Hermoso, V., Kennard, M.J. & Linke, S. (2014). Evaluating the costs and benefits of systematic data acquisition for conservation assessments. *Ecography* DOI: 10.1111/ecog.00792

Published version on journal website available at:
http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1600-0587

Evaluating the costs and benefits of systematic data acquisition for conservation assessments.

Virgilio Hermoso, Mark J. Kennard and Simon Linke

Australian Rivers Institute and Tropical Rivers and Coastal Knowledge, National Environmental Research Program Northern Australia Hub, Griffith University, Nathan, Queensland, 4111, Australia

Short title: Data acquisition for conservation

Corresponding author: Virgilio Hermoso

email: virgilio.hermoso@gmail.com

Tlf: (61) 07 3735 5192

Fax: (61) 07 3735 7615

Abstract

Effective decision-making in conservation is constrained by the quality of the data available. Uncertainties associated with poor quality or sparse data can lead to the misuse of limited resources and the failure of conservation practice. Data acquisition, which can help improve decision-making, is constrained by limited budgets and time. This is especially concerning for rare species, the most in need of conservation, but the most difficult to accurately represent in conservation plans. Here we test the suitability of three different sampling design strategies (two systematic vs. random) specially focused on relatively rare species to improve the quality of information available for conservation planning. We modelled the spatial distribution of freshwater fish species in a data rich area in northern Australia using a large dataset (representing the best attainable data or true distribution) and simulate increasing subsets of data acquired through the three alternative sampling designs. We then evaluated omission and commission errors in conservation planning outcomes, efficiency and return on investment of data acquisition for conservation planning outcomes obtained from the different data availability x sampling design strategies. Even though we were able to find new species more effectively through systematic sampling designs, this did not directly translate into reduced errors in conservation planning outcomes for rare species. This led us to miss our goal of enhancing cost-effectiveness of conservation planning. Our results suggest that collecting more biodiversity data, irrespective of the sampling design used, does not necessarily reduce data uncertainty issues and could lead to the misuse of the limited resources available and ultimately the failure of conservation practice.

Keywords: Omission error, commission error, efficiency, conservation planning, Marxan, systematic sampling.

Introduction

The success of truly achieving conservation objectives is constrained by the accuracy of the data used in conservation assessments portraying the patterns in species distribution on the ground (called quality of data hereafter). Poor quality or sparse data on the presence and absence of species can potentially lead to high uncertainty in our estimates of spatial distribution of biodiversity, which can in turn lead to poor decision-making, the misuse of the limited resources available and ultimately the failure of conservation practice (Possingham et al. 2007). Errors derived from poor quality data can decrease effectiveness (e.g., commission errors, arising when a species is erroneously thought to be present within a reserve) and efficiency (e.g., omission errors, arising when a species is erroneously thought to be absent forcing the selection of additional and unnecessary areas; Loiselle et al. 2003) of conservation planning outcomes. Although these errors could be minimised by investing in data acquisition (Balmford and Gaston 1999, Hirzel and Guisan 2002), our capacity to improve decision-making by adding new and better quality data is limited by the cost and time required to collect it (Halpern et al. 2006, Grantham et al. 2008).

Rare species pose a crucial problem in conservation planning as there is usually a high uncertainty in their spatial distribution unless survey data used to estimate their distribution is intensive and extensive. The effectiveness of conservation plans (measured as the proportion of species that are adequately protected) can be enhanced by including more species in the planning process (Gaston and Rodrigues 2003) as they can be more accurately represented in the plan. However, data addition does not always result in a reduction in uncertainties. Additional sampling records could help improve data for already known species and add some new species not discovered so far, although submitted to high uncertainty due to data scarcity. This might lead to a decline in efficiency at least when following random data acquisition approaches (Gaston and Rodrigues 2003, Hermoso et al. 2013a). Systematic sampling designs, where sampling sites are strategically located to cover more efficiently the full range of environmental conditions that drive biodiversity patterns, have proved to be a good option to sample rare species and accurately represent their spatial distribution (Hirzel and Guisan 2002, Vaughan and Ormerod 2003). This could help reduce the errors associated with rare species and

enhance the cost-effectiveness of conservation planning. However, the trade-offs between investment in data acquisition through these systematic sampling designs and the effectiveness and efficiency of conservation planning outcomes have never been explicitly quantified.

Here we test the suitability of three different sampling design strategies (two systematic strategies vs. random sampling) for reducing commission and omission errors in conservation planning outcomes, as well as their effect on efficiency of priority area selection. The systematic sampling designs include an environmentally informed strategy, where new sites are added to maximise the environmental diversity covered in the dataset, and geographical-distance strategy, where new sites are added to maximise the geographical area covered in the data set. We would expect to find records for rare species more effectively when using a systematic sampling design approach than when using a random allocation of sampling sites. This should lead to better informed selection of priority areas and reduced errors (higher effectiveness and efficiency).

We use a data-rich area in northern Australia as case study to demonstrate the trade-offs associated with data acquisition through different sampling design strategies. We model the spatial distribution of freshwater fish species using the complete dataset and use it as the true distribution of each species representing the best available information (Grantham et al. 2009, Langford et al. 2011). We then use these predictions to assess the performance of alternative distribution maps obtained from models built on different subsets of the database (sub-sampled from the complete data set using different sampling design strategies and amounts of data). These maps are assumed to represent the information that a stakeholder would have available if the area had not been extensively surveyed. We evaluate performance through measures of commission and omission errors (*sensu* Rondinini et al. 2006), efficiency (ratio species representation/ cost in terms of number of planning units required) and return on investment (ratio species representation/ sampling cost).

Methods

Spatial framework, biodiversity and environmental data

The study area spans across all of northern Australia's coastal river basins from the Fitzroy River in the Kimberley region eastwards to the Jardine River in Cape York Peninsula. It encompasses a total area of about 1.19 million km^2; about 15% of the Australian continental area (Supplementary material Appendix 1. The region supports approximately 60% of Australia's freshwater fish diversity (Pusey et al. 2011).

We compiled all available sampling records for 104 freshwater fish species across the study area (see Hermoso et al. 2012; 2013a). This dataset contains records for more than 2300 sampling sites, although we retained for further analysis only sites with true presence-absence data (n=714 sites). For subsequent modelling purposes we translated these presence-absence records into a network of predictive units. We delineated 11508 subcatchments (102.7 km^2 on average) using ArcHydro (Maidment 2002) for ArcGIS 9.3 (ESRI 2002) from a 9 second digital elevation model (Geoscience Australia 2011). This will also be our planning units in the conservation planning analyses below. Although equal-sized grid cells are often used as planning units in terrestrial systematic conservation planning, subcatchments are more appropriate for fresh water as they better convey natural boundaries and ecological processes (such as connectivity; Hermoso et al. 2011). There were a total of 498 subcatchments containing at least one sampling site. For those subcatchments with more than one record (n=216) we combined the list of all species reported to produce a single record. We discarded from the dataset all the species with less than five occurrences, due to difficulties in modelling the distribution of these extremely rare species and the potential bias they would introduce to the analyses. Our final dataset comprised 70 fish species with an average frequency of occurrence of 95 subcatchments (range 5-433).

Predictive modelling of species distributions

Nine ecologically-relevant landscape-scale environmental variables were selected from a larger number of candidate variables for use in the predictive models and were derived from the National Environmental Stream Attributes database for rivers (see Geoscience Australia 2011 for details). We used Principal Component Analysis (PCA) to select a set of nine non-redundant environmental

attributes that explain a high proportion of the environmental variability in the study (Supplementary material Appendix 2).

We used Multivariate Adaptive Regression Splines (MARS, Leathwick et al. 2005) to model the spatial occurrence of the 70 fish species, because of its ability to more effectively model rare species than single-species models (see Elith et al. 2006, Elith and Leathwick 2007 for more details on MARS specifications). The model was built on the whole dataset (n=498 subcatchments, hereafter termed 'true model'). Model accuracy was evaluated using the deviance explained and the area under the receiver operating characteristic curve (ROC, Fielding and Bell 1997). The area under the ROC curve (AUC) was assessed through a k-fold cross validation procedure. Following this approach models were validated by dividing the dataset randomly into 10 exclusive subsets. Each subset is successively removed and the model refitted with the remaining data. This model is then used to predict species' probability of occurrence in the removed data. The AUC is calculated for each of this step as normally done by comparing predictions and observed presence-absence data (Fielding and Bell 1997). Model performance was calculated as the average AUC across all the subsets (Leathwick et al. 2005). We retained these measures as an estimate of the uncertainty around the predictions for each species. MARS models were fitted using the code provided by Elith and Leathwick (2007) for the mixture and flexible discriminant analysis (MDA) library in R (R Development Core Team, 2013).

The model was then used to predict the probability of occurrence of each species in all the unsurveyed subcatchments. Probabilities of occurrence were transformed into presence-absence data for posterior analyses using the optimal threshold obtained from the cost method in the presence-absence package in R (Freeman 2007). Given that these predictions were made with the best and most accurate dataset available, we treated them as our true species distribution for subsequent analyses (see Grantham et al. 2009, Langford et al. 2009 for similar approach). Note that this represents the best attainable estimate of species distributions for our study area rather than their true distributions.

Data availability and sampling design strategies

In a conservation planning context where available data are not sufficient and/or funding and capacity for data collection become available, different sampling strategies to collect additional data can be used. Here we tested three alternative sampling data designs, including two systematic approaches (environmentally and geographic distance informed) and random sampling:

1. Under the *environmentally informed design* we selected new sites from the available pool of sites not used yet to maximise the range of environmental conditions encompassed by the models. We used an environmental dissimilarity matrix containing each pair of sites' dissimilarity (based on Euclidean distance) to select those sites with the highest environmental distance to the ones already included in the previous model. Environmental dissimilarity was calculated using the same set of environmental variables used as predictive variables for the models after being standardised (listed in Supplementary material Appendix 2).
2. A similar approach was followed for the *geographic distance-based design*, but using the Euclidean distance between each pair of sites. New sites were added to cover the study area in the most effective way by maximising the distance to those sites already "surveyed" or used in previous models.
3. The third sampling design added new sites randomly selected from the available pool.

We simulated additional sampling of subcatchments using these three strategies, starting from a subset of subcatchments (15%, N=75) of the complete dataset (100%, N=498). All three strategies consisted in selecting and adding more subcatchments to an initial pool of 75 sampled subcatchments, randomly selected. The reduced dataset was sequentially increased in 10% steps up to 85% of the whole dataset (n=498 subcatchments). Each dataset was used as if no additional data were available so we followed the same procedure as detailed above for evaluating model performance (no additional data from the remaining set were used). We used the same set of environmental predictors across all models (same set of variables used for constructing the true model above, Supplementary material Appendix 2). We applied the same minimum threshold of 5 occurrences for a species to be included in the predictive model developed for each dataset. For subsequent analyses and reporting of results we will refer to species as *common* if they could be modelled in the sparsest data set (15% data) and

rare if they are species that we incorporated gradually during further data addition. Similar approaches to defining rare species based on the frequency of occurrence have been used before (Pressey et al. 1999). Note that, even though we used an advanced modelling approach that allowed us to model the distribution of relatively rare species, we had to discard from the analyses some extremely rare species (< 5 records in the full dataset) that could not be modelled, similar to Hermoso et al. (2013a). It was not possible include these species in the models as their low prevalence in the dataset (sometimes just one record) made it impossible to distinguish their occurrence from random occurrence and then difficult to model their habitat preferences (difficult to link their true occurrence to environmental variables through robust models). In all cases we recorded the number of species and the number of occurrences within each subset of data for each survey strategy.

Identification of priority areas for conservation

We used the predictions from each model as surrogates of biodiversity patterns to identify priority areas for conservation. We used the software Marxan (Ball et al. 2009) to find an optimal set of planning units to represent each species' predicted distribution at three target levels of occurrence (10%, 25% and 50%, respectively) at the minimum cost. Marxan uses a heuristic optimisation algorithm to minimize an objective function that includes the cost of subcatchments (our units of planning) in the solution and other penalties for not achieving the conservation target for all the species. Our goal was not to design a reserve network suitable for on-the ground conservation, but to examine the effects of sampling designs for data acquisition on hypothetical reserve networks. Given our special interest in evaluating the effect of data acquisition strategies, we used a constant cost for each planning unit, so our objective translated into finding the minimum set of planning units to achieve the conservation targets (Hermoso and Kennard 2012). We ran a total of 63 different conservation prioritisation scenarios (3 data acquisition strategies x 7 predictive models x 3 target levels; Fig. 1). For each scenario we retained 100 solutions obtained after 1.5 M iterations each for further analyses.

Commission and omission errors in conservation planning outcomes

We measured *commission error rates* (when a species is erroneously thought to be present within a conservation priority area) and *omission error rates* (when a species is erroneously thought to be absent forcing the selection of additional and unnecessary areas) in conservation planning outcomes by comparing the expected and observed representation achieved for each species. We did this for each of the 63 scenarios x 100 solutions obtained from Marxan (total of 6300 solutions). The expected representation of each species was measured as the number of occurrences within solutions according to the surrogate data used in the conservation prioritisation process (each of the 21 different predictive models). This was treated as the expected representation since it resembles the potential representation that would be achieved if the predictions used had no associated errors. The observed representation of each species was measured as the number of occurrences within solutions according to the true spatial distribution (i.e. based on the true model using the entire dataset). We then measured the rate of commission and omission errors independently for each species as the proportion of expected representation that was not truly achieved [$Error_i$= (Expected rep_i-Observed rep_i)/Expected rep_i, being i each of our modelled species]. Whenever the expected and observed representations are similar, the error obtained is close to 0, indicating low commission or omission error. However, when the expected representation is significantly higher or lower than the observed representation, the error value will depart from 0 and be negative or positive, indicating commission and omission errors, respectively (Hermoso et al. 2013a).

In order to explore the effect of potential determinants of the observed errors we used General Linear Models (GLM). We combined the results from the different models for each sampling design and ran separate GLM models using omission and commission errors as dependent variables and species' prevalence in the dataset (number of records for each species), an estimate of the spatial aggregation of each species' occurrences, the proportion of presences and absences correctly predicted, and the proportion of the total dataset used in the model as independent variables. The spatial aggregation of records was measured as the average number of records per unit area within the catchments where each species occurred. The proportion of presences and absences correctly predicted were estimated as the ratio of the true presences and absences predicted by each model. We retained the adjusted R^2 as an indicator of the model fit and the *Beta* coefficient and *P* value of each independent variable in

the model as an estimate of their relative importance at explaining the dependent variable. We would expect important determinants to be included in a model that explains a high proportion of variation in the dependent variable (high adjusted R^2), with a high *Beta* coefficient. All independent variables were minimally correlated and so were appropriate for use in the GLM analyses (Pearson's $r<0.40$ for all pairwise comparisons).

Efficiency and return on data acquisition investment

We measured the efficiency of each solution as the average ratio across all species between the true representation and the total number of planning units required. Efficient solutions achieve high representation with a reduced set of subcatchments. The investment necessary to survey each dataset used for the models was estimated as a combination of the number of sites and the geographic distance between them (assuming that the higher the number of sites and the further apart they are, the more expensive it would be to survey them). Based on our extensive prior sampling experience in Australian rivers, (e.g. Kennard et al. 2006), we estimated the cost to survey each site as the sum of a constant sampling cost per site ($500) plus an additional cost for the relative isolation of that site. We assumed that the farther apart the neighbouring sites were, the lower the chance a given site could be included in a four sampling sites/ day schedule (we assume that a team can survey four sites per day on average). Due to difficulties in estimating travelling distances between sites (most of sampling sites are located in remote areas far from mapped roads and only accessible by unmapped dirt tracks), the isolation cost was based on the euclidean distance between sites. This additional cost was estimated as a weighted sum of distances to the three closest neighbour sites included in each subset of data (Isolation cost= $2*Dist_1+5*Dist_2+10*Dist_3$, where $Dist_1$, $Dist_2$ and $Dist_3$ are the distances in km to the nearest, second nearest and third nearest sampling sites respectively. Note that the weights were set to reflect the relative increment in cost when needing to travel long distances between sampling sites –e.g., time spent -). We estimated the total cost associated with each subset of data as the sum of the individual sampling cost of the sites included. The return on a given investment for each subset of data was assessed as the average ratio across species between the observed representation and cost of survey. Finally, we explored the efficiency of this investment at reducing

omission and commission errors by correlating (Pearson correlation) the magnitude of the error and the cost associated with the survey of each dataset.

Results

Sampling designs and predictive model performance

A total of 46 species fulfilled the prevalence > 5 threshold in the initial randomly selected 15% of data (common species hereafter). Starting from that common point the environmentally informed design was the most effective sampling design at finding rare species (new species found in subsequent data additions with a prevalence > 5; Fig. 2a) and accumulating new records for them at low-data availability scenarios (Fig. 2b). However, the geographical distance and random strategies outperformed the environmentally informed strategy at higher data availability scenarios as they could find at least five records to represent a greater number of rare species.

Predictive models showed high performance with AUC > 0.75 and deviance explained > 0.3 in most cases, even for the low-data availability datasets (Fig. 3, Supplementary material Appendix 3). There was a general improvement in model performance when adding new data across all different survey strategies, but improvements were relatively minor after about 55% of data was used (Fig. 3). This happened even though new species with low prevalence were gradually incorporated to the models as they fulfilled the > 5 occurrences threshold (up to completing the 70 species modelled under the true model). The environmentally informed strategy showed the highest model performance overall, while the random strategy showed the highest increase in performance relative to the initial model with low data availability (Fig. 3, Supplementary material Appendix 3). This improvement in model performance was also followed by higher rates of correctly predicted presences and absences (Supplementary material Appendix 4).

Omission and commission errors in conservation planning outcomes

The improvement in modelling performance did not directly translate into a general reduction in errors in conservation planning outcomes. These results were independent of the target level used, so for the sake of simplicity only the results for the 25% target are presented here. Across all species no

consistent trends in omission and commission error rates were apparent with increasing data addition or data acquisition strategy (Fig. 4a). However, both omission and commission errors were always higher for rare than common species (Fig. 4b, 4c). There was also a major reduction of omission errors for common species with increasing data quantity across all the different sampling strategies (Fig. 4b). Omission errors for rare species followed contrasting patterns to data addition depending on the sampling design tested. There was an increase in omission errors at low-data availability scenarios for the environmentally informed design that decreased towards high-data availability (Fig. 4c). The geographical distance design showed opposite patterns, with low omission errors at low-data availability which increased towards high-data availability. There was no response of omission errors to the random selection design. Commission errors remained invariant to data addition independently of the sampling design used for both common and rare species except for the random sampling design and for common species (Fig. 4b). Across all three data acquisition strategies, the proportion of species' presences correctly predicted was the most important factor explaining omission errors, while commission errors were mostly explained by the proportion of absences correctly predicted and species' prevalence (Table 1).

Cost, efficiency and return on investment

There was a similarly major and rapid increase in cost for new data addition across all three data acquisition strategies (Fig. 5a). Expanding the dataset from the lowest-data to the highest-data availability scenario posed a two-fold increase in cost for the geographical distance and random sampling strategies and a three-fold for the environmentally informed design. In all cases the average cost per site decreased as the dataset was expanded given the reduction in travel expenses between sampling sites (Supplementary material Appendix 5). The investment in new data acquisition had only significant effects on reducing omission errors of common species across all different sampling strategies and commission errors of common species for the random sampling design (Table 2). Increasing data addition led to a general reduction in efficiency for all data acquisition strategies (Fig. 4b), but was more pronounced for the random addition strategy (20% decline in efficiency) than the environmentally informed (14% decline) and geographical distance (13% decline) strategies. The total

number of planning units needed to achieve the targets was similar across different data availability and sampling strategies (2815, 2830 and 2815 planning units for the environmentally informed, geographical distance and random addition respectively), so we rule out the potential effect of differences in number of planning units on these results. There was also a significant decline in return on investment for data acquisition across all sampling strategies (64%, 53% and 57% decline for the environmentally informed, geographical distance and random sampling strategies respectively; Fig. 5c).

Discussion

Investment in data acquisition is widely recognised as an adequate strategy to improve decision-making by reducing potential errors derived from uncertainty in the spatial distribution of biodiversity (Balmford and Gaston 1999). Our results demonstrate that this assumption cannot be generalised as data acquisition is not always the best option to address errors in conservation planning outcomes, especially when dealing with relatively rare species. Given the constraints to data acquisition imposed by limited budget and time, it is crucial to find cost-effective ways to enhance the quality of the data used (Nichols and Williams 2006). Systematic sampling designs have been suggested as an effective strategy to obtain data for rare species which are prone to high uncertainty in their spatial distribution (Wessels et al. 1998, Hirzel and Guisan 2002, Vaughan and Ormerod 2003). Here we show that in our case systematic sampling designs did not improve the ability to represent rare species in spatial prioritisation exercises. Systematic designs were more efficient at finding new species (more species found with less sampling effort) for low sampling effort than random addition of sites. However, this was not translated into a reduction in omission and commission errors, especially for these new species (relatively rare in our data set). These errors led to a decline in efficiency of priority areas and a reduction in the return on investment (comparatively lower true representation of biodiversity for each dollar invested in sampling). Our results suggest that collecting more biodiversity data, irrespective of the sampling design used, does not necessarily reduce data uncertainty issues and associated errors in conservation planning recommendations and could lead to the misuse of the limited resources available and ultimately the failure of conservation practice. This does not

disqualify the value of biodiversity surveys, given that it would be very difficult to detect some of the rarest species in the landscape (most in need of conservation) without intensive surveys. This may particularly apply in biogeographically complex or environmentally heterogeneous areas that may exhibit high species turnover and centres of endemism/rarity. Although we included a large proportion of freshwater fish species inhabiting the study area in our analyses, the number of species accounted for here is small compared to other freshwater ecosystems (e.g., large catchments in tropical areas such as the Amanzon River) or realms (e.g., tropical forests). Further research is required to adequately account for these rare species in conservation planning and confirm our results in more biogeographically complex areas.

Omission and commission errors

Both omission and commission errors were always higher for rare species than for common species irrespective of the sampling design used. These differences were especially marked for omission errors, indicating poor predictions of true occurrences for rare species. This result contrasted with the general improvement in predictive modelling performance experienced when adding data and might reflect over-fitting of models for rare species. Over-fitting produces predictive models that are very accurate at making predictions within a very narrow range of environmental conditions but unable to extrapolate beyond that (predictions are centred around the range of environmental characteristics of occurrences included in the model, which might not necessarily accurately reflect the breath of the species' ecological niche, as the rate of false negative predictions of our models reflect; see Supplementary material Appendix 3 and 4 for details on prediction errors). As a consequence, the spatial distribution of rare species was erroneously considered to be narrower than it actually was leading to inflated omission errors in conservation planning outcomes. The effect of systematic sampling designs was the opposite to expected, as there were either no significant differences between systematic and random addition of sites (e.g., omission errors for common species) or higher errors when using the systematic sampling approach (e.g., commission errors for common and rare species). This result aligns with Hirzel and Guisan (2002) who reported that the most critical parameter when trying to improve predictive modelling performance is sample size, independent of the sampling

design used. However, in our case, data acquisition was clearly beneficial only to address omission errors for common species. The more data added to the models, independent of the design used, the higher the predictive accuracy for common species, which resulted in reduced omission errors. This pattern was not clear for rare species, as sampling designs showed different responses to data acquisition. Allocating sites to maximise the spatial coverage of the area surveyed was the best strategy to minimise omission errors for rare species, especially when the capacity to acquire new data was limited.

Efficiency and return on investment

The investment in data acquisition had a doubly negative effect on conservation planning outcomes from an economic point of view as more expensive surveys led to less efficient priority areas. Our results demonstrate that the use of a large amount data is not always better than fewer data. This highlights the value of comparatively small datasets for conservation planning (Nichols and Williams 2006, Grantham et al. 2008, Williams et al. 2011). Data addition resulted in a decline in efficiency of priority areas, independent of the sampling design used (more areas were required to fulfil the conservation targets, which were adequately achieved in all Marxan solutions). Similar results have been reported in previous studies. For example, Gaston and Rodrigues (2003) showed a reduction in efficiency when deleting species records from a dataset used to identify priority areas for conservation in South Africa. Hermoso et al. (2013a) also showed that random data addition resulted in increased effectiveness as more species were adequately represented within priority areas, but at the expense of reducing efficiency. This was an expected result due to the increase in the number of rare species included in the prioritisation process (Pressey et al. 1999, Rodrigues and Gaston 2001). The decline in efficiency was mainly driven by omission errors related to the rarest species. The underestimation of the spatial distribution of rare species constrained the capacity of the optimisation algorithm to find areas that complement each other (Kirkpatrick 1983). This forced the selection of more areas than needed since the species were erroneously considered to be absent from areas where they actually occurred.

There was also a strong diminishing return on investment for data acquisition (up to two-fold between the largest and smallest datasets). The average species representation (effectiveness) for each dollar spent in data collection declined with increasing amounts of data collected. Grantham et al. (2008) also reported the decline in return on investment for data acquisition for proteas in South Africa. Here we demonstrate that this happened independent of the sampling design used for acquiring the data. In our particular case, the resources that could be spent in data acquisition might be better directed toward other actions, such as implementation of on the ground management. Our measure of return on investment was based on the improved capacity to avoid representation errors in conservation planning outputs for a given investment in data acquisition. However, return on investment conveys a broader concept and additional aspects related to conservation success (e.g., retention of species and habitats or social-economic aspects) should be considered in specific research in the future. Further research would also be required to confirm similar patterns in return on investment for data acquisition strategies in other realms (although see Grantham et al. 2008) to better inform investment of limited resources available for conservation. This would enhance cost-effectiveness of conservation practice, especially in areas where biodiversity is currently under the pressure of increasing threats (Grantham et al. 2009).

Rare species: mind them or neglect them?

Our results demonstrate that the inclusion of rare species in conservation planning must be cautiously considered. This poses a dilemma for conservation practitioners, as rare species are a substantial component of biodiversity, and their conservation is a major objective of many management plans. However, the errors associated with the weak knowledge on their spatial occurrence can severely flaw the success of conservation practice. Systematic sampling designs have been suggested as an efficient way of collecting data to improve the representation of the rarest species in datasets (Margules and Austin 1994, Hirzel and Guisan 2002). These designs bias sampling efforts towards rare environmental or geographical areas with the aim of increasing the chance of finding rare species. We have demonstrated that these strategies are more effective at finding rare species than random allocation of sampling sites. However, even when using these sampling designs, collecting enough

information to obtain a reasonable level of analytical power for conservation prioritisation would require large investment in data acquisition (Vaughan and Ormerod 2003). In our case, data acquisition through systematic sampling designs clearly failed at reducing errors in conservation planning outcomes for rare species due to the errors in their predicted spatial distribution. Even though we were able to find the spots where rare species "hide" more effectively, there was not a direct translation into reduced errors in conservation planning outcomes. This led us to miss our goal of enhancing cost-effectiveness of conservation planning. Moreover, we could not model the spatial distribution of some of the rarest species in the study area and so could not evaluate the effect of including them in conservation planning scenarios. Unfortunately, exclusion of rare (and difficult to model) species is a common practise in conservation planning (e.g., Linke et al. 2008). We acknowledge that these species might be the most in need of conservation actions and so need to be incorporated in the planning process. Alternative approaches to representing biodiversity patterns such as coarse filter surrogates (e.g., habitat types) have also been reported as poor surrogates for rare species (Januchowski-Hartley et al. 2011, Hermoso et al. 2013b). The effectiveness of these surrogate methods is limited as the representation of rare species within priority areas identified by using coarse filter surrogates is generally lower than random. Leaving rare species out of conservation plans could be an option to improve efficiency. However, given their reduced areas of occupancy, the incidental representation of rare species in conservation plans derived for common species would be reduced and, consequently is a risk prone approach. For this reason, new alternatives need to be explored to improve the efficiency of sampling designs (see Pacifici et al. 2012) and testing new approaches to adequately incorporate these species through novel surrogate methods (see Ferrier 2002, Arponen et al. 2008) are required.

Acknowledgements

We thank J. Stein for providing environmental data and B. Pusey and D. Burrows for assistance with compilation of fish data. We acknowledge funding support provided by the Australian Research Council (Discovery Grant DP120103353 to SL and MK; DECRA DE130100565 to SL), the Australian Government Department of Sustainability, Environment, Water, Population and

Communities, the Tropical Rivers and Coastal Knowledge (TRaCK) Research Hub, the National Environmental Research Program Northern Australia Hub, and the Australian Rivers Institute, Griffith University.

References

Arponen, A., Moilanen, A. and Ferrier, S. 2008. A successful community-level strategy for conservation prioritization. Journal of Applied Ecology 45: 1436–1445.

Ball, I. R., Possingham, H. P. and Watts M. 2009. Marxan and relatives: Software for spatial conservation prioritisation. In Moilanen, A., Wilson, K. A. and Possingham, H. P. (eds), Spatial conservation prioritisation: Quantitative methods and computational tools. Oxford University Press, Oxford, UK, pp. 185-195.

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

Elith, J. and Leathwick, J. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 265-275.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, R.J., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. and Zimmermann, N.E. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.

ESRI. 2002. ArcGIS. Environmental Systems Research Institute, Redlands, CA.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Systematic Biology 51: 331–363.

Freeman, E. 2007. Presence/Absence: An R Package for Presence-Absence Model Evaluation. USDA Forest Service, Rocky Mountain Research Station, Ogden, UT, USA.

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.

Geoscience Australia. 2011. Environmental Attributes Database. Available at http://www.ga.gov.au/ (last visited 29 November 2012).

Gaston, K. J. and Rodrigues, A. S. L. 2003. Reserve selection in regions with poor biological data. Conservation Biology 17: 188-195.

Grantham, H. S., Moilanen, A., Wilson, K. A., Pressey, R. L., Rebelo, T. G. and Possingham, H. P. 2008. Diminishing return on investment for biodiversity data in conservation planning. Conservation Letters 1: 190–198.

Grantham, H. S., Wilson, K. A., Moilanen, A., Rebelo, T. and H. P. Possingham. 2009. Delaying conservation actions for improved knowledge: how long should we wait? Ecology Letters 12, 293-301.

Halpern, B. S., Regan, H. M., Possingham, H. P. and McCarthy, M.A. 2006. Accounting for uncertainty in marine reserve design. Ecology Letters 9: 2–11.

Hermoso, V., Linke, S., Prenda, J. and Possingham, H.P. (2011) Addressing longitudinal connectivity in the systematic conservation planning of fresh waters. Freshwater Biology 56: 57–70.

Hermoso, V. and Kennard, M. J. 2012. Uncertainty in coarse conservation assessments hinders the efficient achievement of conservation goals. Biological Conservation 147: 52-59.

Hermoso, V., Kennard, M. J. and Linke, S. 2013a. Data acquisition for conservation assessments: is the effort worth it? PLoS ONE 8(3): e59662. doi:10.1371/journal.pone.0059662.

Hermoso, V., Januchowski-Hartley, S. R. and Pressey, R. L. 2013b. When the suit does not fit biodiversity: loose surrogates compromise the achievement of conservation goals. Biological Conservation 159: 197-205.

Hirzel, A. and Guisan, A. 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157: 331-341.

Januchowski-Hartley, S. R., Hermoso, V., Pressey, R. L., Linke, S., Kool, J., Pearson, R. G., Pusey, B. J. and VanDerWal, J. 2011. Coarse-filter surrogates do not represent freshwater fish diversity at a regional scale in Queensland, Australia. Biological Conservation. 144: 2499–2511.

Kennard, M. J., Pusey, B. J., Harch, B. H., Dore, E. and Arthington, A.H. 2006. Estimating local stream fish assemblage attributes: sampling effort and efficiency at two spatial scales. Marine and Freshwater Research. 57: 635–653.

Kirkpatrick, J. B. 1983. An iterative method for establishing priorities for the selection of nature reserves: an example from Tasmania. Biological Conservation 25: 127–134.

Langford, W. T., Gordon, A. and Bastin, L. 2009. When do conservation planning methods deliver? Quantifying the consequences of uncertainty. Ecological Informatics 4: 123–135.

Langford, W. T., Gordon, A., Bastin, L., Bekessy, S. A., White, M. D. and Newell, G. 2011. Raising the bar for systematic conservation planning. Trends in Ecology and Evolution 26: 634-640.

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J. and Hastie, T. 2005. Using multivariate adaptive regression splines to predict the distribution of New Zealand´s freshwater diadromous fish. Freshwater Biology 50: 2034-2052.

Linke S., Norris, R.H. and Pressey, R.L. 2008. Irreplaceability of river networks: Towards catchment-based conservation planning, Journal of Applied Ecology 45: 1630-1638.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G. and Williams, P. 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology 17: 1591–1600.

Maidment, D. R. 2002. Arc Hydro: GIS for Water Resources. ESRI Press, Redlands, CA.

Margules, C. R. and Austin, M. P. 1994. Biological models for monitoring species decline: the construction and use of data bases. Philosophical Transactions of the Royal Society of London Series B 344: 69–75.

Nichols J. D. and Williams, B. K. 2006. Monitoring for conservation. Trends in Ecology and Evolution 21: 668–673.

Pacifici, K., Dorazio, R. M. and Conroy, M. J. 2012. A two-phase sampling design for increasing detections of rare species in occupancy surveys. Methods in Ecology and Evolution 3: 721-730.

Possingham, H. P., Grantham, H. and Rondinini, C. 2007. How can you conserve species that haven't been found? Journal of Biogeography 34: 758–759.

Pressey, R. L., Possingham, H. P., Logan, V. S., Day, J. R. and Williams, P. H. 1999. Effects of data characteristics on the results of reserve selection algorithms. Journal of Biogeography 26: 179–191.

Pusey, B., Kennard, M.J., Burrows, D., Perna, C., Kyne, P., Cook, B. and Hughes, J. 2011. Freshwater fish. Pages 71–92, in B.J. Pusey, editor. Aquatic Biodiversity in Northern Australia: Patterns, Threats and Future. Charles Darwin University Press, Darwin.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna.

Rodrigues, A. S. L. and Gaston, K. J. 2001. How large do reserve networks need to be? Ecology Letters 4: 602–609.

Rondinini, C., Wilson, K., Boitani, L., Grantham, H. and Possingham, H.P. 2006. Tradeoffs of different types of species occurrence data for use on systematic conservation planning. Ecology Letters 9: 1136–1145.

Vaughan, P. and Ormerod, S. J. 2003. Modelling the distribution of organisms for conservation: optimising the collection of field data for model development. Conservation Biology 17: 1601–1611.

Wessels, K.J., Van Jaarsveld, A. S., Grimbeek, J. D. and Van der Linde, M. J. 1998. An evaluation of the gradsect biological survey method. Biodiversity and Conservation 7: 1093-1121.

Williams, B.K., Eaton, M. J. and Breininger, D. R. 2011. Adaptive resource management and the value of information. Ecological Modelling 222: 3429–3436.

Figure 1. Flow diagram of analyses. We created 21 different datasets by sequentially adding sites to a small random sample of sites (15% of the original dataset composed of 498 sites) simulating three different sampling design strategies (maximising geographical coverage, environmentally informed and random addition). We then used these datasets to build predictive models to estimate the probability of occurrence of fish species across Northern Australia. We also built an additional model where we used the entire dataset available (true model), which will be used as an indicator of the true distribution of species, or at least the best attainable estimates. The model outputs from the 7 subsets/ sampling design strategy were then used in Marxan to identify priority areas for conservation. Results from Marxan were compared against the true distribution to obtain estimates of three different performance measures: 1) commission and omission errors, and 2) efficiency.

Figure 2. Accumulation of rare species (a) and average number of records (b; Mean ± SE) to the datasets for different sampling designs strategies (environmental dissimilarity, geographical distance and random in black, grey and white bars respectively).

Figure 3. Predictive models performance (Average AUC ± SE) for increasing data available gained through different sampling designs (environmental dissimilarity, geographical distance and random in black, grey and white bars respectively).

Figure 4. Average (± SE) omission and commission errors across all species (a), common species (b) and rare species (c) for increasing data available gained through different sampling designs (environmental dissimilarity, geographical distance and random in black, grey and white bars respectively).

Figure 5. Cost (a), efficiency (b; Mean ± SE) and return on investment (c; Mean ± SE) for increasing data available gained through different sampling designs (environmental dissimilarity, geographical distance and random in black, grey and white bars respectively).

Table 1. GLM models to explore the importance of four different variables at explaining commission and omission errors for each data acquisition strategy (PCP= Proportion of presences correctly predicted, PCA= 1-Proportion of absences correctly predicted, Spatial dist= Measure of spatial aggregation of presences for each species –average of occurrences for catchments where the species was found-).

	Data acquisition strategy										
Error	**Environmental**				**Geographical**				**Random**			
Omission	**Variable**	**Beta**	**P**	**Adj R^2**	**Variable**	**Beta**	**P**	**Adj R^2**	**Variable**	**Beta**	**P**	**Adj R^2**
	PCP	-0.97	<0.001	0.90	*PCP*	-0.91	<0.001	0.79	*PCP*	-0.86	<0.001	0.68
	PCA	0.27	<0.001		*PCA*	0.20	<0.001		*PCA*	0.48	<0.001	
	Prevalence	0.04	0.15		*Prevalence*	-0.01	0.84		*Prevalence*	0.19	<0.001	
	Spatial dist	0.09	<0.001		*Spatial dist*	0.08	0.04		*Spatial dist*	0.03	0.32	
	Amount data	0.004	0.87		*Amount data*	0.04	0.33		*Amount data*	-0.16	<0.001	
Commission	**Variable**	**Beta**	**P**	**Adj R^2**	**Variable**	**Beta**	**P**	**Adj R^2**	**Variable**	**Beta**	**P**	**Adj R^2**
	PCP	0.34	<0.001	0.47	*PCP*	0.31	<0.001	0.46	*PCP*	0.32	<0.001	0.39
	PCA	-0.61	<0.001		*PCA*	-0.62	<0.001		*PCA*	-0.64	<0.001	
	Prevalence	-0.47	<0.001		*Prevalence*	-0.55	<0.001		*Prevalence*	-0.45	<0.001	

Spatial dist	-0.02	0.15	*Spatial dist*	-0.07	0.13	*Spatial dist*	-0.25	<0.001
Amount data	0.18	<0.001	*Amount data*	0.14	0.001	*Amount data*	0.05	0.45

Table 2. Relationships (Pearson's R) between the cost of data acquisition and commission and omission errors for each of three data acquisition strategies. Significance values (P) are also shown.

		Data acquisition strategy					
Species	Error	Environmental		Geographical		Random	
		R	P	R	P	R	P
Common	Omission	**-0.91**	**0.005**	**-0.87**	**0.01**	**-0.74**	**0.05**
	Commission	0.36	0.44	-0.10	0.82	**-0.91**	**0.003**
Rare	Omission	-0.40	0.37	0.60	0.15	-0.68	0.09
	Commission	0.30	0.20	-0.58	0.17	0.45	0.31

Figure 1

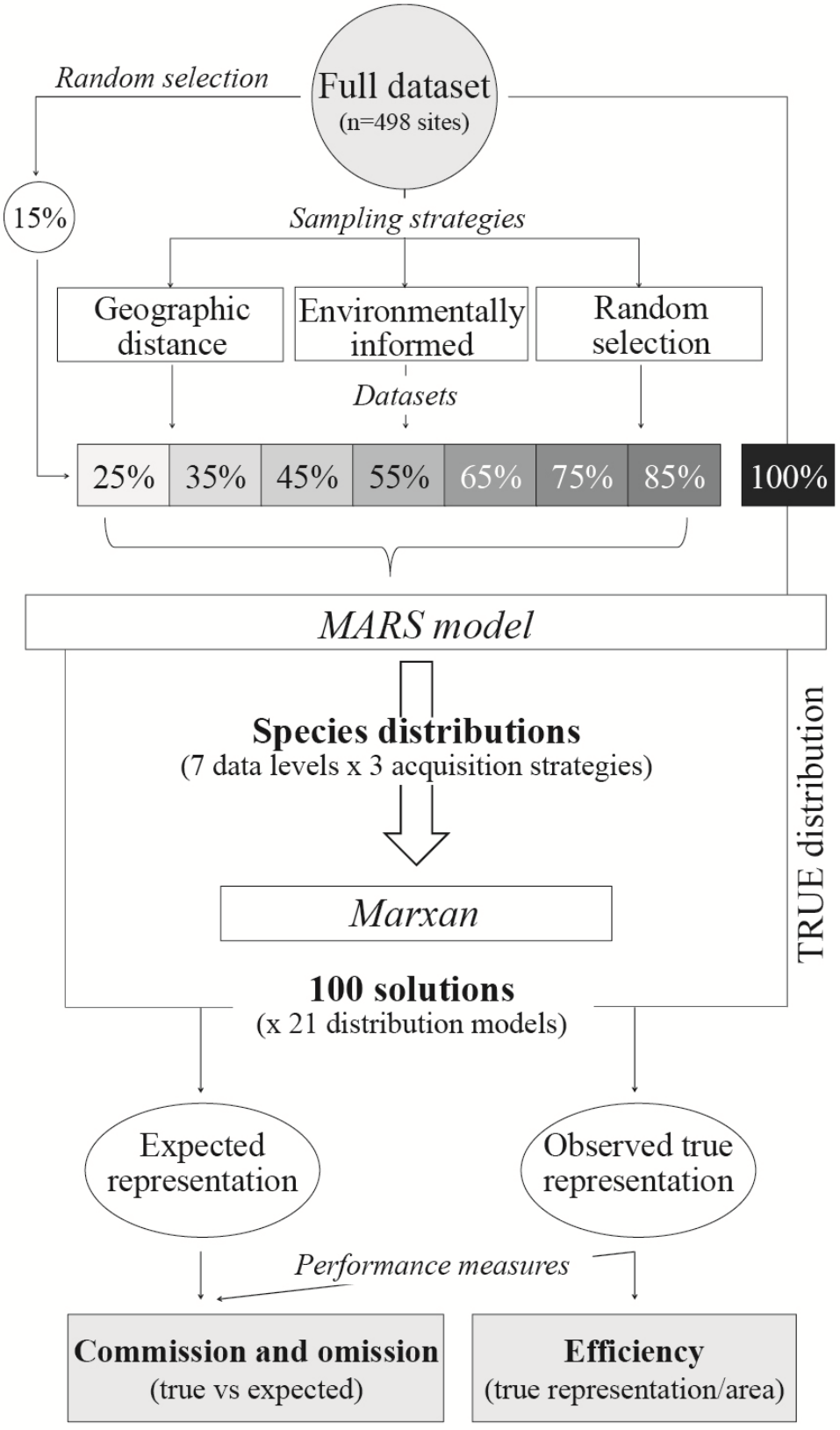

Random selection
Full dataset
(n=498 sites)
15%
Sampling strategies
Geographic distance
Environmentally informed
Random selection
Datasets
25%
35%
45%
55%
65%
75%
85%
100%
MARS model
Species distributions
(7 data levels x 3 acquisition strategies)
Marxan
100 solutions
(x 21 distribution models)
TRUE distribution
Expected representation
Observed true representation
Performance measures
Commission and omission
(true vs expected)
Efficiency
(true representation/area)

Figure 2

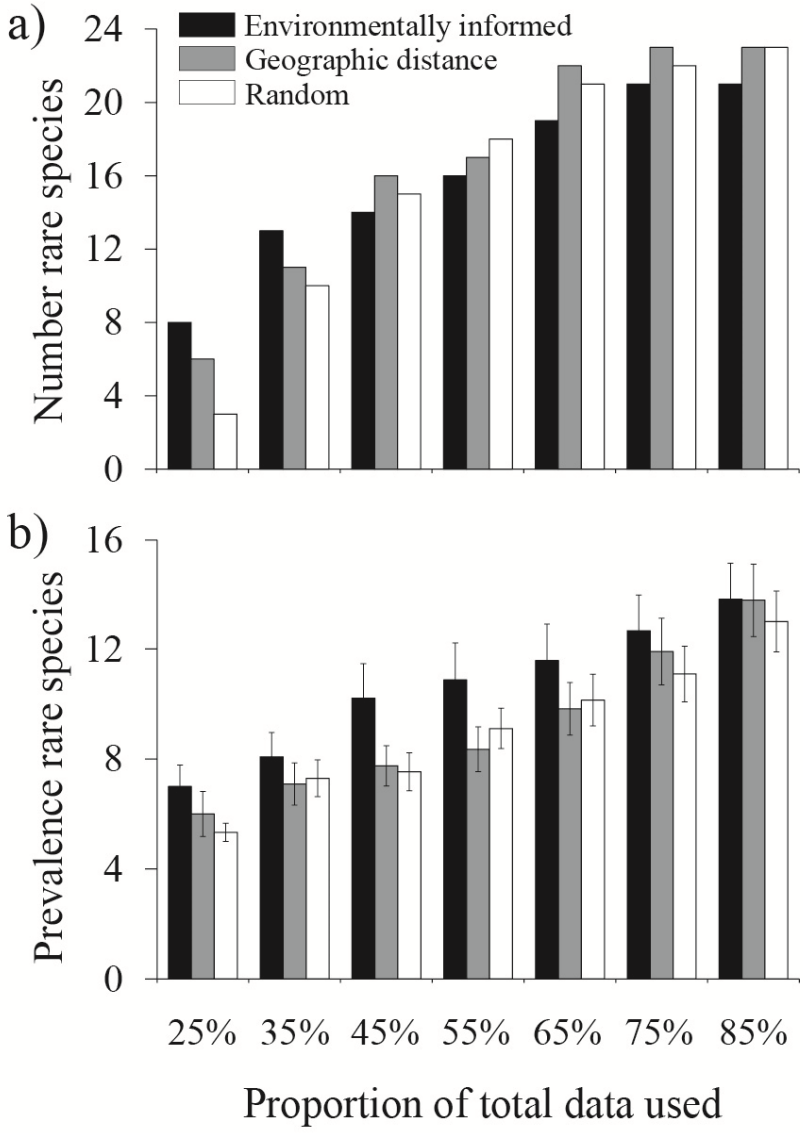

a)
Environmentally informed
Geographic distance
Random
Number rare species
0
4
8
12
16
20
24
b)
Prevalence rare species
0
4
8
12
16
25%
35%
45%
55%
65%
75%
85%
Proportion of total data used

Figure 3

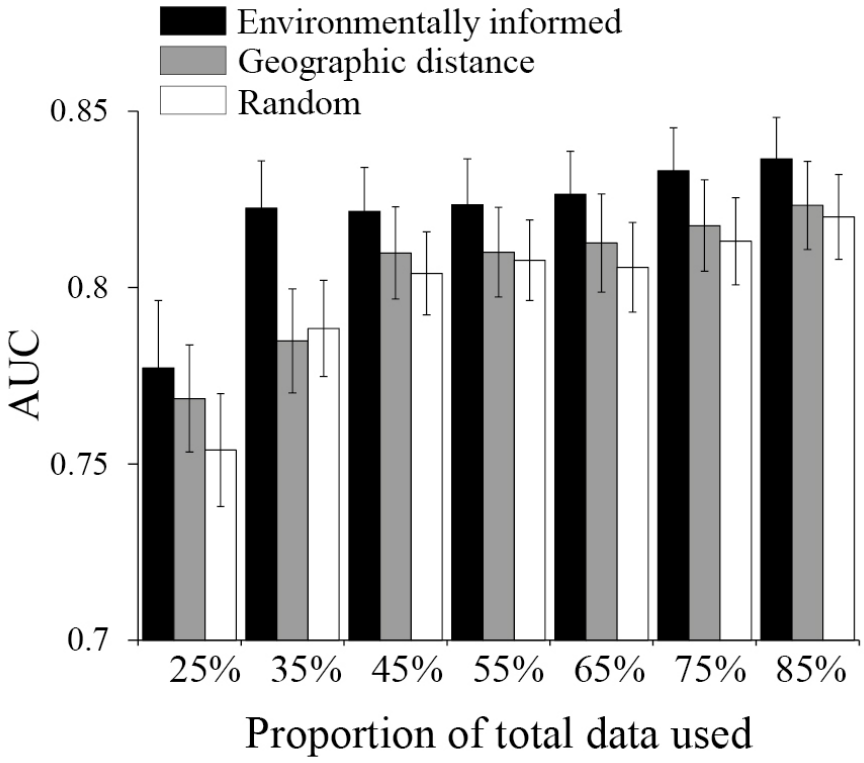

Environmentally informed
Geographic distance
Random
0.85
0.8
0.75
0.7
AUC
25%
35%
45%
55%
65%
75%
85%
Proportion of total data used

Figure 4

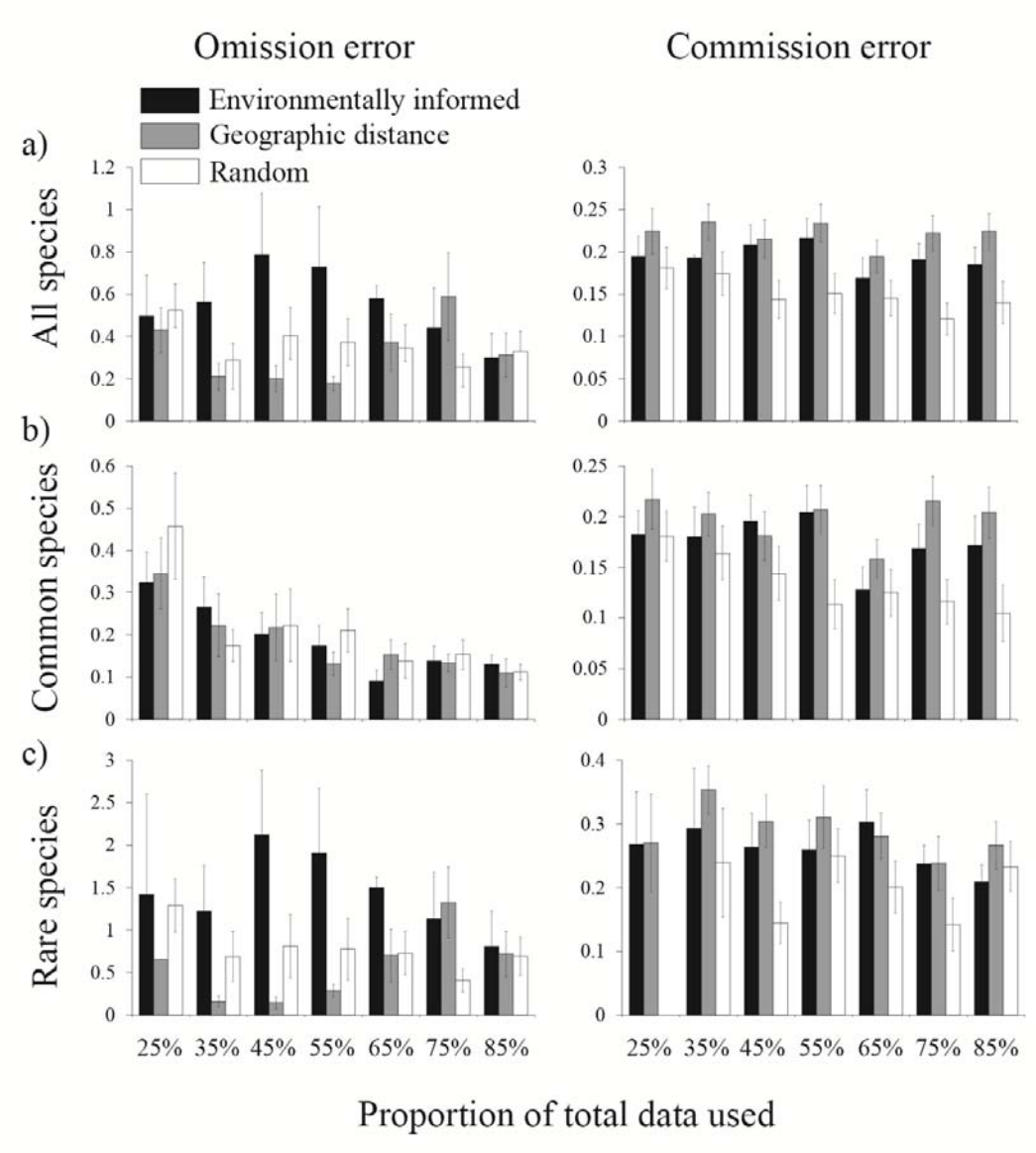
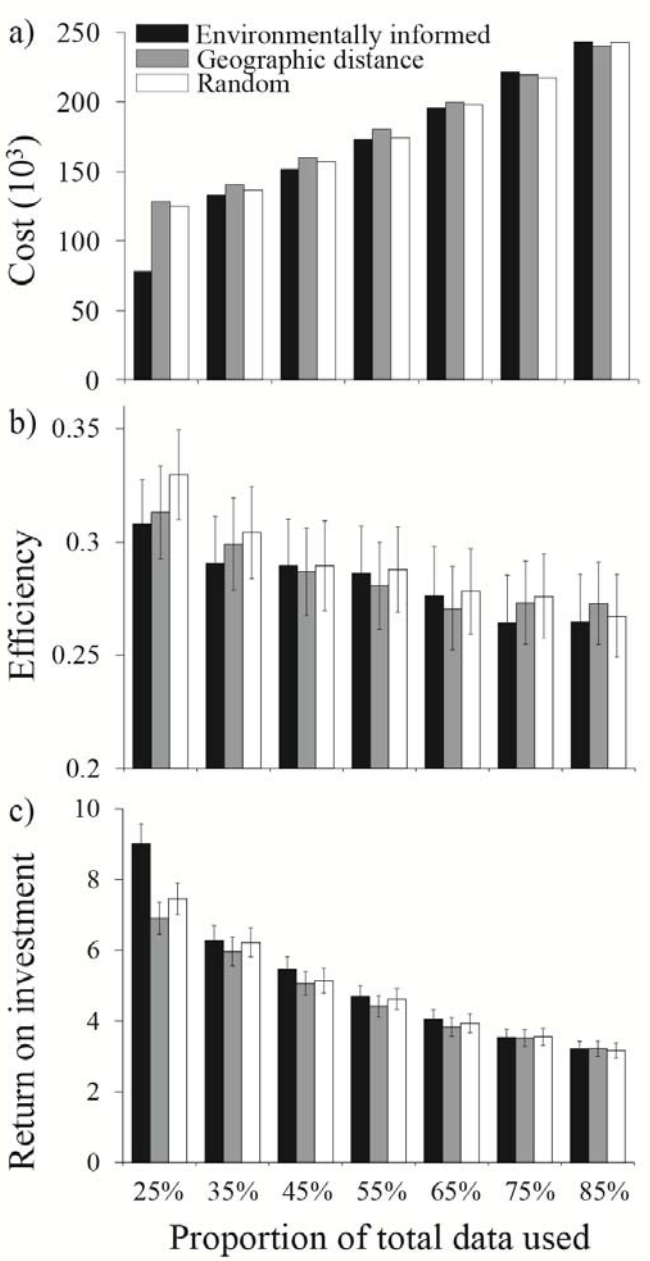

Omission error
Commission error
Environmentally informed
Geographic distance
Random
a)
All species
b)
Common species
c)
Rare species
25% 35% 45% 55% 65% 75% 85%
Proportion of total data used

Figure 5


a)
Environmentally informed
Geographic distance
Random
Cost (10^3)
250
200
150
100
50
0
b)
Efficiency
0.35
0.3
0.25
0.2
c)
Return on investment
10
8
6
4
2
0
25%
35%
45%
55%
65%
75%
85%
Proportion of total data used