual precipitation	Millimetres	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean precipitation seasonality	Coefficient of variation	Hijmans <i>et al.</i> 2005	30 arc seconds
Minimum elevation	Meters	Hijmans <i>et al.</i> 2005	30 arc seconds
Elevation range	Meters	Hijmans <i>et al.</i> 2005	30 arc seconds
Mean consumptive water loss	CDF-standardized water consumption through irrigation, thermoelectric and manufacturing industries divided by contemporary discharge	Vorosmarty <i>et al.</i> 2010	30 arc seconds
Mean wetland disconnectivity	CDF-standardized proportion of wetland occupied by cropland or impervious surface area	Vorosmarty <i>et al.</i> 2010	30 arc seconds
Mean river fragmentation	CDF-standardized proportion of each drainage basin that is accessible from a given grid cell	Vorosmarty <i>et al.</i> 2010	30 arc seconds
Mean mercury deposition	CDF-standardized difference between present-day and pre-industrial Hg deposition	Vorosmarty <i>et al.</i> 2010	30 arc seconds
Mean pesticide loading	CDF-standardized country-based pesticide application to croplands	Vorosmarty <i>et al.</i> 2010	30 arc seconds
Mean sediment loading	CDF-standardized total suspended solids	Vorosmarty <i>et al.</i> 2010	30 arc seconds

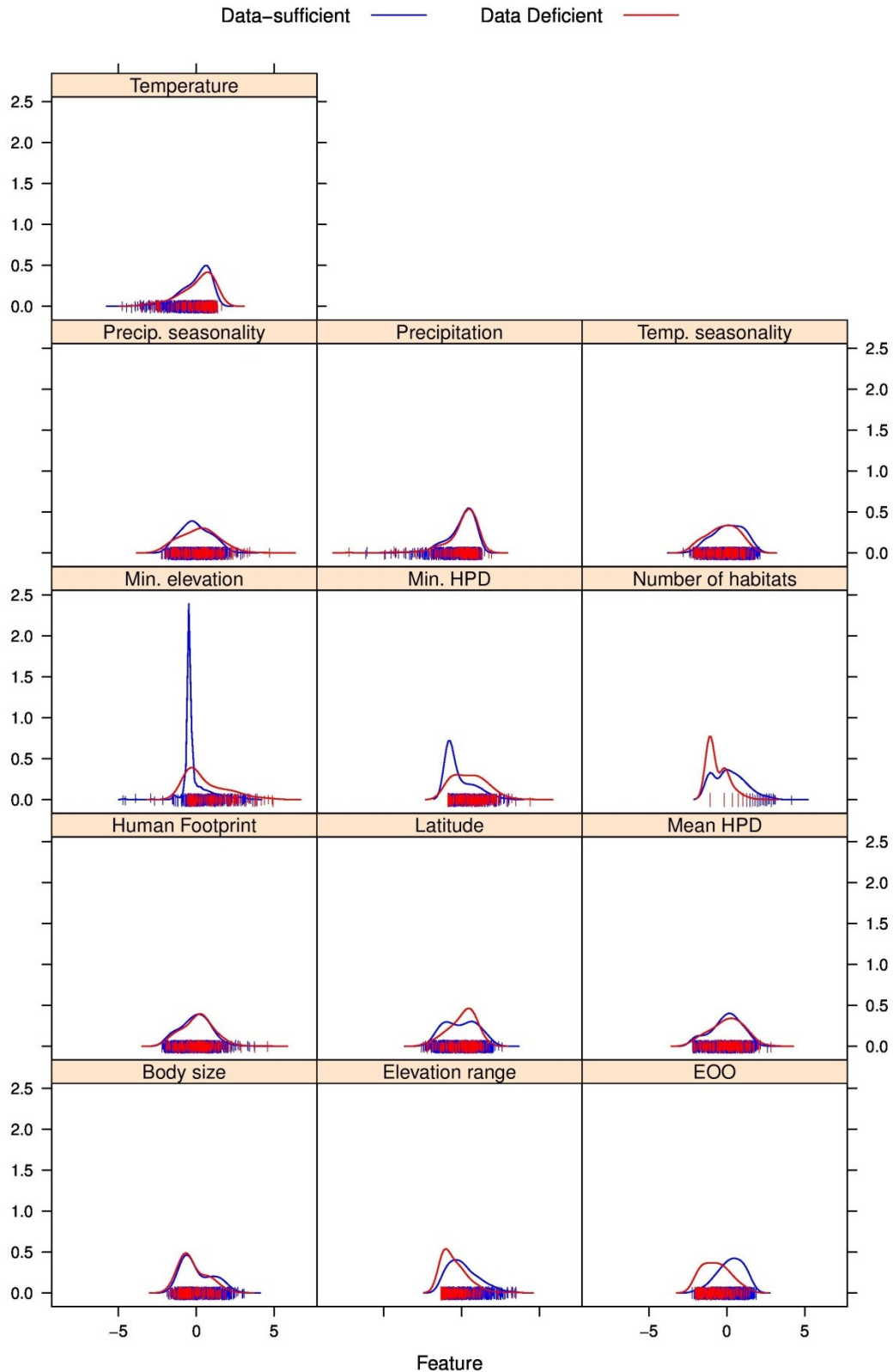


Figure S5.1 Overlay density plot of numeric explanatory variables in Data Deficient (n=229) and data-sufficient species (n=982) reptiles. HPD: human population density. Temp: temperature. Precip: precipitation. Min: minimum.

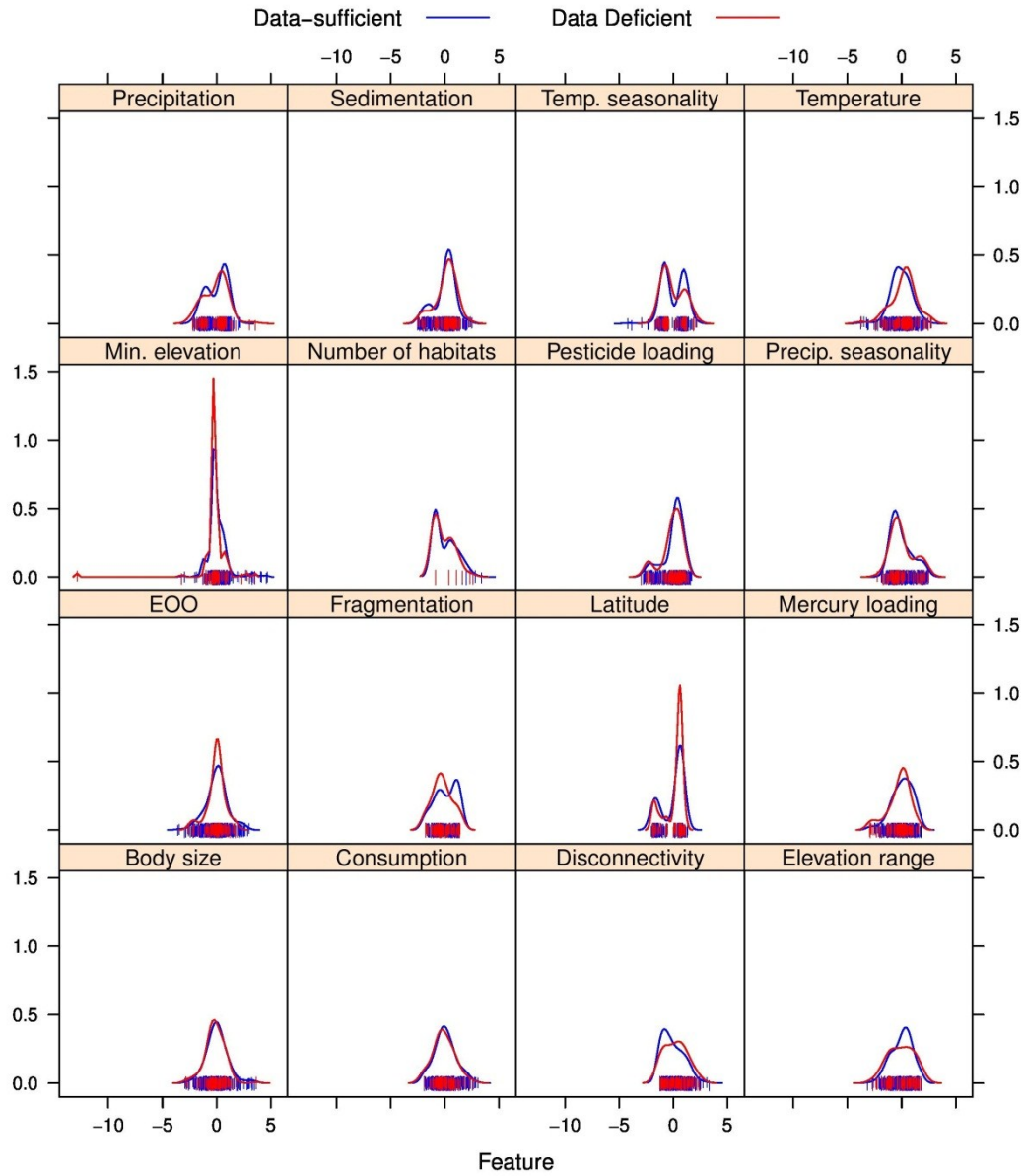


Figure S5.2 Overlay density plot of numeric explanatory variables in Data Deficient (n=118) and data-sufficient species (n=440) crayfish. Temp: temperature. Precip: precipitation. Min: minimum.

Table S5.2 AUC for amphibian and crayfish models with and without data pre-processing (encoding of multinomial categorical variables as dummy variables). AUC: area under the receiver operating characteristic curve.

	With pre-processing	Without pre-processing
Amphibians		
Classification tree	0.828	0.898
Random forest	0.916	0.953
Boosted trees	0.892	0.949
Decision stump	0.842	0.842
Crayfish		
Classification tree	0.813	0.874
Random forest	0.892	0.919
Boosted trees	0.877	0.927
Decision stump	0.698	0.698

Table S5.3 Optimal tuning parameters for models of extinction risk in mammals, reptiles, amphibians and crayfish. CT: classification tree, RF: random forests, BT: boosted trees, KNN: k-nearest neighbours, SVM: support vector machines, NNET: neural networks.

	CT	RF	BT			KNN	SVM	NNET		
	Cost parameter	Number of variables randomly sampled	Number of trees	Tree depth	Learning rate	Number of neighbours	Sigma inverse kernel width	Cost of constraint violation	Number of units in the hidden layer	Weight decay
Full models										
Mammals	0	6	212	11	0.1	29	0.0244	1	1	0.1
Reptiles	0.0437	6	188	7	0.01	51	0.0306	1	1	0.1
Amphibians	0	6	29	9	0.1					
Crayfish	0.0029	11	181	19	0.01					
Coarsened range size										
Mammals	0	3	54	14	0.1	29	0.0248	1	3	0.1
Reptiles	0.00215	3	467	8	0.01	57	0.0306	0.5	31	0.0611
Amphibians	0	9	33	14	0.1					
Crayfish	0.0015	9	94	4	0.05					

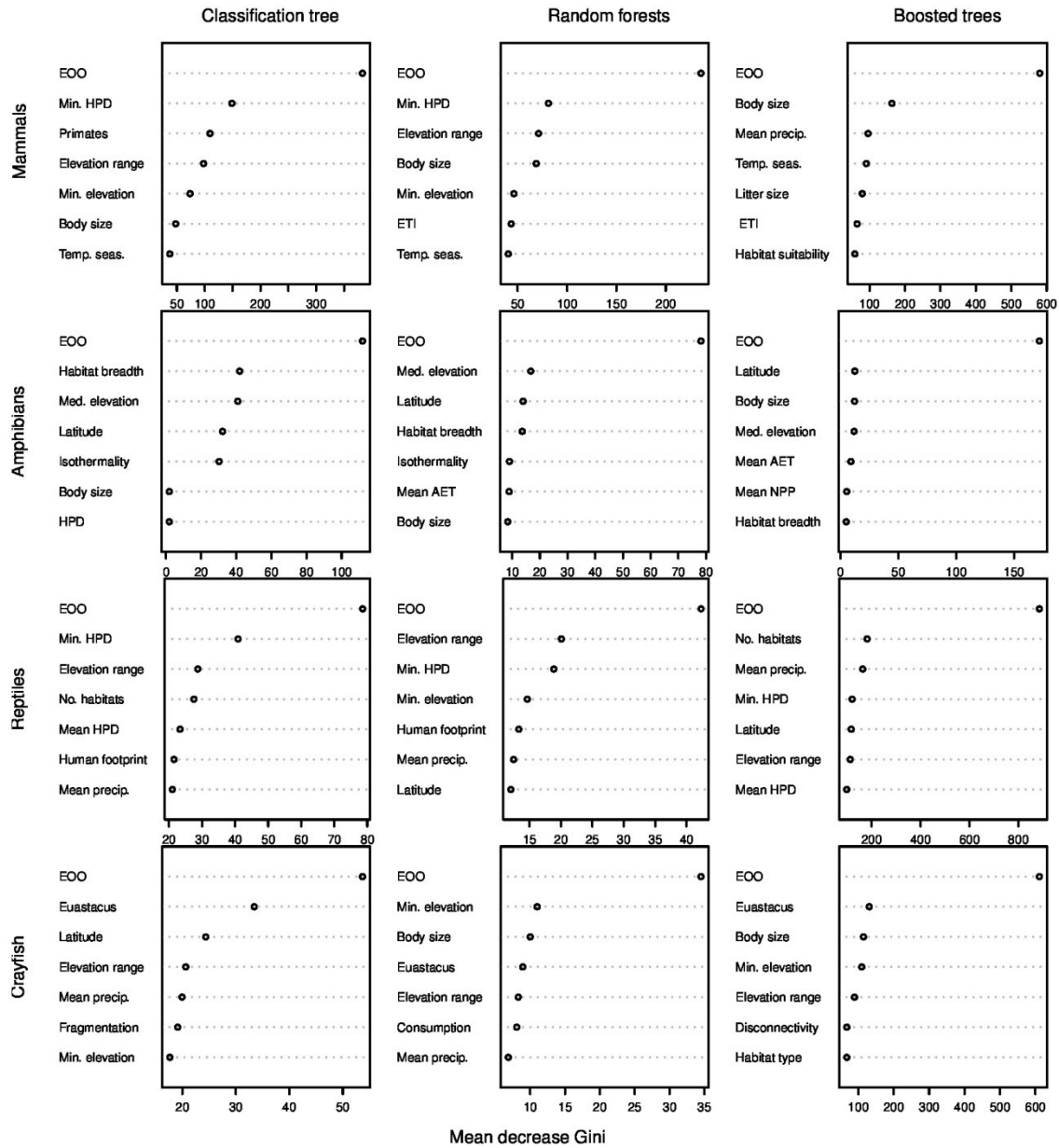


Figure S5.3 Variable importance in classification trees, random forests and boosted trees among taxonomic groups. Data are shown for the seven most important variables in models calibrated on fine geographical range size. Variable importance is determined by mean decrease in Gini index (node impurity). EOO: extent of occurrence. ETI: estimated threat index. HPD: human population density. Precip: precipitation. Temp: temperature. AET: actual evapo-transpiration. NPP: net primary productivity. Seas: seasonality. Med: median. Min: minimum. No: number of.

Cost ratio R

I outline the approach to computing the cost ratio R (Table S5.4). I converted all costs in pounds sterling to American dollars with a conversion rate of 1.55 (as of 12/08/2013). Assigning a DD species to a data-sufficient category relies on collecting information to a level suitable for the application of IUCN Red List criteria and re-assessment by the IUCN. I derived monitoring costs from published literature, grant reports and personal communications (Table S5.5). I computed three survey costs for mammals in three bins of range sizes: 0-10,000 km²: \$25,000; 10,000-100,000 km²: \$50,000; 100,000-8,000,000 km²: \$80,000. These were respectively represented by 35.5%, 48.3%, and 16.2% of species. I derived an average cost per DD mammal of \$45,994 by weighting the three cost bins by their relative contribution. There is less variation in range size in DD amphibians, reptiles and crayfish; hence I assigned them a survey cost of \$25,000. I derived IUCN Red List assessments costs from published sources (Stuart *et al.* 2010) and consultation with IUCN assessors (mammals: B. Collen, reptiles: M. Böhm, amphibians; A. Angulo, crayfish: N. Richman).

I computed the costs of predicting species conservation status with predictive models. Model building involves the following stages: collection of species trait data, GIS extractions of species range maps, data preparation, and Machine Learning model calibration and interpretation. The cost of the panTHERIA database for mammals was estimated at £700,000 (\$1,085,000) for 44 variables across 5,416 species. I used only four life-history variables; hence the cost per variable per species is \$18.2. The value is a low estimate as the costs of collecting extra data are not additive, but likely to diminish after a certain number of variables. Hence I computed the cost of collecting all variables per species (\$200.3). I also computed a middle cost estimate (\$24.4) based on the cost of crayfish life-history variables, as I personally collected those. I derived cost estimates for the GIS extractions of 12 maps across 3,967 species, with a mean time per map per species of 0.0007 hour. I assumed that such work would be conducted by a postgraduate research assistant, and derived costs for such staff from the Institute of Zoology salary schemes (£16.6 per hour). An analogous salary scheme for University College London is freely available at:

http://www.ucl.ac.uk/hr/salary_scales/final_grades.php . I computed the cost of data preparation for terrestrial mammals (16 hours) from the hourly staff cost of a postdoctoral researcher at the Institute of Zoology (£18.8 per hour). I computed the cost of Machine Learning model calibration, interpretation and report drafting (40 hours) from the staff cost of a postdoctoral researcher.

I repeated these costing methods for amphibians, reptiles and crayfish, with minor differences. Amphibian data (life-history and geographical variables) for 478 species were collected during a Masters project (J. Bielby, *pers. comm.*): I costed these as three months (528 hours) by a postgraduate research assistant (\$28.4 per species). I computed the cost of data preparation (8 hours by a postdoctoral researcher: \$0.5 per species), and computed the cost of Machine Learning model calibration, interpretation and report drafting (24 hours by a postdoctoral researcher: \$1.5 per species). Reptile life-history data were collected in a collaborative effort among the National Autonomous University of Mexico, Stony Brook University, Nature Serve and the Institute of Zoology. Collection of life-history data for 660 squamate species during two Masters projects at the Institute of Zoology ran over 11 weeks (440 hours): extrapolating these figures to the 982 species used in modelling corresponds to 655 hours by a postgraduate research assistant (\$17.1 per species). I computed the cost of data preparation (8 hours by a postdoctoral researcher: \$0.24 per species), and computed the cost of Machine Learning model calibration, interpretation and report drafting (24 hours by a postdoctoral researcher: \$0.71 per species). I collected crayfish life-history data for 558 species over two months (352 hours; \$16.2 per species). I computed the cost of GIS layer extraction (\$0.22 per species), data preparation (\$0.4 per species), and computed the cost of Machine Learning model calibration, interpretation and report drafting (\$1.3 per species). For all groups, I computed the cost of building models based on geographical range size alone based on 2 hours for a postgraduate research assistant (cost per species: mammals: \$0.02; amphibians: \$0.1; reptiles: \$0.05; crayfish: \$0.09).

Table S5.4 Calculation of the cost ratio R in mammals, amphibians, reptiles and crayfish. The three scenarios in the cost ratio calculations for mammals are: (1) cost of four variables from the panTHERIA database, (2) cost of all variables from the panTHERIA database and (3) cost of life-history data identical to crayfish. RA: research assistant. PD: postdoctoral researcher.

a) Mammals

Risk assessments (c₁)	Range size (km ²)	Cost per species (\$)	Number of species	Cost per species (\$)	Scenario
	0-10 ³	25,000	175		
Field surveys	10 ³ -10 ⁶	50,000	238	45,994 (mean)	
	10 ⁶ - 10 ⁸	80,000	80		
IUCN Red List assessments				900	
Total cost				46,893	
Predictive models (c₂)	Number of variables	panTHERIA cost (\$)	Number of species		
Life-history variables	44	108,500	5,416		
				18.2	(1)
				200.3	(2)
				24.4	(3)
	Total time (hours)	Cost per hour (RA or PD) (\$)	Number of species		
GIS layer extraction	33.3	25.7	3,967	0.22	
Data cleaning	16	29.2	3,967	0.1	
Model calibration and report writing	40	29.2	3,967	0.3	
Total cost				18.8	(1)
				200.9	(2)
				24.9	(3)
Cost ratio (c₁/ c₂)					
				2,489	(1)
				233	(2)
				1,877	(3)

b) Amphibians

Risk assessments (c_1)				Cost per species (\$)
Field surveys				25,000
IUCN Red List assessments				400
Total cost				25,400
Predictive models (c_2)	Total time (hours)	Cost per hour (RA or PD) (\$)	Number of species	
Data collection	528	25.7	478	28.4
Data cleaning	8	29.2	478	0.5
Model calibration and report writing	24	29.2	478	1.5
Total cost				30.4
Cost ratio (c_1/ c_2)				836

c) Reptiles

Risk assessments (c_1)				Cost per species (\$)
Field surveys				25,000
IUCN Red List assessments				230
Total cost				25,230
Predictive models (c_2)	Total time (hours)	Cost per hour (RA or PD) (\$)	Number of species	
Data collection	655	25.7	982	17.1
GIS layer extraction	5.5	25.7	982	0.3
Data cleaning	8	29.2	982	0.24
Model calibration and report writing	24	29.2	982	0.71
Total cost				18.35
Cost ratio (c_1/ c_2)				1,375

d) Crayfish

Risk assessments (c_1)				Cost per species (\$)
Field surveys				25,000
IUCN Red List assessments				100
Total cost				25,100
Predictive models (c_2)	Total time (hours)	Cost per hour (RA or PD) (\$)	Number of species	
Life-history variables	352	25.7	558	16.2
GIS layer extraction	3.12	25.7	558	0.22
Data cleaning	8	29.2	558	0.4
Model calibration and report writing	24	29.2	558	1.3
Total cost				18
Cost ratio (c_1/ c_2)				1,401

Table S5.5 Cost of species field surveys, derived from published literature, grant reports and personal communications.

	Distribution	IUCN Red List status	Geographical range size (km ²)	Study type	Value	Source
Mammals						
15 mammals	Global	NA	1-1,000,000	Rediscovery	\$8,696 (researchers in same country) \$32,541 (researchers in another country)	Fisher (2011), <i>Biological Conservation</i> , 144(5), 1712-18
<i>Santamartamys rufodorsalis</i> (Red-crested tree Rat)	Colombia	DD (CR in 2010 Red List)	121	Field survey	\$5,000	Mohamed bin Zayed Species Conservation Fund
<i>Lepilemur sahamalazensis</i> (Sahamalaza sportive lemur)	Madagascar	DD	9,527	Field survey	\$10,000	Mohamed bin Zayed Species Conservation Fund
<i>Nyctophilus sherrini</i> (Tasmanian long-eared bat)	Australia	DD	48,861	Field survey	\$15,000	Mohamed bin Zayed Species Conservation Fund
Four rodents and a squirrel of Sulawesi	Sulawesi	DD		Distribution survey and occupancy modelling	\$20,000	Alessio Mortelliti, Australian National University
<i>Taeromys arcuatus</i>			415			
<i>Taeromys microbullatus</i>			395			
<i>Maxomys dollmanni</i>			736			
<i>Ratus salocco</i>			1,400			
<i>Prosciurillus abstrusus</i>			1,530			
<i>Salanoia durrelli</i> (Durrell's vonsira)	Madagascar	NE	NA	Field survey	\$12,000	Mohamed bin Zayed Species Conservation Fund
<i>Euchoreutes naso</i> (Long-eared jerboa)	China; Mongolia	LC	1,675,132	Field survey	\$9,820	Jon Bielby, Zoological Society of London
<i>Choeropsis liberiensis</i> (Pygmy hippo)	Côte d'Ivoire; Guinea; Liberia; Sierra Leone	EN	143,638	Distribution survey	\$48,000	Ben Collen, University College London
<i>Choeropsis liberiensis</i> (Pygmy hippo)	Côte d'Ivoire; Guinea; Liberia; Sierra Leone	EN	143,638	Population trend survey	\$180,000	Ben Collen, University College London
<i>Choeropsis liberiensis</i> (Pygmy hippo)	Côte d'Ivoire; Guinea; Liberia; Sierra Leone	EN	143,638	Threat survey	\$7,000	Ben Collen, University College London
<i>Loxodonta africana</i> (African elephant)	Africa	VU	6,344,920	Field survey	\$400,000	Monitoring the Illegal Killing of Elephants
Armenian Myotis bats	Armenia	EN, DD	NA	Field survey	\$9,200	Astghik Ghazaryan & Rufford Foundation
<i>Aproteles bulmerae</i> (Bulmer's fruit bat)	Papua New Guinea	CR	59.7 km ²	Field survey	\$1,000,000	Ken Aplin, Smithsonian Institution

Rodents in Sulawesi	Sulawesi	DD	NA	Field survey	\$15,000 for 5 to 10 species, depending on location	Kevin Rowe, Museum Victoria
<i>Pteropus rufus</i> (Madagascar Fruit Bat)	Madagascar	VU	185,399	Field survey	\$100,000	Paul Racey, Aberdeen University
<i>Myotis csorbai</i> (Csorba's Mouse-eared Myotis)	Nepal	DD	33,432	Field survey	\$38,000	Sanjan Thapa, Small Mammals Conservation & Research Foundation
<i>Notoryctes caurinus</i> (Northwestern Marsupial Mole)	Australia	DD	451,245	Field survey	\$100,000	Joe Benshemesh, La Trobe University
<i>Notoryctes typhlops</i> (Southern Marsupial Mole)	Australia	DD	921,468	Field survey	\$100,000	Joe Benshemesh, La Trobe University
Amphibians and Reptiles						
<i>Osaecilia osae</i> (Caecilian)	Costa Rica	DD	468	Field survey	\$3,000	Mohamed bin Zayed Species Conservation Fund
<i>Boulengerula changamwensis</i> (Changamwe caecilian)	Malawi	DD	989	Field survey	\$3,000	Mohamed bin Zayed Species Conservation Fund
<i>Xenophrys parallela</i>	Indonesia	DD	10	Field survey	\$3,000	Mohamed bin Zayed Species Conservation Fund
<i>Bolitoglossa insularis</i> (Lungless salamander)	Nicaragua	NE	NA	Field survey and conservation	\$13,000	Mohamed bin Zayed Species Conservation Fund
More than 55 threatened amphibian species	Madagascar	VU, EN, CR	NA	Field survey	\$300,000 (~\$5,400 per species)	Sahonagasy Action Plan
Vietnamese amphibians	Vietnam	VU, EN, CR	NA	Acoustic survey	\$5,000	Jodi Rowley, Australian Museum
DD amphibians	Global	DD	NA	Population trend monitoring (up to 10 years)	4,000-\$20,000 per year	Peter Paul van Dijk, IUCN/SCC Tortoise & Freshwater Turtle Specialist Group
DD tortoises and freshwater turtles	Global	DD	NA	Population trend monitoring (up to 30 years)	\$20,000-\$50,000 per year	Peter Paul van Dijk, IUCN/SCC Tortoise & Freshwater Turtle Specialist Group
Crayfish						
North American crayfish	United States	DD	NA	Field survey	\$25,000-\$100,000	Jim Fetzner, Carnegie Museum
North American and Australian crayfish	United States and Australia	DD	NA	Field survey	\$10,000+	Zachary Loughman, West Liberty University

Case study

I choose n , the number of species assessed with risk assessments and N , the number of species assessed with predictive models, to minimize the variance in the estimation of the proportion of threatened species in the Data Deficient sample. I minimize the measurement costs subject to the constraint that the variance of the proportion of threatened species is less than the variance of a binomial estimate based on n_v error-free risk assessments. K is the coefficient of reliability between risk assessments and predictive models.

Following Tenenbein (1970):

$$n = \frac{-Nn_v(1-K)}{n_vK-N} \quad \text{Equation S5.1}$$

In this example, I set $n_v = 100$. I consider the determination of n and N for terrestrial mammals with the best predictive model and the minimum cost ratio between risk assessments and predictive models. The random forest model achieved a coefficient of reliability K of 0.7. I consider the minimum cost ratio $R = 250$, with the cost of risk assessments $c_1 = \$50,000$ and cost of predictive models $c_2 = \$200$. The value of N is 677, the number of Data Deficient mammal species.

$$n = \frac{-677 * 100(1 - 0.7)}{100 * 0.7 - 677} = 33.45 \approx 34 \text{ species}$$

The cost C of the double sampling scheme is:

$$C = c_1n + c_2N = \$1,848,600$$

The cost C_v of the single sampling scheme is:

$$C_v = c_1n_v = \$5,000,000$$

The reduction in cost under double sampling is:

$$\lambda = 1 - \frac{C}{C_v} = 0.633$$

Which corresponds to a reduction in cost of 63.3%.

Application of double sampling to small population sizes

The expected variance of \hat{p} (the proportion of threatened species in the sample) as calculated by Tenenbein (1970) relies on a binomial sampling distribution for an infinite population.

With n the sample size and p the proportion of threatened species in the sample, the expected variance for a binomial distribution is:

$$V(\hat{p}) = \frac{p(1-p)}{n} \quad \text{Equation S5.2}$$

However, populations of Data Deficient species are finite, e.g. 677 Data Deficient mammals and 125 Data Deficient crayfish. If we assess all species with error-free IUCN Red List assessments, the variance in the estimated proportion of threatened DD species is zero. The hypergeometric distribution is more appropriate than a binomial distribution when assessing a finite population, as the expected variance \hat{p} decreases faster as the sample size approaches the population size. With N the total population size, the expected variance for hypergeometric distribution is:

$$V(\hat{p}) = \frac{p(1-p)}{n} * \frac{N-n}{N-1} \quad \text{Equation S5.3}$$

The variance for the double sampling strategy is a weighted average of the variance of a binomial estimate of p based on n true measurements, and the variance of a binomial estimate of p based on N fallible measurements (Tenenbein, 1970). With K the coefficient of reliability for the fallible measurements, the expected variance for double sampling with a binomial distribution is:

$$V(\hat{p}) = \frac{pq}{n} * (1 - K) + \frac{pq}{N} * K \quad \text{Equation S5.4}$$

From exploratory analyses, I observe that the difference in variance between single and double sampling changes as a function of the sample size and the population size (Figure S6). When the sample size is small, the estimate variance for sampling with a hypergeometric distribution (green line; Figure S5.5) approximates the variance for binomial sampling (red line; Figure S5.5). As the sample size approaches the population size, the variance decreases faster for a hypergeometric distribution. Under these conditions, single sampling variance for a given budget can switch from being above the double sampling estimate to below (Figure S6d). For example, as of 2013 1,629 amphibian species are listed as DD, including 361 in the Indomalayan realm (IUCN, 2013). Under a budget of 188 units (188 species), double sampling is more cost-effective than single sampling with a binomial distribution (Figure S5.5d). Double sampling with a binomial distribution remains more cost-effective than hypergeometric single sampling for a population of 1,629 Data Deficient

species. However, for a population of 361 Data Deficient species binomial sampling is no longer cost-effective, as the variance of single hypergeometric sampling is lower than the variance for double binomial sampling.

To date the variance for the double sampling estimate based on a hypergeometric distribution has not been mathematically determined, but we may expect a lower variance compared to binomial sampling, given a correction factor for population size analogous to single hypergeometric sampling. As a consequence, for a given population size of DD species a comparison between single sampling with a hypergeometric distribution and double sampling with a binomial distribution provides an upper boundary of the sample size under which double sampling is known to be cost-effective. I determined such boundaries for a range of K , R , and total population sizes of DD species (Table S5.6). For example, with $K = 0.7$, $R = 3,000$, $p = 0.25$ and a total population size of 500 DD species, double sampling is cost-effective if I can only afford to red list fewer than 342 species (Table S5.6).

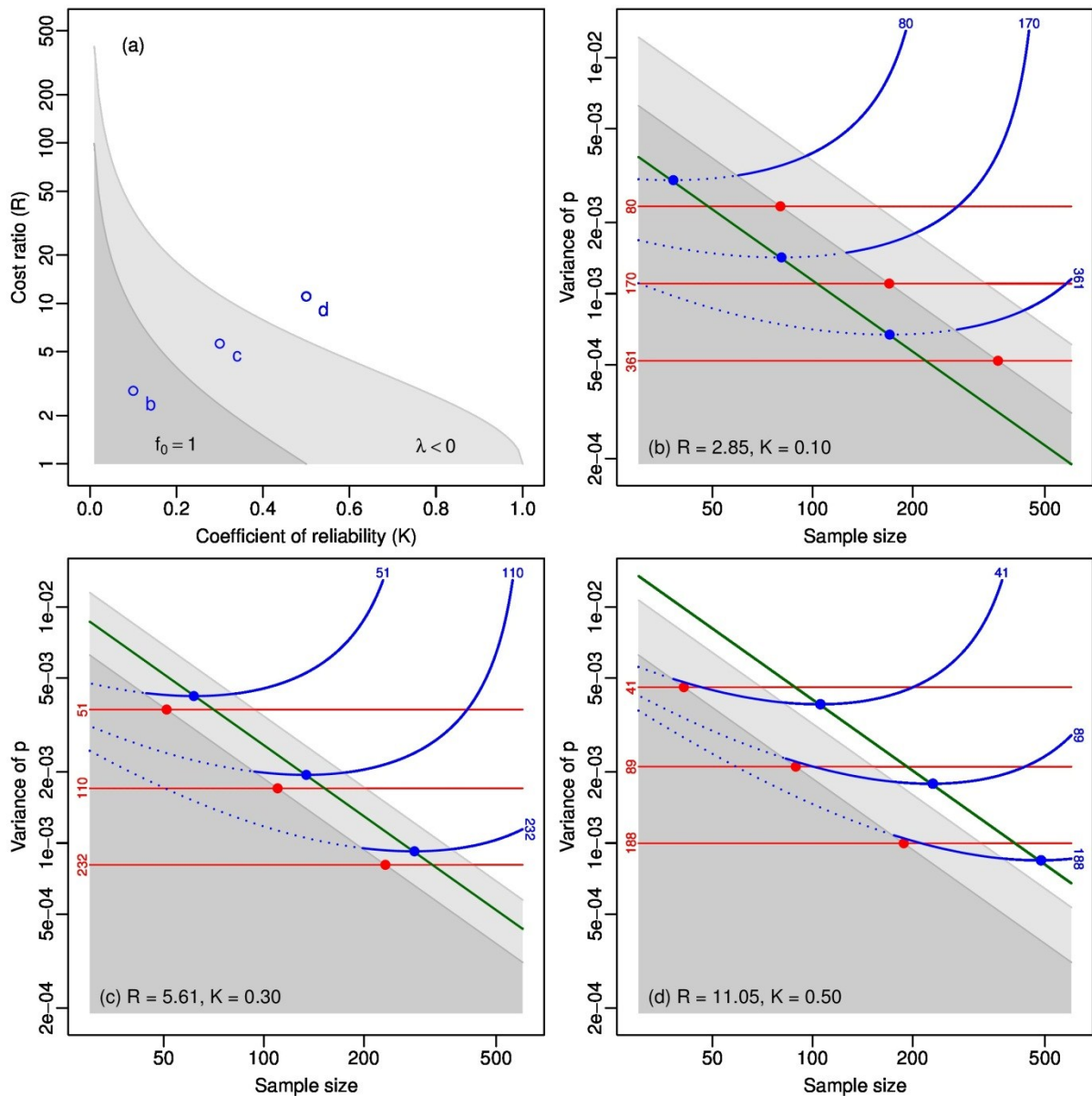


Figure S5.4 (a) Key regions for assessment of double sampling, given model reliability (K) and cost ratio (R), showing $f_0=1$ (darker grey) and $\lambda < 0$ (lighter grey). Panels (b) to (d) show how the variance of estimated p changes with sample size for each of the combination of K and R shown. Each panel shows the region where $f_0=1$ (dark grey) and $\lambda < 0$ (light grey). The line showing variance at a given sample size for the optimum f_0 (dark green) intercepts contours of equal cost under double sampling with varying f_0 (blue) at the points of minimum variance for a given budget (solid blue dots). The variance estimates from single sampling for each of the two cost budgets are shown as red dots and horizontal lines. (b) There is no valid double sampling solution ($f_0=1$) (c) There are valid double sampling solutions, but for a given budget, the estimate of p (the proportion of threatened DD species) is less precise than can be achieved by single sampling with the same budget (d) There are valid solutions and they can yield better estimates of p for a given budget than single sampling alone.

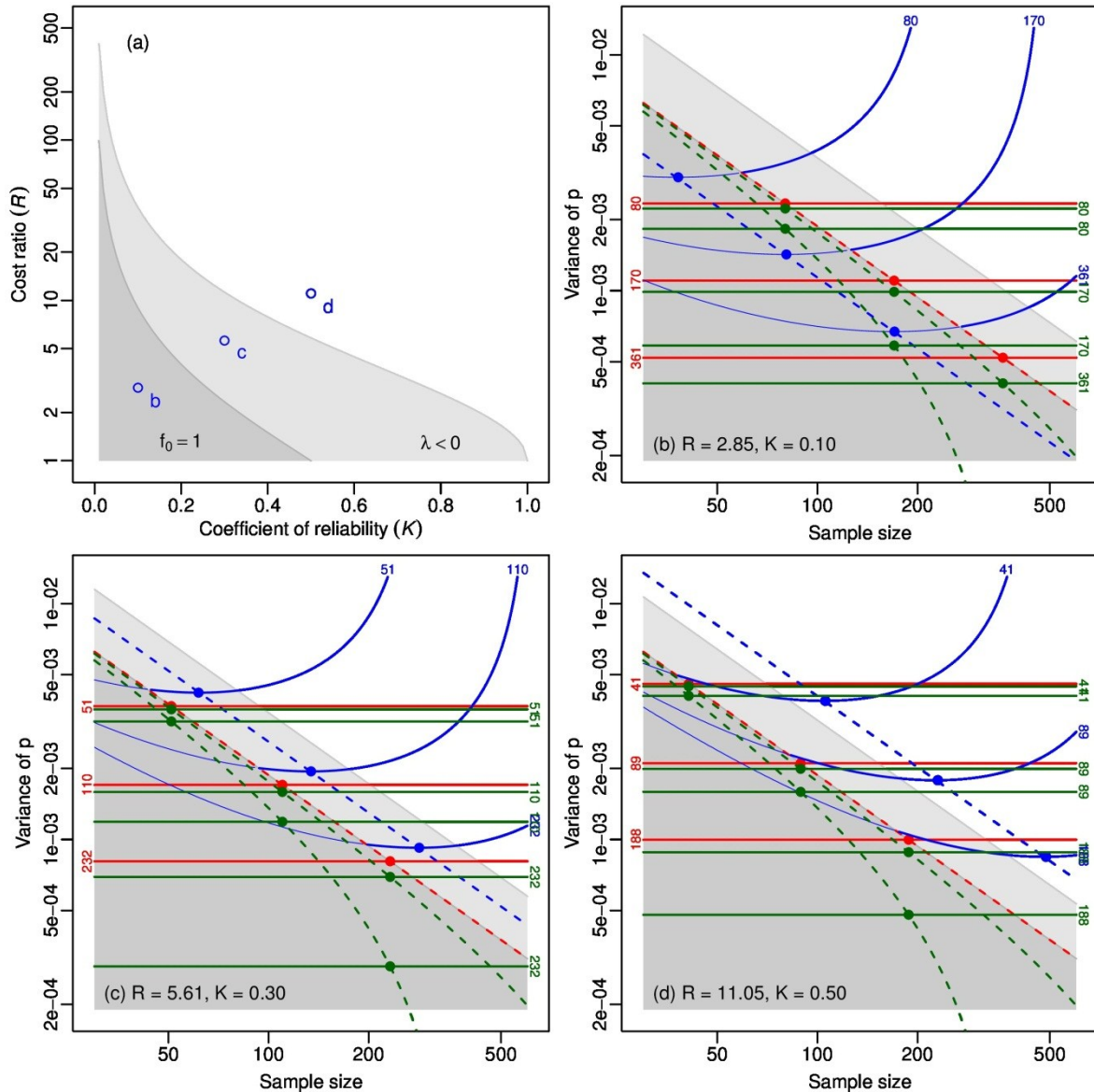


Figure S5.5 (a) Key regions for assessment of double sampling, given model reliability (K) and cost ratio (R), showing $f_0=1$ (darker grey) and $\lambda < 0$ (lighter grey). Panels (b) to (d) show how the variance of estimated p changes with sample size for each of the combination of K and R shown. Each panel shows the region where $f_0=1$ (dark grey) and $\lambda < 0$ (light grey). The line showing variance at a given sample size for the optimum f_0 (dark green) intercepts contours of equal cost under double sampling with varying f_0 (blue) at the points of minimum variance for a given budget (solid blue dots). The variance estimates from single sampling with a binomial distribution for each of the three cost budgets are shown as red dots and horizontal lines. The variance estimates from single sampling with a hypergeometric distribution for each of the three cost budgets as shown as green dots and horizontal lines. The fast-declining dotted green line indicates a population size of 361 DD species; the slow-declining dotted green line a population size of 1,629 DD species. (b) There is no valid double sampling solution ($f_0=1$) (c) There are valid double sampling solutions, but for a given budget, the estimate of p (the proportion of threatened DD species) is less precise than can be achieved by single sampling with the same budget (d) There are valid solutions and they can yield better estimates of p for a given budget than single sampling alone.

Table S5.6 Sample size (n) for which single sampling with a hypergeometric distribution is more cost-effective than double sampling with a binomial distribution, shown for a range of total population sizes of Data Deficient species (N), coefficients of reliability (K) and cost ratios (R). The proportion of threatened species is set at $p = 0.25$. The sample size (n) provides an upper boundary for the use of double sampling with a binomial distribution; double sampling with a hypergeometric distribution may prove cost-effective for higher sample sizes (budgets).

Population size (N)	K	R	Sample size (n)
250	0.4	250	84
250	0.7	250	161
250	0.4	3000	97
250	0.7	3000	172
500	0.4	250	169
500	0.7	250	320
500	0.4	3000	192
500	0.7	3000	342
1000	0.4	250	338
1000	0.7	250	640
1000	0.4	3000	383
1000	0.7	3000	684