Natureza & Conservação
Brazilian Journal of Nature Conservation

Research Letters

Natureza & Conservação 8(2):145-150, December 2010
Copyright© 2010 ABECO
Handling Editor: Rafael Dias Loyola
doi: 10.4322/natcon.00802007

Dealing with Data Uncertainty in Conservation Planning

Kerrie Ann Wilson

The University of Queensland, School of Biological Sciences, Brisbane, Queensland, Australia

Abstract

Conservation planning analyses often employ data on biodiversity and sometimes vulnerability and these data are generally assumed to be accurate and correct. Here, different ways of exploring uncertainty associated with typical input data used for conservation planning are illustrated. First the uncertainty associated with predicted species distribution data is measured, summarised, and visualised. Second the uncertainty associated with the choice of vulnerability model is evaluated using Bayesian Model Averaging and the implication of this uncertainty on inference about key relationships associated with native forest conversion is assessed. The approaches used to assess uncertainty are applicable to any conservation planning exercise and such assessments will increase confidence in the products developed and reduce the risk that conservation effort is misdirected.

Key words: Conservation Planning, Bayesian Model Averaging, Species Distribution Models, Uncertainty, Vulnerability.

Introduction

Conservation planning is the process of locating and designing networks of terrestrial and marine protected areas to protect biodiversity *in situ*. The aim is to efficiently meet quantitative targets within a system of representative and complementary areas. Such analyses often rely upon data on biodiversity values and the vulnerability of these values to threatening processes (Moilanen *et al.* 2009). A commonly disregarded source of uncertainty in the planning process is the uncertainty associated with the input data.

In the face of uncertainty, conservation planners could adopt a risk-averse attitude. Being risk-averse is to avoid risks associated with uncertainty and preferentially seek circumstances in which the risk is minimised. One risk-averse response to errors and uncertainty is to undertake further data collection. In a global and dynamic economic climate, however, such an adjournment entails uncertainty and risks of its own: ecosystems may be degraded and species may go extinct. For this reason, and in line with the precautionary principle, conservation decisions must be made and actions must be initiated in the face of uncertainty (Ludwig *et al.* 1993; Moilanen *et al.* 2006). Conservation planners should therefore identify the errors and uncertainties in the planning process and where necessary, evaluate the sensitivity of conservation planning outcomes to these.

Errors and uncertainty associated with biodiversity data

Species locality data are commonly used for conservation planning. However, these are often biased to parts of a region or towards particular species, are incomplete, or contain errors. The selection of conservation areas cannot generally be delayed pending acquisition of improved species locality data so predicted species distribution data are increasingly relied upon (Elith & Leathwick 2009). Predicted species distributions can be derived by modelling the relationship between species locality data and mapped environmental information, such as climate, terrain, and soil (Guisan & Zimmermann 2000). Predicted species distribution data can contain errors and exhibit uncertainty, due to errors in the species locality data and mapped environmental information (McKelvey & Noon 2001). In addition, errors might be introduced due to decisions made during the modelling process (Diniz-Filho *et al.* 2009; Wilson *et al.* 2005c). Often, however, predicted species distribution data are presented as accurate digital representations without measures of precision. While measures of model accuracy based on misclassifications can be useful to identify spatial locations where errors occur, most model evaluation statistics assess overall model performance and do not provide information about the spatial distribution of prediction uncertainties (Fielding & Bell 1997).

Confidence intervals, which express the uncertainty associated with parameter estimation, can be generated for probabilities of species occurrence. These intervals can

***Send correspondence to:** Kerrie Ann Wilson
The University of Queensland, School of Biological Sciences,
Brisbane, Queensland, 4072, Australia
E-mail: k.wilson2@uq.edu.au

be used to summarise and present spatially the uncertainty associated with predicted species distribution data. For example, conservation planners might be interested in predictions that have wide confidence intervals, as these might be the most unreliable sources of data for that species and therefore require further field sampling. Alternatively, conservation planners might be most interested in areas where the certainty is greatest as these might represent lower risk options for investment. The first objective of this paper is to measure, summarise, and visualise the uncertainty associated with predicted species distribution data.

Errors and uncertainty associated with vulnerability data

Areas of the landscape that are priorities for conservation should be those that are both vulnerable to threats and that if lost or degraded, will result in conservation targets being compromised (Araújo *et al.* 2002). The vulnerability of sites can be predicted using quantitative models that relate the extent of a past threat to characteristics believed to have predisposed areas to the threat (for example, proximity to roads or soil type). Presently unaffected areas that share these characteristics are identified as vulnerable (Wilson *et al.* 2005b).

An assessment of the uncertainty associated with vulnerability information is important in order to minimise the misallocation of conservation effort. For example, if vulnerability is overestimated, scarce resources could be allocated to areas that are not in urgent need of protection. Conversely, if vulnerability is underestimated, areas that are threatened could be overlooked possibly resulting in their biodiversity values being reduced or eliminated. Estimating and exploring the uncertainty associated with vulnerability assessments could determine the sensitivity of predictions to input data and assumptions. In particular, model structural uncertainty, which arises when predictions are based on a single, 'best' model of a particular structure could be investigated (Burnham & Anderson 2002, page 154). An approach to dealing with model uncertainty is to avoid selecting a single, 'best' model but rather average over a number of possible models. Model averaging can be applied in both Frequentist and Bayesian frameworks (Araujo & New 2007). Bayesian approaches to model averaging weight each model according to its posterior probability, as determined by the support it receives from the observed data and prior knowledge. Madigan & Raftery (1994) and Raftery *et al.* (1997) found that model-averaged predictions are more accurate than those obtained from a single, 'best' model.

The second objective of this paper is to use Bayesian Model Averaging to assess the uncertainty associated with the choice of vulnerability model describing the conversion of native forest to plantations in Southern Chile. The effect of model uncertainty on inference about native forest conversion and on the vulnerability predictions is investigated.

Materials and Methods

Errors and uncertainty associated with biodiversity data

Predictions of occurrence of *Acacia ausfeldii*, a rare plant endemic to the Box-Ironbark region of Victoria (Australia) were generated using logistic regression by modelling the relationship between survey data for this species and environmental variables (Wilson *et al.* 2005c). In order to quantify the uncertainty associated with the probabilities of species occurrence, Wald statistic confidence intervals for the logit were calculated (Hosmer & Lemeshow 1995). This uncertainty was then summarised and visualised by (1) mapping the upper and lower bounds on the probabilities independently of the probabilities of occurrence, (2) mapping the width of the confidence intervals, where a large width indicates high uncertainty, and (3) depicting the probabilities of occurrence and the uncertainty associated with these simultaneously. These methods to summarise and visualise the uncertainty associated with predictions of species occurrence could be applied to the data from any species distribution modeling method that generates predictions of occurrence and associated measures of uncertainty (Elith *et al.* 2006).

Errors and uncertainty associated with vulnerability data

The Bayesian approach to model averaging involves calculating predictions under each possible model. These predictions are then weighted by the posterior probability (degree of belief) of each model (Hoeting *et al.* 1999). A model-averaged prediction for a particular outcome (Δ) is obtained via:

$$P(\Delta \mid D) = \sum_{k=1}^{k} P(\Delta \mid S_k, D)\, P(S_k \mid D) \qquad (1)$$

where $P(\Delta|S_k, D)$ is a posterior prediction of the outcome (Δ) according to model S_k and the data (D) and $P(S_k|D)$ is the posterior probability of model S_k, given the data (D) and prior knowledge.

A single, "best" model describing the conversion of native forest to plantation in south central Chile was developed in order to identify areas of native forest vulnerable to conversion (Wilson *et al.* 2005a). The explanatory variables available to calibrate the model were soil type, annual rainfall, minimum annual temperature, slope, altitude, latitude, distance to towns, distance to roads, and distance to timber mills (with distance to towns, distance to timber mills and minimum annual temperature found to be correlated).

While BMA may appear an intuitively attractive means to account for model uncertainty, there are three main difficulties associated with its implementation. First, when the number of models is large the direct evaluation of model-averaged predictions P(Δ|D) is computationally infeasible. For example, if the number of explanatory variables (v) is 20, the number of possible models is 2^v (approximately 1 million). However, usually only a small number of these models will receive support from the data. Identifying a subset of parsimonious, data-supported models (termed Occam's Window, Madigan & Raftery 1994) greatly reduces the number of models requiring summation. This subset can be identified by removing any model that has a posterior probability far less than the best model and then removing any model that has a lower posterior probability than any simpler sub-model. The leaps and bounds algorithm (Furnival & Wilson 1974) provides a tool for searching for this subset.

The second obstacle to BMA is that the higher order integrals implicit in calculating posterior model probabilities can be analytically intractable, but approximate methods of integration have been developed (Kass & Raftery 1995). One asymptotic approximation is the Schwarz criterion (or BIC approximation, Schwarz 1978), which is reasonably accurate for large samples and computationally efficient (Volinsky & Raftery 2000).

Specification of the prior belief that model S_k is the true model presents the third challenge. Clyde (2000) presented objective prior distributions for BMA of GLMs. These distributions, referred to as the Calibrated Information Criterion prior distributions (referred to herein as Clyde's CIC prior distributions) include standard model selection criteria such as BIC, AIC (Akaike Information Criterion), and RIC (Risk Inflation Criterion). The use of Clyde's CIC prior distributions permits model inference based on maximum likelihood theory.

To perform the BMA analysis, the S-PLUS function [BMA.GLM] (available from http://www.research.att.com/~volinsky/bma.html) was used. Three model spaces were evaluated. Each model space contains one of the correlated explanatory variables. To assess the fit of the models, the deviance was converted into an estimated D^2 with values between 0.2 and 0.4 representing a very good model fit (Wrigley 1985). The Receiver Operating Characteristic (ROC) curve was employed to measure the discrimination ability of the models (Pontius & Schneider 2001), with areas greater than 0.8 indicating good discrimination. A random selection of the data was withheld to validate the models.

Results

The probability of occurrence of *A. ausfeldii* was mapped along with the upper and lower bounds on the predictions (Figure 1). The range of the confidence interval width was from 0.00 to 0.42, with the upper end of this range indicating greatest uncertainty (Figure 2). The data and its uncertainty were displayed simultaneously (Figure 3). No areas were identified as having high probability of occurrence and low uncertainty. This is likely to be partially related to the greater certainty associated with areas predicted to have a high probability of occurrence, but also because confidence intervals around predicted probabilities of 0.5 tend to be wider than those closer to one or zero. This effect is due to the binomial nature of the response variable.

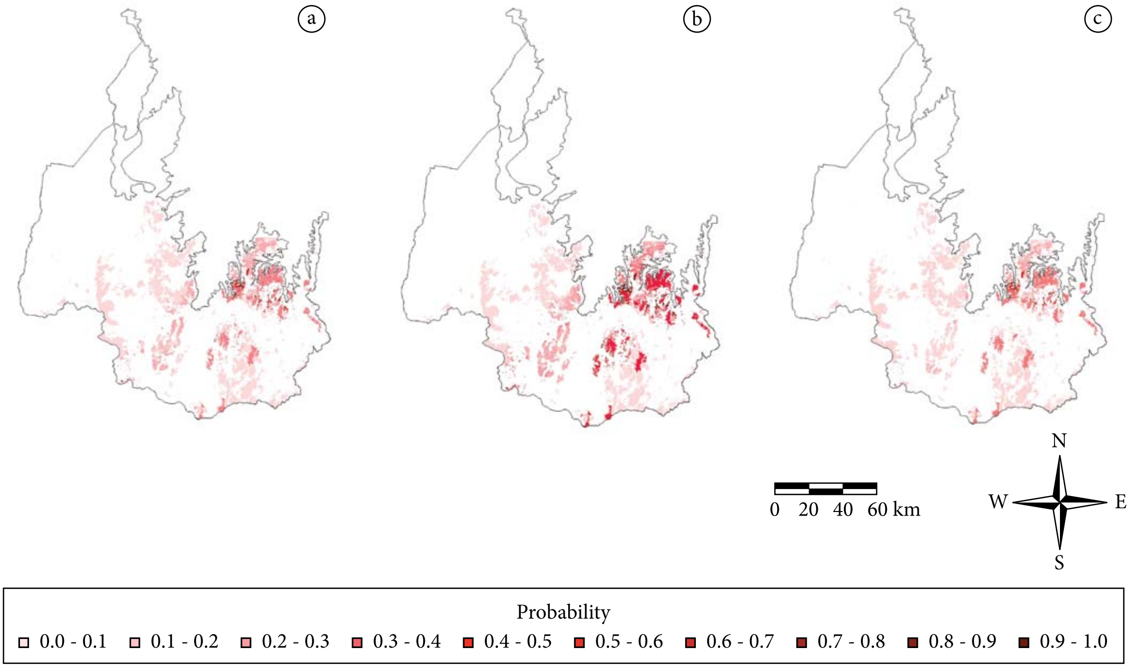

Figure 1. a-c) The probability of occurrence of *A. ausfeldii* a) expected probability of occurrence b) upper probability of occurrence c) lower probability of occurrence. The darker areas indicate a higher probability of occurrence.

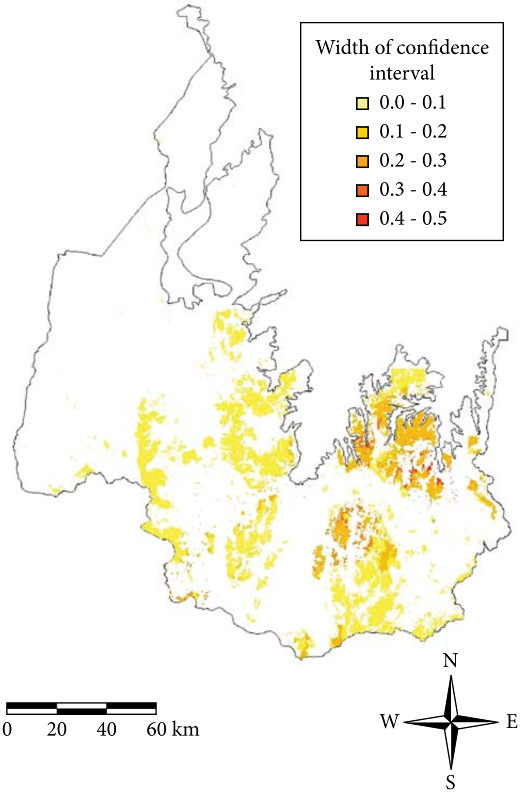

Figure 2. Width of the confidence interval associated with the probabilities of occurrence of *A. ausfeldii*. Darker areas depict a greater confidence interval width.

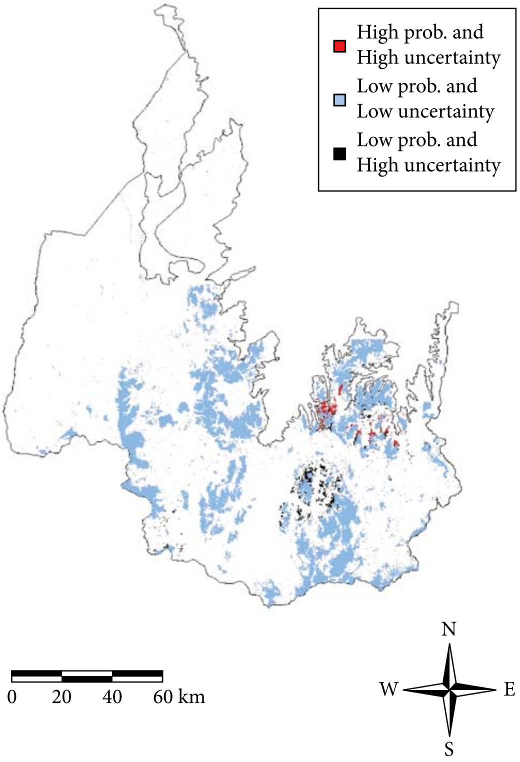
🟥 High prob. and High uncertainty
🟦 Low prob. and Low uncertainty
■ Low prob. and High uncertainty
0 20 40 60 km
N
W
E
S

Figure 3. The probabilities of occurrence of *A. ausfeldii* and the associated uncertainty displayed simultaneously. A high probability of occurrence was anything greater than 0.5. A threshold of 0.2 was chosen for the high uncertainty values, as the range of the confidence interval width for the probabilities was from 0 to 0.42.

When the three model spaces were submitted to BMA, no model subsets were identified. That is, for each of the three model spaces, the model with all the variables included was identified as the favoured model. The model which included distance to towns was identified to be the best model in terms of deviance explained. However, the three models performed equally well (deviance explained ranged from 42.1 to 42.7%). The models also had equal discrimination ability as measured by the area under the ROC curve of 0.95 for each. Under the three models, the relationships between each of the explanatory variables and the presence or absence of native forest conversion were consistently positive or negative and the parameter coefficients were almost identical. Generally, the model that included minimum annual temperature gave odds that were slightly more rational (for example, greater odds of conversion occurring at low elevations).

Discussion

Research within the field of conservation planning has focused on the development of theories and tools to design reserve networks that protect biodiversity in an efficient and representative manner. Whilst much progress has been made in this regard, within the field of conservation planning there has been limited explicit assessment of the uncertainty associated with input data.

Here, the uncertainty associated with predicted species distribution data is summarised and visualised in order to highlight information that is important to conservation planners. Displaying separately, the predicted probability of species occurrence and the upper and lower bounds on these predictions was the simplest means employed to do this. This mapping procedure will be cumbersome when there are many species of interest. Further, little information is provided on how the probabilities of occurrence and their associated uncertainty correspond. Combining the probabilities of occurrence and the estimates of their uncertainty using Boolean operators enabled the predicted species distribution data and its associated uncertainty to be depicted on a single map. This approach takes advantage of subjective associations by which people deal with a sequence of hues or particular hues. For example, red might represent areas that are important but have high risk (high probability and high uncertainty). The difficulty associated with this procedure is choosing an appropriate threshold for the display of the different categories.

An alternative procedure would involve using a continuum of values to represent the probabilities of occurrence and their associated uncertainty by employing varying colour hues and saturations respectively (Davis & Keller 1997). For example, high probability of occurrence could be represented by green and low probability by red. The amount of uncertainty associated with the predictions could be represented by the colour saturation, with hazy colours representing greater uncertainty and vivid colours representing greater certainty. This approach provides the most information on how the probabilities of occurrence and their associated uncertainty correspond and avoids the need to define thresholds, but produces a complex interpretation key.

Whilst a single, "best" model of native forest conversion was obtained (Wilson *et al.* 2005a), it was recognised that this might be only one of many possible models that perform equally well, but result in divergent predictions. Here, the single, "best" model of vulnerability was extended to incorporate model structural uncertainty, specifically that associated with the choice of explanatory variables included in the model. While there is uncertainty associated with model choice, each of the models performed well and the impact on inference and prediction was negligible.

The approaches used here to assess the uncertainty associated with vulnerability data and and to visualise the uncertainty associated with predicted species distribution data are broadly applicable to conservation planning exercises. Other sources of uncertainty in conservation planning include those associated with the scale, resolution, and accuracy of input data and these also require consideration. The assessment of uncertainty in conservation planning can be used to increase confidence in the use of species and vulnerability data in conservation planning, help reduce the risk that conservation effort is misdirected, and increase the likelihood that conservation decisions are made that are optimal for biodiversity conservation.

References

Araujo MB & New M, 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22:42-47.

Araújo MB, Williams PH & Turner A, 2002. A sequential approach to minimise threats within selected conservation areas. *Biodiversity and Conservation*, 11:1011-1024.

Burnham KP & Anderson DR, 2002. *Model selection and multi-model inference*: a practical information-theoretic approach. 2nd ed. New York: Springer.

Clyde M, 2000. Model uncertainty and health effect studies for particulate matter. *Environmetrics*, 11:745-763.

Davis TJ & Keller CP, 1997. Modelling and visualising multiple spatial uncertainties. *Computers and Geosciences*, 23:397-408.

Diniz-Filho JAF *et al.*, 2009. Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, 32:897-906.

Elith J & Leathwick JR, 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677-697.

Elith J *et al.*, 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29:129-151.

Fielding AH & Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24:38-49.

Furnival GM & Wilson RW, 1974. Regressions by leaps and bounds. *Technometrics*, 16:499-511.

Guisan A & Zimmermann NE, 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135:147-186.

Hoeting JA, Madigan D, Raftery AE & Volinsky CT, 1999. Bayesian Model Averaging: a tutorial. *Statistical Science*, 14:382-417.

Hosmer DW & Lemeshow S, 1995. Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine*, 14:2161-2172.

Kass RE & Raftery AE, 1995. Bayes Factors. *Journal of the American Statistical Association*, 90:773-795.

Ludwig D, Hilborn R & Walters C, 1993. Uncertainty, resource exploitation, and conservation: lessons from history. *Science*, 260:17-36.

Madigan D & Raftery AE, 1994. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89:1535-1546.

McKelvey KS & Noon BR, 2001. Incorporating Uncertainties in Animal Location and Map Classification into Habitat Relationships Modeling. In: Hunsaker CT, Goodchild MF, Friedl MA & Case TJ (Ed.). *Spatial Uncertainty in Ecology*: Implications for Remote Sensing and GIS Applications. New York: Springer-Verlag. p. 72-90.

Moilanen A *et al.*, 2006. Planning for robust reserve networks using uncertainty analysis. *Ecological Modelling*, 199:115-124.

Moilanen A, Wilson KA & Possingham HP, 2009. Spatial conservation prioritisation: quantitative methods and computational tools. Oxford: Oxford University Press.

Pontius RG & Schneider LC, 2001. Land cover change model validation by an ROC method for the Ipswich watershed Massachusetts, USA. *Agriculture, Ecosystems & Environment*, 85:239-248.

Raftery AE, Madigan D & Hoeting JA, 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179-191.

Schwarz G, 1978. Estimating the dimension of a model. *Annals of Statistics*, 6:461-464.

Volinsky CT & Raftery AE, 2000. Bayesian information criterion for censored survival models. *Biometrics*, 56:256-262.

Wilson KA *et al.* 2005b. Measuring and incorporating vulnerability into conservation planning. *Environmental Management*, 35:527-543.

Wilson KA, Newton AN, Echeverría C, Weston CJ & Burgman MA, 2005a. A vulnerability analysis of the temperate forests of south central Chile. *Biological Conservation*, 122:9-21.

Wilson KA, Westphal MI, Possingham HP & Elith J, 2005c. Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, 122:99-112.

Wrigley N, 1985. *Categorical Data Analysis for Geographers and Environmental Scientists.* New York: Longman.

Received: August 2010
First Decision: September 2010
Accepted: September 2010