Cost-effective assessment of extinction risk levels with limited information

Standard Paper

Lucie M. Bland¹²,²*, C. David L. Orme²†, Jon Bielby³§, Ben Collen⁴‡, Emily Nicholson¹¶, Michael A. McCarthy³§

¹ School of Botany, The University of Melbourne, Parkville, Victoria 3010, Australia
² Division of Biology, Imperial College London, Silwood Park, Ascot, SL5 7PY, UK
³ Institute of Zoology, Zoological Society of London, Regent’s Park, London, NW1 4RY, UK
⁴ Centre for Biodiversity and Environment Research, University College London, Gower Street, London, WC1 E6BT, UK

* l.bland@unimelb.edu.au, † d.orme@imperial.ac.uk, ‡ b.collen@ucl.ac.uk, § jon.bielby@ioz.ac.uk,
¶ emily.nicholson@unimelb.edu.au, \ mamcca@unimelb.edu.au

Corresponding author: Lucie M. Bland, l.bland@unimelb.edu.au, T: +61 (0)4 5112 9967

Running title: Cost-effective assessment of extinction risk

Number of words in the summary: 346
Number of words in the main text: 4,553
Number of words in the acknowledgments: 99
Number of words in the references: 1,847
Number of words in the tables: 600
Number of words in the figure legends: 144
Number of tables: 2
Number of figures: 1
Number of references: 45
Summary

1. Cost-effective reduction of uncertainty in global biodiversity indicators is a central goal of conservation. Comprising a sixth of the 74,000+ species currently on the IUCN Red List, Data Deficient species contribute to considerable uncertainty in estimates of extinction risk. Estimating levels of risk in Data Deficient species will require large resources given the costs of surveys and Red List assessments. Predicting extinction risk from species traits and geographical information could provide a cheaper approach for determining the proportion of Data Deficient species at risk of extinction.

2. We use double sampling theory to compare the cost-effectiveness of predictive models and IUCN Red List assessments for estimating risk levels in Data Deficient terrestrial mammals, amphibians, reptiles and crayfish. For each group, we calibrate Machine Learning models of extinction risk on species of known conservation status, and assess their cost and reliability relative to field surveys followed by Red List assessments.

3. We show that regardless of model type used or species group examined, it is always more cost-effective to determine the conservation status of all species with models and assess a small proportion of species with IUCN criteria (double sampling), rather than spend the same resources on field surveys and Red List assessments alone (single sampling).

4. We estimate that surveying and re-assessing all Data Deficient species currently listed on the IUCN Red List (12,206 species) with IUCN criteria would cost a minimum of US $323 million. Double sampling reduces the cost of determining the proportion of Data Deficient species at risk of extinction by up to 68%, because less than 6% of Data Deficient species would need to be surveyed and assessed with IUCN criteria.

5. Synthesis and applications. Double sampling with models cost-effectively estimates extinction risk levels in poorly-known species, and can be used to reduce the impact of uncertainty in the Red List and Red List Index. We provide recommendations for uptake by managers and a sampling planner spreadsheet. 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.

Keywords: Aichi biodiversity targets, Amphibians, Biodiversity indicators, Cost-effectiveness, Crayfish, IUCN Red List, Mammals, Reptiles.
Introduction

Global indicators of biodiversity status are central to monitoring progress towards the 2020 Aichi targets (Convention on Biological Diversity 2010), and assessing the outcomes 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 needs to inform conservation decisions cost-effectively (McDonald-Madden et al. 2010). Representativeness and reliability have been identified as desirable properties of successful 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 74,000 species assessed as of 2014 (IUCN 2014). However, one in six species on the Red List are too poorly known to assign to a category of extinction risk, and are listed as Data Deficient (DD). This gap in knowledge contributes to considerable uncertainty in global patterns of extinction risk (Bland et al. 2012) and conservation prioritization (Trindade-Filho et al. 2012). Re-assessment of the 12,206 species currently listed as DD to data-sufficient categories will require substantial financial resources, given the costs of biodiversity surveys (Gardner et al. 2008) and Red List assessments (Stuart et al. 2010). As a consequence, cost-effectively estimating risk levels in DD species is crucial to reducing uncertainty in the IUCN Red List and Red List Index (Bailie et al. 2008).

Comparative studies of extinction risk based on species trait data have yielded insight into the determinants of risk among groups (Purvis et al. 2000; Cardillo & Meijaard 2012), and could underpin a preliminary re-assessment of DD species (Davidson et al. 2009; Jones & Safi 2011). Good coverage of species’ trait data is available for a large number of DD species and includes life-history, ecological and phylogenetic information. The geographic distribution of many DD species is known, allowing inference of species’ geographical range size, environmental niche and exposure to
anthropogenic threats. These data alone are insufficient for making a decision on formal Red List status, but could be used to inform global estimates of 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 DD mammals (Bland et al. 2014).

Models may be cheaper to apply than collecting field-based data to update Red List assessments of Data Deficient species, but model predictions may be inaccurate and bias estimates of extinction risk. 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 DD species? We use sampling theory to answer this question. Specifically, we compare the variance in the estimated proportion of DD species threatened with extinction with two methods:

i) **Single sampling.** The proportion of DD species at risk of extinction is inferred from surveying and updating Red List assessments for a random subset of DD species.

ii) **Double sampling.** The same financial resources are shared between developing predictive models of extinction risk based on species data for all species, and updating Red List assessments for a smaller set of species. Given the relative costs of these two procedures and expected accuracy of model classifications, double sampling theory (Tenenbein 1970) identifies both the optimal allocation of funds to each process and the resulting variance in the estimated proportion of threatened species. If model development is sufficiently cheap and accurate, double sampling can give more precise estimates of risk prevalence than single sampling. Double sampling theory is frequently used in medical diagnostics (Zhou, McClish & Obuchowski 2002) and quality control (Poduri 2005), but few ecological applications exist (Harper et al. 2004; Rayner, Ellis & Taylor 2011).
In the context of this paper, species can be assessed as threatened or non-threatened as defined within the Red List categories (threatened: Vulnerable, Endangered and Critically Endangered. Non-threatened: Near Threatened and Least Concern; IUCN 2001), with Red List assessments or with predictive models of extinction risk. We 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, we calibrate Machine Learning models of extinction risk on species of known conservation status (data-sufficient species), and assess their reliability compared to Red List assessments. We compute the costs of field data collection and updating Red List assessments, and compare them with the costs of model development. We then devise the most cost-effective strategy for determining the proportion of DD species threatened with extinction in each group.

**Materials and methods**

**Double sampling**

We estimate the proportion of threatened species ($p$) and its associated variance with double sampling theory. In practice an investigator may wish to minimize the variance in the estimation of $p$ for a given budget; alternatively, she may wish to obtain a given variance in the estimation of $p$ for a minimum budget. Tenenbein (1970) derives identical solutions for these two problems; we outline his main results and refer the reader to Tenenbein (1970) for complete statistical derivations. We compare two estimates of variance of $p$:

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

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

First, the variance under single sampling ($V_s$) is the binomial variance: we 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 \( V_d \), we 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, we cannot afford to assess as many species.

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

To use this approach in practice, we derive:

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

The example below shows the cell probabilities, and a confusion matrix for the classification of 109 crayfish species in the validation set by assessments (rows) and the best model (columns):

<table>
<thead>
<tr>
<th></th>
<th>thr.</th>
<th>n.thr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>thr.</td>
<td>( p(1-\phi) )</td>
<td>( p\phi )</td>
</tr>
<tr>
<td>n.thr.</td>
<td>( q\theta )</td>
<td>( q(1-\theta) )</td>
</tr>
</tbody>
</table>

| thr. | 28  | 6   | 34  |
| n.thr.| 6  | 69  | 75  |
With thr.: threatened; n.thr.: non-threatened species. From this we calculate:

\[
p = \frac{34}{109} = 0.312, \quad q = \frac{75}{109} = 0.688, \quad \phi = \frac{6}{34} = 0.176, \quad \theta = \frac{6}{75} = 0.08, \quad \pi = \frac{34}{109} = 0.312, \quad \text{hence}
\]

\[
K = 0.555.
\]

ii) The costs of risk assessments \((c_1)\) and modelling \((c_2)\) per species, and their cost ratio \((R = c_1/c_2)\).

Below, we estimated these values from the cost of previous assessments and the combined costs of collating databases and modelling.

iii) The sampling ratio \((f_0)\), giving an optimal division of costs between modelling and assessment \((n_d = N^*f_0)\) that minimizes the variance \(V_d\). This is derived by Tenenbein (1970) as:

\[
f_0 = \min \left[ \frac{1-K}{KR}, 1 \right] \quad \text{Equation 4}
\]

If \(f_0\) 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 \(f_0 < 1\) then double sampling may generate more precise estimates for the same cost. A crucial metric is the proportional reduction in cost (or variance) achieved by double sampling \((\lambda)\):

\[
\lambda = 1 - \frac{(R + \frac{1}{f_0})(1-K-Kf_0)}{R} \quad \text{Equation 5}
\]

The threshold \(\lambda > 0\) (Fig. 1a) gives the region in which double sampling is a cost effective alternative to single sampling.

Estimating the coefficient of reliability \(K\)

We developed predictive models of extinction risk for four taxonomic groups: terrestrial mammals (hereafter, mammals), amphibians, reptiles and crayfish (Table 1). We defined data-sufficient species as threatened (Critically Endangered, Endangered or Vulnerable) or non-threatened (Near Threatened or Least Concern). For each group, we predicted the conservation status of data-
sufficient species with life-history, ecological, environmental and threat exposure information. We collected new datasets for reptiles and crayfish. For reptiles, we collected 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, we collected: maximum carapace length, habitat type, and number of IUCN-listed habitats (IUCN 2010) (Appendix S1). Using mean values from within species’ geographic ranges, we compiled species’ spatial data with ArcGIS 9.2 as follows:

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

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

We used an existing mammal dataset (Bland et al. 2014), and collated an amphibian dataset (Appendix S1) based on Bielby et al. (2008) and Cooper et al. (2008). Biological traits were phylogenetically imputed for 32-58% of mammal species (Appendix S1); all other data for all groups were 100% complete. Species data varied among groups, due to differences in variable measurement, variable availability, and variable relevance to risk prediction. Datasets remain comparable in the sense that they use the best macroecological data available to date to predict extinction risk in poorly known species.

Machine Learning (ML) tools are increasingly used in ecology for statistical pattern recognition (Cutler et al. 2007; Olden, Lawler & Poff 2008). For mammals and reptiles, we trained classification trees, boosted trees, random forest, k-nearest neighbours, support vector machines and neural
networks (Bland et al. 2014). For amphibians and crayfish, we only trained classification trees, random forests and boosted trees, as necessary data pre-processing for other ML tools increased model misclassifications (Appendix S1). For all groups, we trained decision stumps based on geographical range size alone (IUCN criterion B) to assess its predictive power. Range boundaries may be more uncertain for DD species than data-sufficient species. To assess the influence of uncertainty in range size on model predictions, we 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). We then recalibrated all models of extinction risk.

We partitioned data-sufficient species into a training set comprising 75% of species and a validation set comprising 25% of species. For each ML tool and dataset in turn, we optimized tuning parameters using ten-fold cross-validation on the training set. For each combination of tuning parameters, we measured area under the receiver operating characteristic curve (AUC) in the cross-validation test folds. We selected AUC as measure of model performance, as it 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 we calculated the reliability coefficient $K$ among all realisable thresholds and selected the threshold maximizing $K$.

**Estimating the cost ratio $R$**

For each taxon we 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. 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, we focused on IUCN criterion B (range size),

which can be predominantly investigated with presence/absence surveys. We estimated

survey costs through consultation with experts from the IUCN/SCC Specialist Groups and

a range of funding bodies for threatened species research (Appendix S1). We computed

three survey costs for mammals based on geographical range size (Appendix S1). We

computed one survey cost for amphibians, reptiles and crayfish as range size is less

variable among species (Appendix S1). We 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).

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

ML model calibration. We computed the project and staff costs of collecting species data

from database compilers for mammals (Jones et al. 2009; Bland et al. 2014), amphibians

(Bielby et al. 2008), reptiles (M. Böhm, pers. comm.), and crayfish (this study). We

computed three costs of mammal trait data as costs for the panTHERIA database were

uncertain (Jones et al. 2009). We computed the cost of data cleaning and ML model

calibration based on the recorded task time and staff costs of a postdoctoral researcher.

Details of costs for both risk assessments and predictive models are available in

Appendix S1.

Results

Estimating the coefficient of reliability $K$

Machine Learning tools achieved high classification performance in all groups as measured by AUC

(Table 2). Values of the coefficient of reliability $K$ ranged between 0 and 0.7 among models and taxa,
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 than models calibrated on raw range size (Table 2). Decision stumps based on geographical range size alone achieved lowest $K$ values in all taxa (Table 2).

**Estimating the cost ratio $R$**

We calculated cost ratios $R (c_1/c_2)$ of 233, 1,877 and 2,489 for mammals, contingent on the three cost estimates of trait data. We computed cost ratios of 836 for amphibians, 1,375 in reptiles, and 1,401 for crayfish. We present results for a medium cost ratio of $R=1,500$ among all groups, as the choice of cost ratio did not qualitatively affect results (see Appendix S1 for alternative cost ratios).

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: $R=272,131$). We present results for a medium cost ratio of $R=250,000$ among all groups, as the choice of cost ratio did not qualitatively affect results (Appendix S1).

**Double sampling**

It was always more cost-effective to determine the status of all DD 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) (Fig. 1). If all DD species were modelled with the best models calibrated on raw range size, assessments by the IUCN were required for a random selection of 11 mammals, 25 amphibians, 8 reptiles, and 3 crayfish. The number of risk assessments increased when models were calibrated on a coarse measure of range size, requiring the random selection of 43 mammals, 68 amphibians, 25 reptiles, and 9 crayfish. For the best model calibrated on precise data, reduction in cost achieved by double sampling was 68% in mammals, 60%
in amphibians, 46% in reptiles, and 53% in crayfish (Fig. 1). Reduction in cost decreased when coarsening range size: the best models achieved reductions in cost of 36–59% among groups. Reductions in cost achieved by models calibrated on range size alone were low: 32% in mammals, 47% in amphibians, 25% in reptiles, and 16% in crayfish (Fig. 1).

Discussion

We find that it is always more cost-effective to model the risk status of all DD species in a group and update IUCN Red List assessments for a small number of DD species (double sampling), compared to allocating all financial resources to updating IUCN Red List assessments (single sampling). Double sampling reduces the cost of determining the proportion of DD species at risk of extinction by up to 68%, as pre-existing biological data are used to minimize the number of field surveys to perform.

Assuming DD species not included in this study can be surveyed and assessed for similar costs as amphibians, reptiles and crayfish (US $25,400 per species), we estimate the total cost of surveying and risk assessments for all 12,206 DD species on the Red List (IUCN 2013) to US $323 million. Our figure does not reflect efficiencies in surveying multiple species simultaneously (Gardner et al. 2008), or the costs of assessing species under criteria other than B, which may provide more complete information on risk status. The cost of increasing the number of species on the Red List to 160,000 has been estimated at US $60 million (Barometer of Life: Stuart et al. 2010 also see Collen & Bailie 2010). Many invertebrate, plant and fungi species to be included in the Barometer of Life are not well studied and may be assessed as DD, so the initiative is likely to require considerable additional investment in field surveys. 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. 2014). Limited resources for tracking the status of biodiversity create a trade-off between expanding the taxonomic coverage of biodiversity assessments (Collen et al. 2009), keeping assessments up-to-date (Rondinini et al. 2014), and ensuring reliability of risk assessments (Bland et al. 2012). Improving our understanding of the costs and trade-offs involved in creating biodiversity indicators is therefore a
key topic for further research. 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 contribute to considerable uncertainty in conservation prioritization (Bland et al. 2012; Trindade-Filho et al. 2012) and may jeopardize the measurement of progress towards Aichi targets (Tittensor et al. 2014), particularly towards improving the conservation status of threatened species (Target 12: Convention on Biological Diversity 2010). Yet, they receive very little conservation investment: for example 3% of the awards from the Mohamed bin Zayed Species Conservation Fund are directed toward DD species (MBZSCF 2013). We show that using existing biological data and conducting risk assessments for poorly known species could enable the cost-effective monitoring of progress towards international biodiversity targets. Extinction risk models are not only a cheaper option than risk assessments for monitoring broad-scale changes in risk; they are also more likely to be developed within time scales relevant to biodiversity targets. Group assessments require extensive workshops, administration and training and typically take several years to complete (Rondinini et al. 2014). On the other hand, models require collection of data from species descriptions and other natural history resources, which can be carried out rapidly by non-experts. Whilst calibration of ML tools requires statistical expertise, accessibility could be improved by developing user-friendly platforms.

Comparison of models and taxonomic groups

The utility of risk models for conservation depends on their reliability and cost relative to risk assessments. In the focal groups considered in this study, models achieved very high AUC in validation sets, indicating excellent discrimination between threatened and non-threatened species (Table 2). Predictive performance varied among groups (Table 2) likely due to complex interactions among taxon size, risk prevalence and species data availability. Most models performed better on mammals and amphibians than crayfish or reptiles, likely due to the high number of mammals
modelled and the high prevalence of risk in amphibians. Random Forests, Boosted Trees, Neural Networks and Support Vector Machine achieved high coefficients of reliability. We recommend testing multiple ML tools for predicting risk in new species groups.

The biggest savings were achieved by improving model performance, whilst savings were less sensitive to the estimate of risk assessment and model costs. Approximate cost ratios may therefore be sufficiently informative when designing double sampling schemes (Appendix S1). Double sampling remains cost-effective under poor data quality: models calibrated on a coarse measure of range size still achieved 34–56% reduction in cost among groups. Reductions in cost achieved by models calibrated on range size alone were smaller (15–47%), indicating that collecting biological data is necessary to achieve the highest cost savings.

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), which is unlikely to occur. With poor models (K<0.1), reductions in cost are small (<10%) so managers may decide the overhead costs of calibrating models are not worthwhile. 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. To facilitate exploration of possible savings, we provide an Excel planning model (Appendix S2).

**Limitations and prospects**

We 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, Tibshirani & Friedman 2009). A multinomial double sampling scheme (Tenenbein 1972) could investigate the cost-effectiveness of estimating the prevalence of individual Red List
categories. We assume that the relationship between predictor variables and extinction risk is similar in data-sufficient and DD species. Accurate predictions require the range of predictor values exhibited by DD species to be represented by modelled data-sufficient species (Appendix S1; DD data not available for amphibians). Modelled species are also assumed to be representative of the wider data-sufficient species pool (Table 2). It should be noted that the utility of models is contingent on the quality of IUCN Red List assessments. The Red List status of some species may change due to genuine improvements and deteriorations in conservation status, as well as previous misclassifications (non-genuine reasons; Butchart et al. 2004).

Estimation of predictor variables may be less accurate in DD species, which could affect model performance. We 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. We also assume that geographical range maps are available for all species to assess in a sample, which may not be the case for all DD species (although only 3 DD crayfish species could not be mapped; B. Collen pers. comm.) or for species not assessed by the IUCN (e.g. species not selected in the Sampled Red List assessment of their taxonomic group; Bailie et al. 2008). 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.

The estimation of $K$ may depend on the number of species used to calibrate models of extinction risk. To deal with this problem, a three-stage sampling scheme (Tenenbein 1971) can determine the size of a pilot study to estimate $K$. Rarefaction analyses indicate that fewer than 1,000 mammals could have been assessed to obtain $K>0.5$ (unpublished data). Understanding the relationship between $K$ and sample size is therefore a useful avenue for further research, especially for species assessed with the Sampled Red List approach. Similarly, future studies could incorporate dynamic updating of $K$ as species of unknown conservation status are assessed by the IUCN.
The double sampling scheme relies on binomial sampling from an infinite population (Tenenbein 1970). In reality, DD species represent populations of finite size, which are more adequately modelled by a hypergeometric distribution. As the single sample size approaches the total population size, the variance in the estimated proportion of DD species at risk decreases faster for a hypergeometric distribution than for a binomial distribution, eventually reaching zero when all species have been assessed. Double sampling theory has not been extended to the hypergeometric distribution, but we have included single sampling under a hypergeometric model in our planning model (Appendix S2). 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 when funds for red listing are small (e.g. fewer than 188 out of 500 DD species can be assessed). 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).

Finally, our study addresses only one objective of the IUCN Red List, the quantification of global patterns and trends in extinction risk globally (IUCN 2014). The IUCN Red List also aims to pinpoint individual species at high risk of extinction (IUCN 2014). Models of extinction risk could address this objective by identifying high-risk DD 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), hence survey costs and likelihood of survey success should be taken into account during prioritization (Joseph et al. 2009).

Conclusions

To measure progress towards international targets and halt the current loss of biodiversity, reliable indicators of biodiversity status are needed. We show that double sampling with predictive models cost-effectively estimates the proportion of IUCN DD species at risk of extinction, and reduces assessments costs by up to 68%. Double sampling remains cost-effective under poor data quality and availability, demonstrating the method’s capacity to cheaply determine extinction risk levels in...
poorly known groups of plants and invertebrates. We conclude that double sampling could reduce
the impact of uncertainty in the Red List and Red List Index, and cost-effectively monitor progress
towards Target 12 of the Aichi Biodiversity Targets. The technique could also be applied to local and
national risk assessment programmes, and to species excluded from Sampled Red List assessments.

Double sampling schemes are available for multinomial data (Tenenbein 1972), continuous data
(Gilbert 1987), and for designing pilot studies in multiple stages (Tenenbein 1971). 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. Given the urgency of the
biodiversity crisis and the limited availability of conservation funds and biological data, designing
efficient monitoring schemes is imperative.

Acknowledgements

This research was supported by a Travel Fellowship to L.B. from the Australian Research Council
(ARC) Centre of Excellence for Environmental Decisions, an ARC Future Fellowship to M.A.M. and a
Centenary Research Fellowship to E.N. We thank the collaborative effort among the National
Autonomous University of Mexico, Stony Brook University, Nature Serve and the Institute of Zoology
for the collection of reptile data, particularly Andres Garcia, Monika Böhm and Ana Davidson. We
thank the curation staff of the Natural History Museum London, Musee d’Histoire Naturelle Paris,
Museum Victoria and the Australian Museum for assistance in collecting crayfish data.
References


IUCN. (2014) Red List Overview, http://www.iucnredlist.org/about/red-list-overview, Accessed on the 10th of...
October 2014.


Supporting Information

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

Appendix S1. Supplementary methods and results.

Appendix S2. Sampling planner for double sampling with a binomial distribution, single sampling with a binomial distribution and single sampling with a hypergeometric distribution.
Tables

Table 1 Description of IUCN Red List assessments and predictive models of extinction risk for terrestrial mammals, amphibians, reptiles and crayfish

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

* 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.

Sampled Red List of 1,500 randomly selected reptiles.
Table 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.
### a) Fine geographical range size

<table>
<thead>
<tr>
<th>Mammals</th>
<th>AUC*</th>
<th>Cutoff†</th>
<th>θ‡</th>
<th>φ¶</th>
<th>π§</th>
<th>K**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Stump</td>
<td>0.75</td>
<td>0.731</td>
<td>0.05</td>
<td>0.447</td>
<td>0.161</td>
<td>0.32</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>0.895</td>
<td>0.3</td>
<td>0.102</td>
<td>0.233</td>
<td>0.249</td>
<td>0.406</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.935</td>
<td>0.317</td>
<td>0.069</td>
<td>0.201</td>
<td>0.231</td>
<td>0.515</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.971</td>
<td>0.604</td>
<td>0.014</td>
<td>0.196</td>
<td>0.189</td>
<td>0.7</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.932</td>
<td>0.385</td>
<td>0.059</td>
<td>0.21</td>
<td>0.221</td>
<td>0.533</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>0.922</td>
<td>0.448</td>
<td>0.082</td>
<td>0.242</td>
<td>0.231</td>
<td>0.443</td>
</tr>
<tr>
<td>K-Nearest Neighbours</td>
<td>0.906</td>
<td>0.345</td>
<td>0.069</td>
<td>0.333</td>
<td>0.201</td>
<td>0.383</td>
</tr>
</tbody>
</table>

### Reptiles

| Decision Stump        | 0.726| 0.731   | 0.136| 0.18 | 0.569| 0.467|
| Classification Tree   | 0.898| 0.846   | 0.1  | 0.196| 0.5  | 0.485|
| Boosted Trees         | 0.949| 0.269   | 0.03 | 0.2  | 0.638| 0.629|
| Random Forests        | 0.953| 0.428   | 0.045| 0.18 | 0.621| 0.625|

### Crayfish

| Decision Stump        | 0.698| 0.731   | 0.026| 0.706| 0.11 | 0.157|
| Classification Tree   | 0.874| 0.828   | 0.053| 0.382| 0.229| 0.388|
| Boosted Trees         | 0.927| 0.38    | 0.093| 0.176| 0.321| 0.527|
| Random Forests        | 0.919| 0.456   | 0.08 | 0.176| 0.312| 0.555|

### b) Coarse geographical range size

<table>
<thead>
<tr>
<th>Mammals</th>
<th>AUC*</th>
<th>Cutoff†</th>
<th>θ‡</th>
<th>φ¶</th>
<th>π§</th>
<th>K**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Stump</td>
<td>0.718</td>
<td>0.731</td>
<td>0.038</td>
<td>0.525</td>
<td>0.135</td>
<td>0.28</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>0.875</td>
<td>0.75</td>
<td>0.045</td>
<td>0.411</td>
<td>0.165</td>
<td>0.368</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.912</td>
<td>0.456</td>
<td>0.062</td>
<td>0.297</td>
<td>0.204</td>
<td>0.436</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.927</td>
<td>0.408</td>
<td>0.046</td>
<td>0.279</td>
<td>0.196</td>
<td>0.497</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.915</td>
<td>0.394</td>
<td>0.058</td>
<td>0.301</td>
<td>0.199</td>
<td>0.441</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>0.892</td>
<td>0.36</td>
<td>0.096</td>
<td>0.292</td>
<td>0.231</td>
<td>0.363</td>
</tr>
<tr>
<td>K-Nearest Neighbours</td>
<td>0.897</td>
<td>0.276</td>
<td>0.124</td>
<td>0.228</td>
<td>0.267</td>
<td>0.368</td>
</tr>
</tbody>
</table>

### Amphibians

| Decision Stump        | 0.769| 0.731   | 0.4  | 0.06 | 0.706| 0.344|
| Classification Tree   | 0.9   | 0.286   | 0.12 | 0.12 | 0.551| 0.571|
| Boosted Trees         | 0.94  | 0.69    | 0.08 | 0.151| 0.517| 0.58 |
| Random Forests        | 0.946| 0.666   | 0.06 | 0.167| 0.5  | 0.587|

### Reptiles

| Decision Stump        | 0.5  | 0       | 0   | 0   | 0   | 0   |
| Classification Tree   | 0.854| 0.09    | 0.147| 0.219| 0.253| 0.298|
| Boosted Trees         | 0.901| 0.162   | 0.152| 0.171| 0.265| 0.331|
| Random Forests        | 0.89  | 0.242   | 0.157| 0.146| 0.273| 0.343|
| 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|

### Crayfish

| Decision Stump        | 0.633| 0.731   | 0.133| 0.471| 0.256| 0.141|
| Classification Tree   | 0.823| 0.727   | 0.12 | 0.323| 0.294| 0.322|
| Boosted Trees         | 0.868| 0.432   | 0.107| 0.206| 0.321| 0.467|
| Random Forests        | 0.883| 0.38    | 0.2  | 0.088| 0.422| 0.447|
* 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. The true proportion of threatened
species in the sample (p) for each group is: mammals = 0.221, amphibians = 0.568, reptiles = 0.169,
crayfish = 0.312.
** K: coefficient of reliability of the model.
Figures

Figure 1 Proportional reduction 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). (a) and (b): horizontal dotted line indicates the estimated cost ratio ($R = 1,500$). (c) Proportional reduction in cost and (d) optimal sampling proportion among predictive models and taxonomic groups. (c) and (d); circles: models calibrated on fine geographical range size. Crosses: models calibrated on coarsened geographical range size. Triangles: models calibrated on range size alone (left: fine geographical range size; right: coarse geographical range size). CT: classification tree. RF: random forests. BT: boosted trees. SV: support vector machines. NN: neural networks. KN: k-nearest neighbours. DS: decision stumps.