SC/66b/EM/04 Rev1

Progress report of the intersessional corresponding group Applications of species distribution models (SDMs) since 66a IWC/SC

Hiroto Murase, Ari Friedlaender, Natalie Kelly, Toshihide Kitakado, John McKinlay, Daniel M. Palacios, Debra Palka

INTERNATIONAL
WHALING COMMISSION

Progress report of the intersessional corresponding group "Applications of species distribution models (SDMs)" since 66a IWC/SC

HIROTO MURASE[1], ARI FRIEDLAENDER[2], NATALIE KELLY[3], TOSHIHIDE KITAKADO[4], JOHN MCKINLAY[3], DANIEL M. PALACIOS[2], DEBRA PALKA[5]

[1] National Research Institute of Far Seas Fisheries, Japan Fisheries Research and Education Agency (FRA), 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan
[2] Marine Mammal Institute, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, 3020 Marine Science Drive, Newport, OR 97365 USA.
[3]Australian Antarctic Division, 203 Channel Highway, Kingston, Tasmania 7050, Australia
[4]Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato-ku, Tokyo, 108-8477, Japan
[5]NOAA Fisheries, Northeast Fisheries Science Center, 166 Water Street, Woods Hole, Massachusetts, 02543, USA

ABSTRACT

The intersessional corresponding group "Applications of species distribution models (SDMs)" was established in 65b IWC/SC. The primary task is to "develop guidelines and recommendations for best practices in modelling steps". During the intersessional period from IWC/SC 66a to 66b, the group conducted preliminary reviews of machine learning methods which are commonly used as SDMs. The following machine learning methods were reviewed: maximum entropy models (MAXENT), genetic algorithms (GA), support vector machines (SVMs), Bayesian networks (BNs) and random forest (RF). The group also considered preliminary framework guideline for SDMs applied to cetaceans. The group intends to complete the review of additional machine learning methods by 67a. and guideline by 67b.

BACKGROUND

Application of species distribution models (SDMs) to cetaceans has proliferated since the 1990s in parallel with the advancement of computing power, software such as geographic information systems (GIS) and statistical techniques. SDMs represent a collection of various statistical models rather than a specific technique. In this manuscript, a statistical model relating occurrence of a species to its environment at a certain time period is termed an SDM. Such a model can be used to predict spatial distributions of the target species once the model is constructed. An SDM is not a mechanistic model that can deal with driving processes of the spatial distributions but an empirical model that can incorporate observed relationships between occurrence of species and their environment at a certain time period (Palacios *et al.*, 2013). A SDM can also be called as a snapshot model, being static rather than dynamic.

Within the Scientific Committee of the International Whaling Commission (IWC/SC), a generalized additive model (GAM) based SDM was developed in the late 1990s for the purpose of generating a spatially explicit abundance estimate for Antarctic minke whales (*Balaenoptera bonaerensis*) (Hedley *et al.*, 1999). Since then, SDMs have been applied to a variety of species and regions in the IWC/SC to address questions such as reasons for changes in abundance and spatial distribution of baleen whales (e.g. Beekmans *et al.*, 2010; Murase *et al.*, 2013; Williams *et al.*, 2014). These results are used for in-depth assessment of whale stocks (IA). Traditionally, abundance of baleen whales for the purpose of management under the Revised Management Procedure (RMP) have been estimated using a statistical design-based methods, such as the DISTANCE sampling, according to a guideline from the IWC/SC (IWC, 2012). The Sub-Committee on the RMP of the IWC/SC is currently trying to develop a guideline for model-based abundance estimation methods, mainly focusing on GAMs (Hedley and Bravington, 2014). It is expected that the review and developing a guideline are completed by the 2016 annual meeting (IWC, 2016). Although a workshop for the review and training was planned as a pre-meeting to IWC/SC 66a, it was postponed.

Statistical models other than GAMs are also used as SDMs, mainly focusing on prediction of spatial distribution. The Working Group on Ecosystem Modelling (EM) of the IWC/SC recognized the necessity for the development of a guideline on the techniques and underlying assumptions of SDMs based on up-to-date and comprehensive knowledge (IWC, 2015). An intersessional correspondence group was established during in IWC/SC 65b to facilitate this work. The primary task of the group was to "develop guidelines and recommendations for best practices in modelling steps". Estimation of abundance is the main focus in the context of RMP, while explanatory investigation on spatial distribution in relation to the environment is the main focus of EM, although the distinction is not clear-cut. A preliminary review of SDMs applied to baleen whales was carried out by the group intersessionally between IWC/SC 65b and 66a (Murase *et al.*, 2015).

PROGRESS SINCE 66A IWC/SC

In the intersessional period between IWC/SC 66a and 66b, the group conducted (1) preliminary review of machine learning methods applied as SDMs and (2) preliminary considerations on a guideline framework for SDMs applied to cetaceans. The following machine learning methods are reviewed: maximum entropy models (MAXENT), genetic algorithms (GA), support vector machines (SVMs), Bayesian networks (BNs) and random forest (RF). This paper presents the results of these works. These reviews are presented as Appendices of this document. Preliminarily review of boosted regression tree (BRT) would be presented next year. In addition, some thought on a guideline framework for species distribution models (SDMs) applied to cetaceans is provided in a separate Appendix, based on the ten iterative steps in development and evaluation of models proposed by Jakeman *et al.* (2006).

WORK PLAN

The group intends to complete the review of machine learning methods by 67a and guideline by 67b. The group also intends to complete a review of SDMs applied to baleen whales (i.e., an extension of Murase *et al.*, 2015) and submit it to a peer-reviewed journal by 67a.

A guideline for model-based abundance estimation method are currently being developed in the RMP Sub-Committee. A draft guideline submitted to 65b IWC SC (Hedley and Bravington, 2014) mainly dealt with general issues. Although it has on its merit, we consider that development of a guideline specific to GAM in the RMP sub-committee is also beneficial. A review of ensemble modelling is also conducted in the EM working group. These topics together with our review and guideline are interrelated and a synthesis is required within a few years to develop a comprehensive guideline. It is worth to consider a coordination among three groups (model-based abundance [RMP], ensemble modelling [EM] and SDM guideline [EM]) for the development.

REFERENCES

Beekmans, B.W.P.M., Forcada, J.F., Murphy, E.J., de Baar, H.J.W., Bathmann, U.V., Fleming, A.H. 2010. Generalised additive models to investigate environmental drivers of Antarctic minke whale (*Balaenoptera bonaerensis*) spatial density in austral summer. *J. Cetacean Res. Manage.* 11: 115-29.

IWC. 2012. Requirements and guidelines for conducting surveys and analysing data within the revised management scheme. *J. Cetacean Res. Manage.* 13 (Suppl.): 509-17.

IWC. 2015. Annex K1 Report of the Working Group on Ecosystem Modelling. *J. Cetacean Res. Manage.* 16 (suppl.): 277-90.

IWC. 2016. Annex D: Report of the Sub-Committee on the Revised Management Procedure (RMP) *J. Cetacean Res. Manage.* 17 (suppl.) (in press)

Hedley, S., Bravington, M. 2014. Comments on design-based and model-based abundance estimates for the RMP and other contexts. Paper SC/65b/RMP11presented to the 65b IWC Scientific Committee, May 2014. (unpublished). 9pp.

Hedley, S.L., Buckland, S.T., Borchers, D.L. 1999. Spatial modelling from line transect data. *J. Cetacean Res. Manage.* 1: 255-64.

Murase, H., Friedlaender, A., Kelly, N., Palacios, D.M. and Palka, D. 2015. A preliminary review of species distribution models (SDMs) applied to baleen whales. Paper SC/66a/EM3 presented to the 66a IWC Scientific Committee, May 2015. (unpublished). 18pp.

Murase, H., Kitakado, T., Hakamada, T., Matsuoka, K., Nishiwaki, S., Naganobu, M. 2013. Spatial distribution of Antarctic minke whales (*Balaenoptera bonaerensis*) in relation to spatial distributions of krill in the Ross Sea, Antarctica. *Fish. Oceanogr.* 22: 154-73.

Palacios, D.M., Baumgartner, M.F., Laidre, K.L., Gregr, E.J. 2013. Beyond correlation: integrating environmentally and behaviourally mediated processes in models of marine mammal distributions. *Endanger. Species Res.* 22: 191-203.

Williams, R., Kelly, N., Boebel, O., Friedlaender, A.S., Herr, H., Kock, K.H., Lehnert, L.S., Maksym, T., Roberts, J., Scheidat, M., Siebert, U., Brierley, A.S. 2014. Counting whales in a challenging, changing environment. *Scientific Reports* 4.

Appendices

Appendix 1
Friedlaender, A. S. A preliminary review of Maximum Entropy Models (MAXENT) and the applicability to cetacean studies.

Appendix 2
McKinlay, J. A preliminary review of Genetic Algorithms (GA) applied as species distribution models (SDM) and their applicability to cetacean studies.

Appendix 3
Palka, D. A preliminary review of Support Vector Machines (SVMs) applied as species distribution models (SDMs) and the applicability to cetacean studies.

Appendix 4
Murase, H. A preliminary review of Bayesian Networks (BNs) applied as species distribution models (SDMs) and the applicability to cetacean studies.

Appendix 5
Palacios, D.M. A preliminary overview of random forests (RF) and its applicability to species distribution modelling (SDM) with cetaceans

Appendix 6
Murase, H. Some thought on framework of guideline for species distribution models (SDMs) applied to cetaceans.

Appendix 1

A preliminary review of Maximum Entropy Models (MAXENT) and the applicability to cetacean studies

ARI S. FRIEDLAENDER

Marine Mammal Institute, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, 3020 Marine Science Drive, Newport, OR 97365 USA.

INTRODUCTION & OVERVIEW OF SPECIES DISTRIBUTION MODELS WITH MAXENT

Species distribution models aim to estimate the relationship between records of a given species and the environmental characteristics of that location (Franklin 2009, Elith et al. 2011). Predictive models of the geographic distribution of a species have broad application in ecology and conservation in both terrestrial and marine ecosystems (Graham et al. 2004, Phillips and Dudik 2008). What follows is a brief review of one particular method of maximum entropy modeling, Maxent. Based largely on the comprehensive reference article by Phillips et al. (2006) I will present the data requirements, advantages and disadvantages of this technique, and its application to cetacean studies.

When both presence and absence occurrence data exist for a given species and study, more general-purpose statistical methods can be used to quantify the relationships between a species and its environment (Guisan and Zimmerman 2000, Phillips et al. 2006), and these can be used to make predictions about space use, habitat, niche modeling, etc. However, for many systems, only presence data (georeferenced location) are available. For such data, maximum entropy models are one method for determining a species' environmental requirements from a set of occurrence localities together with a set of environmental variables that describe some of the factors that likely influence the suitability of the environment for the species (Brown and Lomolino 1998, Root 1988, Phillips et al. 2006).

Maxent is, at its most basic level, a method for making predictions or inferences from incomplete information (Phillips et al. 2006). Maxent generates presence-only models of species distributions by estimating the probability of distribution relative to maximum entropy (i.e. uniformity). The probability of a species occurrence is constrained as a function of environmental variables included as predictor variables. More precise and detailed explanations of the mathematical models used in Maxent are reviewed, Phillips et al. 2006, Phillips and Dudik 2008, and in Elith et al. 2011.

USABLE EXPLANATORY VARIABLES (PRESENCE/ABSENCE)

In order to generate a model of a species' environmental requirements, Maxent uses a set of occurrence localities (presences). The environmental features that can be used in Maxent to predict a species distribution can be derived from both continuous and categorical variables. Maxent employs a number of features to fit a function of the covariates that include linear, product, quadratic, hinge threshold, and categorical. Explanations of the differences in how these are used by Maxent to derive relationships to covariates are provided in Elith et al. 2011.

ADVANTAGES AND DISADVANTAGES OF BN_S

The following are advantages and disadvantages of Maxent as discussed in Phillips et al. 2006:

Advantages

- Maxent requires only presence data and environmental information for the study area rather than full/dedicated abundance estimates from surveys.
- It can integrate both continuous and categorical environmental variables and incorporate interactions between variables.
- Deterministic algorithms have been developed that converge to the optimal (maximum entropy) probability distribution.
- The resulting probability distributions have a concise mathematical definition that is amenable to further analysis.
- Over-fitting of a species' probability of distribution can be avoided using regularization algorithms.

- The resulting probability distributions based on the distribution of occurrence localities are explicit and can allow for more formal examination of sampling bias.
- Model outputs are continuous, allowing for fine distinctions to be made the model suitability of different areas.
- Maxent can also be applied to presence/absence data by using a conditional model.
- Maxent is a generative, rather than discriminative approach, which has advantages when the amount of training data (e.g. smaller data sets) used is limited.
- As a general purpose modeling method, Maxent has broad appeal across a wide range of applications.

Disadvantages

- Maxent is relatively new and not as mature as GLM/GAM so there are fewer guidelines and methods for estimating error.
- More work needs to be done to determine the effectiveness of avoiding of over-fitting compared with other variable-selection methods.
- Maxent uses an exponential model for probabilities that can give large predicted values for environmental conditions outside the range present in the study area.
- As a stand-alone package, Maxent software is required.

SOFTWARE

Maxent software can be downloaded easily. A number of sites provide links to previous versions 'as-is' with no warranty/guarantees, tutorials, and discussion forums: https://www.cs.princeton.edu/~schapire/maxent/, http://homepages.inf.ed.ac.uk/lzhang10/maxent.html.

APPLICABILITY OF MAXENT TO CETACEAN STUDIES

To date, Maxent has been used a in number of cetacean studies. Because of the types of data that are required (or not required), it is amenable to a wide range of objectives, study areas, and species. Environmental variables may be limiting if remote-sensing is not available for a given area or there is a lack of other data sources to link with occurrence data. In the references below are a number of studies that have used Maxent to describe the distribution of cetaceans or conducted ecological niche modeling of cetaceans in a Maxent framework to generate probabilities of occurrence. Smith et al. (2012) use Maxent to identify humpback whale breeding and calving habitat around the Great Barrier Reef, Funayama et al. (2012) modeled the potential distribution of northern elephant seals, Bombosch et al. (2014) developed habitat models to calculate prediction maps to evaluate how species-specific habitat conditions evolve for humpback and Antarctic minke whales, Pendleton et al. (2012) modeled relationships between right whale occurrence and environmental covariates, Lindsay et al. (2016) developed predictive habitats models using sightings data and environmental varibles to help develop marine protected areas in the South Pacific, and Friedlaender et al. (2011) developed ecological niche models of krill predators in the Antarctic using Maxent to develop indices of niche stability and potential for competition.

REFERENCES

Bombosch, Annette, et al. "Predictive habitat modelling of humpback (*Megaptera novaeangliae*) and Antarctic minke (*Balaenoptera bonaerensis*) whales in the Southern Ocean as a planning tool for seismic surveys." *Deep Sea Research Part I: Oceanographic Research Papers* 91 (2014): 101-114.

Brown, J.H., Lomolino, M.V., 1998. Biogeography, 2nd ed.. Sinauer Associates, Sunderland, Massachusetts.

Elith, Jane, et al. "A statistical explanation of MaxEnt for ecologists." *Diversity and distributions* 17.1 (2011): 43-57.

Friedlaender, Ari S., et al. "Ecological niche modeling of sympatric krill predators around Marguerite Bay, Western Antarctic Peninsula." *Deep Sea Research Part II: Topical Studies in Oceanography* 58.13 (2011): 1729-1740.

Franklin, J. (2009) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Funayama, Kota, et al. "Effects of sea- level rise on northern elephant seal breeding habitat at Point Reyes Peninsula, California." *Aquatic Conservation: Marine and Freshwater Ecosystems* 23.2 (2013): 233-245.

Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Lindsay RE, Constantine R, Robbins J, Mattila DK, Tagarino A, Dennis TE (2016) Characterising essential breeding habitat for whales informs the development of large-scale Marine Protected Areas in the South Pacific. Mar Ecol Prog Ser 548:263-275.

Pendleton, Daniel E., et al. "Weekly predictions of North Atlantic right whale Eubalaena glacialis habitat reveal influence of prey abundance and seasonality of habitat preferences." *Endangered Species Research* 18.2 (2012): 147-161.

Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. "Maximum entropy modeling of species geographic distributions." *Ecological modelling* 190.3 (2006): 231-259.

Phillips, Steven J., and Miroslav Dudík. "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation." *Ecography* 31.2 (2008): 161-175.

Root, T., 1988. Environmental factors associated with avian distributional boundaries. J. Biogeogr. 15, 489–505.

Smith, Joshua N., Grantham, Hedley S., Gales, Nick, Double, Michael C., Noad, Michael J. and Paton, David (2012) Identification of humpback whale breeding and calving habitat in the Great Barrier Reef. *Marine Ecology: Progress Series*, *447* 259-272. doi:10.3354/meps09462

Appendix 2

A preliminary review of Genetic Algorithms (GA) applied as species distribution models (SDM) and their applicability to cetacean studies

JOHN MCKINLAY

Australian Antarctic Division, 203 Channel Highway, Kingston, Australia, 7050.
Contact email: john.mckinlay@aad.gov.au

ABSTRACT

Genetic algorithms (GA) are a stochastic search optimisation technique that iteratively develop a solution using analogues of mechanisms that operate in genetic evolution of natural populations. In the context of species distribution modelling (SDM), they develop rules for probabilistic classification of species presence across a study domain based on observed species presences, absence data (often inferred) and environmental covariates. GA have been applied widely to the problem of SDM, in large part due to the availability of software tailored for this purpose. Approaches for evaluating SDMs have been developing rapidly over the last decade, but despite variation in approaches to model evaluation across different studies there seems to be a consensus of evidence that a popular implementation of GA predicts poorly compared with many other SDM approaches. It is unclear the degree to which this poor performance is a failing of particular software, or genetic algorithms more generally. Applications of GA for cetaceans are rare, and are at the present time unknown for baleen whale species.

INTRODUCTION

Originally developed by Holland (1975), genetic algorithms (GA) are a rule-based optimisation technique for supervised classification. They are thought to work well in situations involving complex interactions between many variables, particularly when derivative-based techniques prove problematic and complete enumeration of a search space for an optimal solution is not practical (Haupt and Haupt, 2004). Advances in computing power over the last 30 years have seen widespread uptake of GA in a variety of disciplines, including for species distribution modelling (SDM) in ecology.

Stockwell and Noble (1992) introduced the earliest system for GA-based rule-set classification of species distributions according to environmental predictors. Termed GARP (Genetic Algorithm for Rule set Production), the Stockwell framework has proved a popular modelling choice in the applied ecological literature, showing over 1,000 Google Scholar citations over the period 2000-2015 for Stockwell and Peters' (1999) exposition of the technique. As ecological applications of GA outside of the GARP framework are rare (D'Angelo *et al*., 1995 and McClean *et al*. 2005 are notable exceptions), further discussion of GA will focus on GARP since this represents the only software implementation of GA currently tailored to SDM. GARP has been successfully used in many ecological contexts, including for predicting distributions of invasive species (Peterson and Vieglais, 2001; Peterson, 2003; Chen *et al*., 2006), disease vectors (Adjemain *et al*., 2006), terrestrial vertebrates (Raxworthy *et al*., 2003; Martínez-Meyer *et al*., 2004) and marine mammals (MacLeod *et al*., 2008).

BRIEF OVERVIEW OF GA

Genetic algorithms are a stochastic search algorithm motivated by the success of biological evolution and the processes of natural selection. They work to iteratively "evolve" a set of rules, evaluated against an objective function, to achieve some defined optimisation goal (Haupt and Haupt, 2004). For purposes of clarity, it is useful to immediately frame this in the context of SDM, in which:

i) The optimisation goal is to obtain the best predictive model of species occurrence based on point locations of species observations and a suite of geographically referenced environmental covariates;

ii) The objective function is typically formulated as the predictive accuracy of a model evaluated against an independent set of test data (i.e. data not used in forming the model rules); and,

iii) The set of rules defines the way predictors relate to the probability of species' occurrence.

At each iteration (i.e. generation), selection is achieved by choosing optimal rule sets from among many candidates based on improvement of the objective function. To a restricted set of the best solutions, GA then applies analogues of reproductive processes, such as mutation and crossover, to create subsequent generations of solutions (Mitchell, 1996). Mutation modifies one or more aspect of a rule from its initial state in order to decrease the likelihood that candidate solutions become trapped at local minima. Genetic recombination is

simulated via crossover operations, in which characteristics from two or more "parent" solutions are combined to form a "child" for the next generation. The basic premise of crossover is that combining several good solutions has the potential to come up with an even better one. Candidate solutions evolve through iterative modification of rule sets, either for a pre-specified (large) number of iterations or until no further improvements in the objective function can be found.

APPLICATIONS OF GAs AS SDMs

GARP belongs to the class of SDM known as presence/background methods (sometimes called presence/pseudo-absence; see Renner *et al*., 2015 for a review of terminology) that make use of species occurrence data but must infer species' absences from areas within the study domain where the species has not been reported (Franklin, 2009). Generally speaking, these methods have arisen to make use of a wide class of opportunistically collected data (e.g. citizen science programs) or historical collections (e.g. museums or herbaria) for which a requirement for absence data was never anticipated. Absences inferred in this way have been coined pseudo-absences since their role is usually to mimic true absences for model estimation purposes. GARP chooses these points randomly from locations where the species has not yet been detected, but for which environmental data are available. Many alternative methods for choosing pseudo-absence points have been suggested since GARP first appeared, and the impact of different approaches on results remains an active SDM research topic (Barbet-Massin *et al*., 2012; Lobo *et al*., 2010; Senay *et al*., 2013). Presence/background data are in many cases known to be suboptimal for the purposes of estimating species' distributions since sightings data can be subject to (often inestimable) observer or sampling biases (Hastie and Fithian, 2013; Renner *et al*., 2015). Nonetheless, presence-background methods have attracted considerable research attention in recent decades, with almost equal attention afforded to overcoming deficiencies of the approach (Phillips *et al*., 2009; Lobo *et al*., 2010; Dorazio, 2014). While the current implementation of GARP on MS Windows systems is constrained to operate on presence-background data (Franklin, 2009), this constraint was not apparent in the original UNIX implementation and so is not an inherent limitation of GARP (Stockwell and Peters 1999), nor of GA in general.

GARP proceeds in an iterative fashion by randomly splitting presence data from within the study domain to create training and test data sets which are used to fit and evaluate the model, respectively, at each iteration. It then applies steps for rule selection, evaluation, testing, and incorporation or rejection of rules (Peterson, 2003). Classification rules are developed based on four methods (Stockwell *et al*., 2006): i) Atomic rules that assume binary decisions based on single values of explanatory variables (e.g. if average temperature < x and habitat type = y then species is present); ii) bioclimatic envelope rules that define environmental tolerances within which the species occurs; iii) range rules that extend ii) above to exclude or deem irrelevant variables not associated with the rule definition; and, iv) logit rules to combine several variables through logit regression defining a species' probability of presence based on environmental gradients. Stockwell and Noble (1992) and Stockwell *et al*. (2006) provide some practical examples of forming each of these rules. At each iteration rules are "evolved" by means of truncation, inducing point changes, or by combining several rules. Predictive accuracy is evaluated at each iteration against the test data (presences) combined with an equal number of randomly selected "pseudo-absences". This process halts once a pre-specified number of iterations is reached, or when no appreciable improvement in predictive accuracy is observed. Due to the stochastic nature of GARP, this iterative process is typically repeated for a large number of replications, with a final result obtained as the average of the best 10-100 model replicates based on lowest omission error (Franklin, 2009).

Approaches for assessing predictive performance of SDM vary dependent on whether a threshold value is chosen to convert continuous predictions of occurrence (akin to probabilities) into binary classifications of presence/absence (Liu *et al*., 2009). Thresholds may be chosen because of the research question being asked (e.g. do we wish to reflect actual or potential distribution? Is a hard classification required for spatial management?), or with reference to balancing omission and commission errors (predicted false positives and false negatives, respectively). Performance measures for threshold-applied model output generally involve examining the misclassification (confusion) matrix, and may include sensitivity and specificity, Cohen's Kappa statistic (Cohen, 1960) that simultaneously considers both omission and commission errors, and more recently the True Skill statistic (TSS). The TSS was proposed by Allouche *et al*. (2006) as an alternative to Cohen's Kappa, which was demonstrated to be dependent on prevalence (McPherson *et al*., 2004). TSS is defined as sensitivity + specificity – 1, and ranges from −1 to +1 where values of 1 indicate perfect agreement and values ≤ 0 indicate discrimination no better than random. Threshold choice naturally impacts assessments of a model's predictive power, so it can be preferable to examine values of TSS over a range of threshold values in order to choose the optimal threshold value maximising the statistic (Liu *et al*., 2011). More recent work suggests that high max(TSS) scores may not necessarily guarantee good performance for the intended purpose of a study, and that TSS profiles over the full range of threshold values, along with spatial maps of uncertainty, should be used to choose a threshold tailored to the study objectives (Ruete and Leynaud, 2015).

The traditional approach for assessing non-threshold model output is to examine the ROC/AUC (Area Under the Curve of the Receiver Operating Characteristic plot) (Hanley and McNeil, 1982), which plots the false positive rate (1- specificity, or commission error) against the true positive rate (sensitivity). This approach has the advantage of removing any subjectivity in deciding a threshold value since the ROC is evaluated across all possible choices of threshold. AUC values of 1 indicate perfect discrimination, while values ≤ 0.5 are no better than random. AUC values > 0.75 have usually been assumed to indicate a model has reasonable predictive power (e.g. sufficient for conservation planning; Elith *et al*., 2006). AUC has a direct probabilistic interpretation, at least for models that make use of true absences, in that it reflects the probability that the model will rank a randomly chosen 'presence' site higher than a randomly chosen 'absence' site (Pearce and Ferrier, 2000). Lobo *et al*. (2008) criticise the use of AUC on several grounds, perhaps most importantly on the basis that the measure can be misleading when false absences are present in data (as might be expected when using techniques that rely on pseudo-absences) and that a single, summary statistic takes no account of the spatial distribution of error. More detailed treatments of ROC/AUC and confusion matrices (and derivatives) in the context of both threshold-dependent and -independent results from presence-absence models are provided by Fielding and Bell (1997), Lobo *et al*. (2008) and Liu *et al*. (2009).

Comparative studies have evaluated the predictive performance of GARP in relation to other methods for SDM. Inviting the participation of experienced SDM researchers, Elith *et al*. (2006) compared 16 methods for their ability to predict the distributions of 226 species from six regions from around the world. More widely known techniques considered included generalised linear models (GLM; McCullagh and Nelder, 1989), generalised additive models (GAM; Hastie, 1993), multivariate adaptive regression splines (MARS; Friedman, 1991), maximum entropy modelling (MaxEnt; Phillips *et al*., 2006), boosted regression trees (BRT; Friedman, 2002), generalised dissimilarity modelling (GDM; Ferrier, 2002; Ferrier *et al*., 2002) and GARP variants DK-GARP (Stockwell and Noble, 1992) and OM-GARP (Muñoz *et al*., 2009). One notable feature of this study was that model performance was assessed against independent test data that comprised both presences and absences. Performance metrics included ROC/AUC, Cohen's Kappa and point bi-serial correlation (COR) (Zheng and Agresti, 2000) between observations in the presence-absence test data and model predictions. Generalised linear mixed models (Breslow and Clayton, 1993) were used to assess aggregate model behaviour across all species considered. Results indicated three broad groups of methods that were characterised as being high, intermediate or poor predictors of occurrence. Along with several presence-only methods, DK-GARP was classified as having poor predictive ability with AUC values typically < 0.7 and COR values < 0.18. OM-GARP fared somewhat better, being judged as having intermediate predictive ability and grouping with methods such as GAM, GLM and MARS. Highest performing methods in this study included BRT, GDM and MaxEnt.

Several other studies seem to confirm the findings of Elith *et al*. (2006) in relation to the performance of GARP. Pearson *et al*. (2006) compared nine SDM methods for modelling the current and potential future distribution under predicted climate change for four species of *Proteaceae*. Methods examined included artificial neural networks, GAM, GLM and GARP, as well as several less common techniques. Model performance was assessed based on AUC and Kappa scores, with GARP performing poorly compared with all other methods. Peterson *et al*. (2007), comparing GARP and MaxEnt based on AUC, found that differences in predictive performance were apparent depending on whether the goal was interpolation within a study domain (MaxEnt did best, GARP tended to over-predict spatially) or prediction was conducted outside the study domain used to train the model (GARP performed better, though not significantly so). However, Phillips (2008), finding these results equivocal, undertook a re-examination of the data used in Peterson *et al*. (2007) to show that the original work inappropriately selected background data. Tsoar *et al*. (2007) examined six presence-only/presence-background methods, including GARP, using data on 42 species of snails, birds and bats in Israel. While GARP proved comparable and sometimes superior to other methods, the suite of methods considered in the study were generally not among the best performers identified in other studies (i.e. GARP did well amongst a group of relative non-performers). Finally, Elith and Graham (2009) compare the predictive performance of five alternate SDM methods, including GARP. Three methods used presence-absence data (logistic regression, BRT and random forests), with the remainder using presence-background data (MaxEnt and GARP). Results indicated the predictive performance of GARP was inferior to other machine learning methods such as random forests, BRT and MaxEnt, as well as more traditional regression-based approaches such as GLMs and GAMs. Of particular note, Elith and Graham (2009) show that GARP has difficulty in correctly modelling unordered categorical predictors, recommending that such variables should be presented in binary format (i.e. using indicator variables to represent category levels).

ADVANTAGES AND DISADVANTAGES OF GA

Advantages

- GA solutions are not reliant on derivative-based optimisation techniques, allowing the method to search highly complex optimization surfaces with multiple local minima.
- Methods lend themselves to distributed computing.
- Alternate, equally plausible solutions are captured for examination.

Disadvantages

- Computationally intensive.
- No direct model-based estimates of uncertainty.
- Care must be taken to resolve issues of premature convergence (a few comparatively good but non-optimal solutions come to dominate the population) and slow finishing (the population of solutions has largely converged but a global solution remains elusive). Beasley *et al*. (1993) give specific advice to address these circumstances.
- Current implementations of GARP do not easily accommodate categorical variables.
- Perhaps most importantly, predictive performance has been shown to be inferior to other available SDM approaches.

SOFTWARE

Part of the popularity of GARP has almost certainly been due to the availability of user-friendly software for Windows platforms (DK-GARP, see http://www.nhm.ku.edu/desktopgarp). No updates to the web site appear to have occurred since 2007, implying limited development of the software since that time.

GARP has also been rewritten in the openModeller cross-platform environment for species distribution modelling (OM-GARP, see http://openmodeller.sourceforge.net/) (Muñoz *et al*., 2009). The developer documentation states that this implementation fixes several errors in the original code, including problems relating to numerical precision and application of rules during solution evolution. However, problems associated with incorporating categorical predictors remain.

Several packages implementing GA are available in R, details of which can be found on the CRAN task view for optimisation (https://cran.r-project.org/web/views/Optimization.html). At the present time, these R packages for GA are not known to have been applied to SDM.

APPLICATION OF GA TO CETACEAN STUDIES

Cetacean SDM based on GA are relatively rare, but not unknown. For a variety of fish and cetacean species, Ready *et al*. (2010) compare several different methods for SDM including GARP, Relative Environmental Suitability (RES) (Kaschner *et al*., 2006), MaxEnt, GLM and GAM. Of direct relevance to this review, data were modelled for southern bottlenose whale (*Hyperoodon planifrons*) and fin whale (*Balaenoptera physalus*) from the Southern Ocean, and harbour porpoise (*Phocoena phocoena*) from the North East Atlantic. Data set sizes varied considerably between species, and estimated prevalence was lower for the whale species (< 0.05) than for harbor porpoise (0.28). OM-GARP models were fitted using the default settings. Environmental predictors available to the study included variants on depth, sea surface temperature, salinity, proportional ice cover and primary productivity. Models results were compared using ROC-AUC based on a 75%-25% training-test split of the data. Using an approach described in Kaschner *et al*. (2006), Monte Carlo methods were used to assess the significance of Spearman's rank correlation between model predictions and relative abundance estimates based on independent effort-corrected sightings data. Results from different SDM methods were variable across fish and mammal species, but GARP consistently performed poorly on both evaluation metrics described above. Of the nine fish and three cetacean species considered, in no instance did the ROC-AUC for GARP models exceed 0.7, and Spearman's correlation between predicted model probabilities and independent survey data was non-significant for all species.

Building on work presented in Mandleberg (2004), MacLeod *et al*. (2008) compared four SDM methods for modelling the occurrence of the harbour porpoise (*Phoceoena phocoena*) in the Sea of Hebrides, Scotland. The methods compared were GARP (using the default settings of DK-GARP), GLM, a PCA-based approach (Robertson *et al*., 2001) and a multivariate technique based on eigen-decomposition termed ecological niche factor analysis (ENFA; Hirzel *et al*., 2002). Of these, only the GLM made use of true absences available to the study by virtue of repeat surveys along five fixed transects. Environmental covariates used in the analysis

included water depth, seabed slope, standard deviation of seabed slope, aspect of seabed, and distance to land. Transects were systematically divided into smaller sampling units (cells) which were randomly divided into training and test sets in a ratio of 2:1. Models were compared based on ROC-AUC and by comparing the spatial distribution of predictions by considering Pearson correlations between the average predicted probabilities of occurrence for 12 relatively homogenous sub-areas of the study region. Based on these evaluation metrics, the study showed that all four techniques produced statistically equivalent results, with point estimates of AUC in the range 0.74-0.82. Similarly, spatial predictions for the 12 sub-areas were strongly and significantly correlated between all four modelling techniques.

There are no known advantages to using GA (including GARP) in relation to SDM studies of cetacean species. Many of the issues associated with applying SDM to cetaceans are unlikely to be able to be directly addressed through a GA framework, including issues related to paucity of data, observer biases, and a lack of direct links between sightings and environmental correlates during migratory behaviour. In light of these limitations, including the poor predictive performance of GARP shown in several studies, the approach is currently not recommended for developing SDM for cetacean species.

REFERENCES

Adjemian, J.C.Z., Girvetz, E.H., Beckett, L., and Foley, J.E. 2006. Analysis of Genetic Algorithm for Rule-Set Production (GARP) modeling approach for predicting distributions of fleas implicated as vectors of plague, *Yersinia pestis. California. J. Med. Entomol*. 43: 93–103.

Allouche, O., Tsoar, A., and Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS): Assessing the accuracy of distribution models. *Journal of Applied Ecology* 43: 1223–1232.

Barbet-Massin, M., Jiguet, F., Albert, C.H., and Thuiller, W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many?: How to use pseudo-absences in niche modelling? *Methods in Ecology and Evolution* 3: 327–338.

Beasley, D., Bull, D.R. and Martin, R.R. 1993. An Overview of Genetic Algorithms: Part 1, Fundamentals. *University Computing* 15(2): 58–69.

Breslow, N. E. and Clayton, D. G. 1993. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88(421): 9–25.

Chen, P., Wiley, E.O., and Mcnyset, K.M. 2006. Ecological niche modeling as a predictive tool: silver and bighead carps in North America. *Biol. Invasions* 9: 43–51.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psyschol. Meas.* 20: 37-46.

D'Angelo, D.J., Meyer, J.L., Howard, L.M., Gregory, S.V., and Ashkenas, L.R. 1995. Ecological uses for genetic algorithms: predicting fish distributions in complex physical habitats. *Can. J. Fish. Aquat. Sci.* 52: 1893–1908.

Dorazio, R.M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data: Imperfect detection and survey bias in presence-only data. *Global Ecology and Biogeography* 23: 1472–1484.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.

Elith, J., and Graham, C.H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* 32: 66–77.

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

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. II. Community-level modelling. *Biodiversity Conservation* 11: 2309-2338.

Fielding, A.H., and Bell, J.F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 01: 38–49.

Franklin, J. 2009. *Mapping species distributions - spatial inference and prediction*. Cambridge University Press. 320p.

Friedman, J.H. 1991. Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics* 19(1): 1–141.

Friedman, J.H. 2002 Stochastic Gradient Boosting. *Computational Statistics and Data Analysis* 38(4): 367-378.

Hanley, J.A. and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36.

Hastie, T.J. 1993. Generalized additive models. In: Chambers, J.M., Hastie, T.J. (Eds.), *Statistical Models in S*, Chapter 7. Chapman & Hall, London. 608p.

Hastie, T., and Fithian, W. 2013. Inference from presence-only data; the ongoing controversy. *Ecography* 36: 864–867.

Haupt, R.L. and Haupt, S.E. 2004. *Practical Genetic Algorithms*, 2nd Ed. John Wiley & Sons, New Jersey. 253p.

Hirzel, A.H., Hausser, J., Chessel, D., and Perrin, N. 2002. Ecological-Niche Factor Analysis: how to compute habitat-suitability maps without absence data? *Ecology* 83: 2027–2036.

Holland, J.H. 1975. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press. 183p.

Kaschner, K., Watson, R., Trites, A.W. and Pauly, D. 2006. Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. *Marine Ecology Progress Series* 316: 285-310.

Liu, C., White, M., and Newell, G. 2009. Measuring the accuracy of species distribution models: a review. In: *Proceedings of the 18th World IMACs/MODSIM Congress*. Cairns, Australia. pp. 4241–4247.

Liu, C., White, M. and Newell, G. 2011. Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography* 34: 232–243.

Lobo, J.M., Jiménez-Valverde, A. and Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151.

Lobo, J.M., Jiménez-Valverde, A. and Hortal, J. 2010. The uncertain nature of absences and their importance in species distribution modelling. *Ecography* 33: 103–114.

MacLeod, C.D., Mandleberg, L., Schweder, C., Bannon, S.M., and Pierce, G.J. 2008. A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia* 612: 21–32.

Mandleberg, L. 2004. *A comparison of the predictive abilities of four approaches for modelling the distribution of cetaceans*. Mres Thesis, University of Aberdeen, UK, 54pp.

Martínez-Meyer, E., Townsend Peterson, A., and Hargrove, W.W. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography* 13: 305–314.

McCullagh, P., and Nelder, J.A. 1989. *Generalized Linear Models*. Chapman & Hall/CRC. 511p.

McClean, C.J., Lovett, J.C., Küper, W., Hannah, L., Sommer, J.H., Barthlott, W., Termansen, M., Smith, G.F., Tokumine, S., and Taplin, J.R. 2005. African plant diversity and climate change. *Annals of the Missouri Botanical Garden* 92(2): 139–152.

McPherson, J.M., Jetz, W., and Rogers, D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology* 41: 811–823.

Mitchell, M. 1996 *An introduction to genetic algorithms*. Massachusetts Institute of Technology. 158p.

Muñoz, M.E.S., Giovanni, R., Siqueira, M.F., Sutton, T., Brewer, P., Pereira, R.S., Canhos, D.A.L. and Canhos, V.P. 2009. "openModeller: a generic approach to species' potential distribution modelling". *GeoInformatica*. DOI: 10.1007/s10707-009-0090-7

Pearce, J., and Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133: 225–245.

Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P. and Lees, D.C. 2006. Model-based uncertainty in species range prediction. *Journal of Biogeography* 33: 1704–1711.

Peterson, A.T., and Vieglais, D.A. 2001. Predicting Species Invasions Using Ecological Niche Modeling: New Approaches from Bioinformatics Attack a Pressing Problem. *BioScience* 51: 363–371.

Peterson, A.T. 2003. Predicting the Geography of Species' Invasions via Ecological Niche Modeling. *The Quarterly Review of Biology* 78: 419–433.

Peterson, A.T., Papeş, M., and Eaton, M. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography* 30: 550–560.

Phillips, S.J., Anderson, R.P. and Schapire, R.E. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.

Phillips, S.J. 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). *Ecography* 31:272-278.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., and Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* 19: 181–197.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A., and Townsend Peterson, A. 2003. Predicting distributions of known and unknown reptile species in Madagascar. *Nature* 426: 837–841.

Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., and Froese, R. 2010. Predicting the distributions of marine organisms at the global scale. *Ecological Modelling* 221: 467–478.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., and Warton, D.I. 2015. Point process models for presence-only analysis. *Methods in Ecology and Evolution* 6: 366–379.

Robertson, M. P., Caithness, N. and Villet, M. H. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* 7: 15–27.

Ruete, A., and Leynaud, G.C. 2015. Goal-orientated evaluation of species distribution models' accuracy and precision: True Skill Statistic profile and uncertainty maps. *PeerJ PrePrints* 3:e1478 https://doi.org/10.7287/peerj.preprints.1208v1.

Senay, S.D., Worner, S.P., and Ikeda, T. 2013. Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE* 8(8): e71218. doi:10.1371/journal.pone.0071218

Stockwell, D.R.B., Beach, J.H., Stewart, A., Vorontsov, G., Vieglais, D., and Pereira, R.S. 2006. The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling* 195: 139–145.

Stockwell, D.R.B. and Noble, I.R. 1992 Induction of sets of rules from animal distribution data – a robust and informative method of data-analysis. *Mathematics and Computers in Simulation*. 33: 385-390.

Stockwell, D., and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* 13: 143-158.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. and Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* 13: 397–405.

Zheng, B. and Agresti, A. 2000. Summarizing the predictive power of a generalized linear model. *Stat. Med.* 19: 1771-1781.

Appendix 3

A preliminary review of Support Vector Machines (SVMs) applied as species distribution models (SDMs) and the applicability to cetacean studies

DEBRA PALKA

Northeast Fisheries Science Center, 166 Water St. Woods Hole, MA 02543 USA

ABSTRACT

Support Vector Machines (SVMs) use a functional relationship known as a kernel to map training data onto a new hyperspace in which complicated patterns between animal distribution and environmental variables can be more simply represented and then used to predict that pattern using data from a test dataset. The response variable has usually been either presence/absence or even just presence, though more complicated categorizations are possible. SVMs have been applied successfully to text categorization, handwriting recognition, gene-function prediction, and remote sensing classification, demonstrating the utility of the method across disciplines, proving that SVMs produce very competitive results with the best available classification methods. They have infrequently been applied to ecological predications only in the last decade. However, there are no known applications to cetaceans, so far. A brief overview of SVMs, examples of ecological applications, advantages and disadvantages, and available software are provided in this paper.

OVERVIEW

The Support Vector Machine (SVM) method is a type of machine learning method for statistical pattern recognition. That is, supervised learning is performed when a training dataset is analyzed to develop an algorithm that is used to assign results to new examples in a test dataset. SVMs were originally introduced as a binary classifier (Vapnik 1998). Since then it has been extended to situations involving multiple classes, 1-class present only (e.g., untrained algorithms), partially indentified classes, and even regressions. Basically SVM uses a functional relationship known as a kernel to map data onto a new hyperspace in which complicated patterns can be more simply represented. Because SVM are not based on characteristics of statistical distributions there is no theoretical requirement for observed data to be independent, thereby overcoming the problem of auto-correlated observations, although model performance will be affected by how well the observed data represent the range of environmental variables.

In its classical implementation, a 2-class SVM uses two classes (e.g., presence/absence) of training samples within a multidimensional feature space to fit an optimal separating hyperplane in each dimension. In this way, SVM tries to maximize the margin that is the distance between the closest training samples, or support vectors, and the hyperplane itself (Figure 1). The classification can be modeled with a linear or non-linear algorithm. For example, presence of known locations of rare tree species and absent locations without these rare trees, along with the physical and biological characteristics of both types of locations, were used to predict the potential spatial distribution of the rare tree species (Pouteau et al. 2012). Distribution maps of a fish species were modeled from presence/absence data and 19 physical-chemical and environmental variables from freshwater rivers in northern Italy (Tirelli et al. 2012).

The 2-class SVM has been generalized to a multiclass SVM to accommodate data that have been labeled into a finite set of classes. The dominant approach for doing so is to reduce the single multiclass problem into multiple binary classification problems, though one step likelihoods have also been attempted.

Typically if absence data are not available or unreliable, then pseudo-absence data are generated. An example of presence only data are museum-collected locations of animals. An example of potentially unreliable absence data is absence of a mobile species since it is possible the survey just by chance did not see a mobile animals in a particular type of habitat or absence of an invasive species that has not yet spread to an area. To analyze the present-only format data Scholkopf et al. (1999) developed a one-class SVM. For example, Guo et al. (2005) used the one-class SVM methods to map the potential distribution in California of a tree virulent pathogen called Sudden Oak Death. Drake et al. (2006) used presence of 106 species in mountains of the Swiss alps along with nine environmental variables to model their distributions, thus interpreting this as the species ecological niche (a multidimensional environmental space).

If the data are not labeled into categories or only some of the data are labeled, the SVM methodology was expanded to support vector clustering (SVC) which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. An advantage of this method is there are no assumptions on the

number or shape of the clusters in the data (Ben-Hur et al. 2001). In SVC, data points are mapped from data space to a high dimensional feature space using a kernel function. In feature space the smallest sphere is searched for that encloses the image of the data using the Support Vector Domain Description algorithm. This sphere, when mapped back to data space, forms a set of contours which enclose the data points. These contours are interpreted as cluster boundaries, and points enclosed by each contour are associated by SVC to the same cluster.

The basic idea behind support vector regression is to map the data into a high-dimensional feature space via a nonlinear mapping and do linear regression in this space. In essence, linear regression in a high dimensional feature space corresponds to nonlinear regression in a low dimensional space. For example the complex distribution simulated in Figure 2a, was able to be modeled using support vector regression (red line in Figure 2b).

APPLICATIONS AS SDMs

SVMs have been applied successfully to text categorization, handwriting recognition, gene-function prediction, and remote sensing classification, demonstrating the utility of the method across disciplines, proving that SVMs produce very competitive results with the best available classification methods. However, they have been applied to ecological predications only in the last decade and not frequently (examples were mentioned above).

ADVANTAGES AND DISADVANTAGES

Advantages

According to Guo et al. (2005; 2015) and Drake et al. (2006), when compared with traditional statistical or learning models which are based on generation of pseudo-absence data, advantages of SVMs include the following.

- The methods are easy to use. Unlike many other machine learning algorithms, which rely on creativity and extensive tuning of parameters by users, SVMs require a minimum of tuning. SVMs are stable and thus require less model tuning and have fewer parameters then other computational optimization methods.
- Because SVSs are theoretically-based models, combining optimization, statistics and functional analysis to achieve maximum separation, they have many appealing characteristics: SVMs are distribution free making no assumption on the underlying probability distribution; they do not require independent input data (and therefore can overcome the autocorrelation problem); are able to represent various data distribution shapes in feature space (e.g., banana shapes, sphere shapes, or even very irregular shapes); results are free from local minima; they are computationally efficient; and they provide outstanding performance in many situations.

Disadvantages

Guo et al. (2015) noted the major disadvantages was it was computationally complex and slow; difficult to determine optimal parameters when training data is not linearly separable, and difficult to understand the structure of the algorithm. Perhaps it could also be used in describing temporal trends that include inter- and intra-annual variabilities.

SOFTWARE

An award winning library for support vector machines is LIBSVM.

SVMs are also available in many machine learning toolkits, including MATLAB; PRCC SVM and PROC SVMSCORE in SAS; package e1071 offers a R interface to libsvm; SVMlight; kernlab; scikit-learn; Shogun; Weka; Shark; JKernelMachines; OpenCV; openModeller; and others.

A SVC toolbox was written by Dr. Daewon Lee under supervision by Prof. Jaewook Lee. The toolbox is implemented by the Matlab and based on the statistical pattern recognition toolbox (stprtool) in parts of kernel computation and efficient QP solving.

Multiclass SVM analyses can be conducted using Matlab and freeware from Cornell, SVM^{multiclass}.

APPLICABILITY TO CETACEAN STUDIES

To date, SVMs have not been used as SDMs of cetaceans. Though, it appears to be an appropriate tool to investigate developing cetacean SDMs for rarely encountered species or in situations with limited or unreliable effort information.

REFERENCES

A. Ben-Hur, D. Horn, H.T. Siegelmann and V. Vapnik. Support vector clustering. Journal of Machine Learning Research 2:125-137, 2001.

Drake, J.M., Randin, C. and Guisan, A. 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43: 424-432.

Guo, C., Park, Y-S, Lui, Y, and Lek S. 2015. Toward a new generation of ecological modeling techniques: Review and bibliometrics. Chapter 2 in Developments in Environmental Modelling, Volume 27. DOI: 10.1016/B978-0-444-63536-5.00002-8.

Guo, Q., Kelly, M., Graham, C.H. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182:75-90.

Pouteau, R. Meyer, J-Y., Taputuarai, R. and Stoll, B. 2012. Support vector machines to map rare and endangered native plants in Pacific islands forests. Ecological Informatics 9: 37-46.

Pouteau, R. and Collin, A. 2013. Spatial location and ecological content of support vectors in an SVM classification of tropical vegetation, Remote Sensing Letters, 4:7, 686-695. To link to this article: http://dx.doi.org/10.1080/2150704X.2013.784848

Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J.,Williamson, R.C., 1999. Estimation the Support of a High-dimensional Distribution.

Technical Report MSR-TR-99-87, Microsoft Research.Tirelli, T.; Gamba, M., and Pessai, D. 2012 Support vector machines to model presence/absence of Alburnus alburnus alborella (Teleostea, Cyprinidae) in North-Western Italy: Comparison with other machine learning techniques, C. R. Biologies (2012), http://dx.doi.org/10.1016/j.crvi.2012.09.001

Vapnik, V., 1998. Statistical learning theory. Support Vector Machines for Pattern Recognition. John Wiley & Sons, New York.

Yang, Z.R. 2004. Biological applications of support vector machines. Briefings in Bioinformatics. 5(4): 328-338.

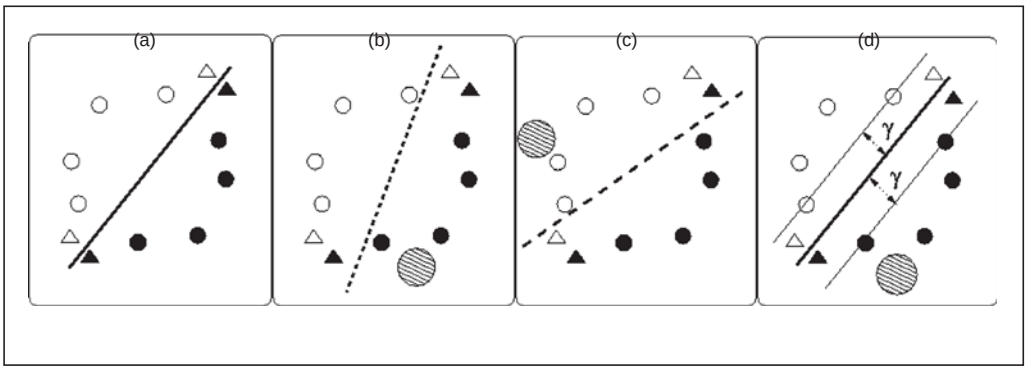

Figure 1: Support vector Kernel function (a) (b) (c) (d) Figure 1: (a) Hyper-plane formed using conventional classification algorithms for the data with a balanced distribution. (b) and (c) Hyper-planes formed using conventional classification algorithms. (d) Hyper-plane formed using SVMs. The open circles represent class A, the filled circles class B and the shaded circle class A or B. The thick lines represent the correct hyperplane for discrimination and the broken thick lines the biased hyper-planes. The thin lines are the margin boundaries. The triangles represent the novel patterns. Gamma (ϒ) means the distance between hyper-plane and the boundary formed by the support vectors. The margin is 2ϒ . From Yang 2004.

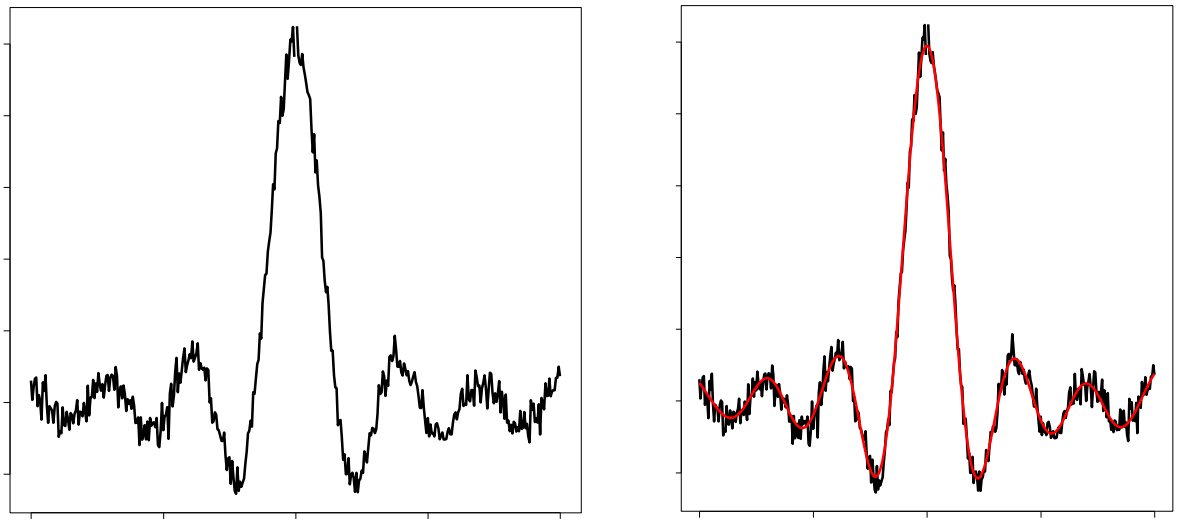

Figure 2: Left side: Pattern simulated using the equation sin(x)/x + rnorm(401,sd=0.03). Right side: Simulated pattern (black line) overlaid with predicted model (red line) using support vector regression estimation.

Appendix 4

A preliminary review of Bayesian Networks (BNs) applied as species distribution models (SDMs) and the applicability to cetacean studies

HIROTO MURASE

National Research Institute of Far Seas Fisheries, Japan Fisheries Research and Education Agency (FRA), 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

ABSTRACT

Bayesian networks (BNs) have been used as species distribution models (SDMs) since early 2000s. This paper presents (1) a brief overview of BNs, (2) examples of applications of BNs as SDMs (3) advantages and disadvantages of BNs, (4) available software and (5) applicability of BNs to cetacean studies. Bayesian networks (BNs) are a kind of probabilistic graphical models that correlative and causal relationship among variables are represented graphically and probabilistically. BNs are categorized as a kind of machine learning methods. BNs have been applied as SDMs for vertebrates but all of them are inland species. The response variables were not abundance but presence and absence. Because of the limitation that variables should be discretized in some extent, utility of BNs for management of cetaceans could be limited as detailed information is lost due the discretization. However, BNs could be useful tool for exploratory research to investigate causal relationship among variables based on expert knowledge which can not be handled by other SDMs.

INTRODUCTION

Bayesian networks (BNs) are a kind of probabilistic graphical models that correlative and causal relationship among variables are represented graphically and probabilistically. BNs are categorized as a kind of machine learning methods. BNs also called in different names like directed Acyclic Graphical Models, Bayesian belief networks and Bayes network. Text books of BNs are available such as Nielsen and Jensen (2009), Pourret *et al.* (2008) and Scutari and Denis (2014). Several reviews and guidelines for BNs in the context of environmental and ecological studies are also available (Aguilera *et al.*, 2011; Chen and Pollino, 2012; Marcot *et al.*, 2006; McCann *et al.*, 2006; Uusitalo, 2007). BNs have used as species distribution models (SDMs) since early 2000s. This paper does not intend to provide full review of BNs because details of BNs can be found in these references. Instead, it provides (1) a brief overview of BNs, (2) examples of applications of BNs as SDMs (3) advantages and disadvantages of BNs, (4) available software and (5) applicability of BNs to cetacean studies.

BRIEF OVERVIEW OF BNS

BNs mainly consist of qualitative and qualitative components. In the qualitative component, causal relationships among variables are represented as directed acyclic graphs (DAGs). A schematic DAG of a simple Bayesian network is shown in Fig. 1. In the graphs, nodes (variables in ellipses) are linked by arcs (also called as edges and arrows) to show causal relationship between nodes. The initial structures of DAGs can be constructed based on known causal relationship (e.g. information from literature) and/or expert knowledge. In the qualitative component, degree of belief expressed as probability of a node in a particular state given states of parents node assuming that conditionally independent of all its non-descendants, given its parents. For example, probability of *C* given *B* in Fig. 1 is calculated based on Bayes' theorem as:

$$: P(D \mid C) = \frac{P(C \mid D)P(D)}{P(C)} \quad (1)$$

APPLICATIONS OF BNS AS SDMS

BNs have been applied as SDMs to a variety of species since early 2000s. Published studies targeting on vertebrates are summarized in Table 1. All of them were targeting inland species. The response variables were not abundance but presence and absence. It seems that number of published papers using BNs as SDMs is small in comparison with other machine learning methods. There is no application to marine vertebrate to date. It should be noted that all studies listed in Table 1 used the software, Netica (Norsys Software Corp., Vancouver, Canada) (see also SOFTWARE section of this paper for related issues).

ADVANTAGES AND DISADVANTAGES OF BNS

The following are advantages and disadvantages of BNs as mentioned in Aguilera *et al.* (2011).

Advantages

- Risk and uncertainty can be estimated more accurately than in models where only values are taken into account because nodes are modelled by means of probability distributions.
- The probability of particular hypothesis can be automatically computed because numeric values are attached to the relationship between the variables.
- The probability distribution of a node given its parents, and even the other way round, the probability distribution of a parents nodes given its child nodes can also obtained once the model is learned. This allows as to know the effects given the causes and the causes given the effects. They can be used as inferential model given this characteristic.
- Expert knowledge can be incorporated in BNs through a participatory modelling procedure because the relations between variables can be visualized easily by the graphical representation of the network and so they can be modified by the experts just by adding or removing variables and links in the graph.
- BNs can model complex systems with a large number of variables.
- BNs can manage missing values in input data and perform the proper predictions with the model built from them.

Disadvantages

- To maintain the accuracy in the estimations and in the network topology, the building process of the network and the parameter estimation requires more data as the number of variables increases.
- BNs can manage continuous data and hybrid of continuous and discrete data but the limitations are too restrictive and the most extend solution is discretization of variables.
- Expert knowledge with an unknown degree of bias and inaccuracy can be easily incorporated in BNs.
- Handling of feedback functions and temporal relationships are not possible.
- Though complex systems can be modelled by BNs, this should be sparingly to avoid crating unwieldy model structures.

SOFTWARE

A number of commercial and noncommercial software is available to build BNs. A list is available from http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html (accessed on 29 April 2016). Some of the software were reviewed in Uusitalo (2007). Several R packages are also available (Scutari and Denis, 2014). However, as mentioned earlier, Netica is only a software to build BNs for SDMs applied to vertebrates.

APPLICABILITY OF BNs TO CETACEAN STUDIES

To date, BNs have not been used as SDMs of cetaceans. Because of the limitation that variables should be discretized in some extent, utility of BNs for management of cetaceans could be limited as detailed information is lost due the discretization. However, BNs could be useful tool for exploratory research to investigate causal relationship among variables based on expert knowledge which can not be handled by other SDMs.

REFERENCES

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R. and Salmerón, A. 2011. Bayesian networks in environmental modelling. *Environ. Modell. Softw.* 26: 1376-88.

Chen, S.H. and Pollino, C.A. 2012. Good practice in Bayesian network modelling. *Environ. Modell. Softw.* 37: 134-45.

Gieder, K.D., Karpanty, S.M., Fraser, J.D., Catlin, D.H., Gutierrez, B.T., Plant, N.G., Turecek, A.M. and Robert Thieler, E. 2014. A Bayesian network approach to predicting nest presence of the federally-threatened piping plover (Charadrius melodus) using barrier island features. *Ecol. Model.* 276: 38-50.

Marcot, B.G., Steventon, J.D., Sutherland, G.D. and McCann, R.K. 2006. Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research* 36: 3063-74.

McCann, R.K., Marcot, B.G. and Ellis, R. 2006. Bayesian belief networks: applications in ecology and natural resource management. *Canadian Journal of Forest Research* 36: 3053-62.

Nielsen, T.D. and Jensen, F.V. 2009. *Bayesian Networks and Decision Graphs*. Springer New York. 448 pp.

Pourret, O., Naïm, P. and Marcot, B. 2008. *Bayesian networks: A practical guide to applications*. Wiley. 446 pp.

Raphael, M.G., Wisdom, M.J., Rowland, M.M., Holthausen, R.S., Wales, B.C., Marcot, B.G. and Rich, T.D. 2001. Status and trends of habitats of terrestrial vertebrates in relation to land management in the interior Columbia river basin. *For. Ecol. Manage.* 153: 63-88.

Rieman, B., Peterson, J.T., Clayton, J., Howell, P., Thurow, R., Thompson, W. and Lee, D. 2001. Evaluation of potential effects of federal land management alternatives on trends of salmonids and their habitats in the interior Columbia River basin. *For. Ecol. Manage.* 153: 43-62.

Scutari, M. and Denis, J.B. 2014. *Bayesian Networks: With Examples in R*. CRC Press. pp.

Smith, C.S., Howes, A.L., Price, B. and McAlpine, C.A. 2007. Using a Bayesian belief network to predict suitable habitat of an endangered mammal – The Julia Creek dunnart (*Sminthopsis douglasi*). *Biol. Conserv* 139: 333-47.

Tantipisanuh, N., Gale, G.A. and Pollino, C. 2014. Bayesian networks for habitat suitability modeling: a potential tool for conservation planning with scarce resources. *Ecol. Appl.* 24: 1705-18.

Uusitalo, L. 2007. Advantages and challenges of Bayesian networks in environmental modelling. *Ecol. Model.* 203: 312-8.

Table 1.

Reference	Region	Number of species considered in studies			
		Fish	Reptile	Bird	Mammal
Raphael *et al.* (2001)	Columbia River basin, US	-	1	16	11
Rieman *et al.* (2001)	Columbia River basin, US	6	-	-	-
Smith *et al.* (2007)	Queensland, Australia	-	-	-	1
Chen and Pollino (2012)	Tasmania, Australia	1	-	-	-
Gieder *et al.* (2014)	Maryland, US	-	-	1	-
Tantipisanuh *et al.* (2014)	Thailand	-	2	15	4

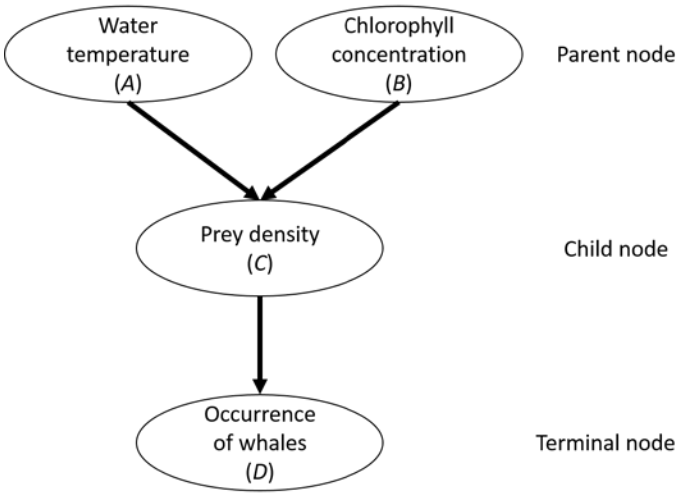

Fig 1. A schematic representation of a simple Bayesian network. In this case, water temperature (*A*) and chlorophyll concentration (*B*) have causal influence on prey density (*C*) which in turn has causal influence on occurrence of whales (*D*). Each node (ellipse representing variable) is linked by arc.

Appendix 5

A preliminary overview of random forests (RF) and its applicability to species distribution modelling (SDM) with cetaceans

DANIEL M. PALACIOS

Marine Mammal Institute, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, 3020 Marine Science Drive, Newport, Oregon, 97365, USA
Contact email: daniel.palacios@oregonstate.edu

ABSTRACT

Random forests (RF) is a machine learning technique that combines many single decision trees in an embedded way to calculate the importance of each predictor. The technique combines the ideas of bagging and random selection of features. From a bootstrap sample, a large number of regression trees are fitted using randomly chosen covariates on each node. Trees are fully grown (rather than pruned), and the results of all trees are averaged for the final prediction. RF performs well in relation to other classification techniques. Use of RF in species distribution modelling (SDM) has proven robust and stable. The technique is being widely used, both as stand-alone and as part of ensemble distribution forecasting on a variety of plant and animal taxa. Software for RF is well developed in the R statistical language. Although RF has apparently not been used in SDM with cetacean survey data to date, the technique is well suited for this purpose, and existing studies from the seabird literature should serve as excellent background.

INTRODUCTION

Random forests (RF) is part of a family of robust methods known as non-parametric. RF was developed by Brieman (2001) and, like other machine learning techniques, it has quickly become popular among the data science community because of its ability to model the complex structure of high-dimensional data sets. At its core, RF is a classification technique that combines many single decision trees in an embedded way to calculate the importance of each predictor. RF is also considered an ensemble method because it aggregates the results of multiple, independently generated classification trees into an averaged prediction. RF performs well compared to other classification techniques such as discriminant functions and neural networks. It has been used for feature selection in bioinformatics (Saeys et al. 2007). The technique was introduced in ecology by Prasad et al. (2006) and by Cutler et al. (2007), and it has become the *de facto* method for supervised dive classification in diving vertebrates (Thums et al. 2008, McIntyre et al. 2011, Photopoulou et al. 2015).

BRIEF OVERVIEW OF RANDOM FORESTS

RF was developed by L. Breiman (2001) as a classification and regression tree (CART) technique. Breiman's layman explanation of RF is as follows:

"Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest)."
https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

RF is a CART method based on bagging. Bagging generates *n* bootstrap samples, builds a model for each, and then averages the resulting models across bootstrap aggregates. The RF algorithm is executed by bootstrapping (with replacement) 63% of the data and generating a "weak learner" based on a CART for each bootstrap replicate. Within the pre-set specification (e.g. node depth and number of samples per node) each CART is unconstrained (grown to fullest) and prediction is accomplished by taking the "majority votes" across all nodes in all random trees. At each replicate the data not used to construct the tree [out of bag (OOB)] are used for validation, providing a quasi-independent validation of model fit. Covariates are randomly selected at each node and variable importance is assessed using the mean decrease in accuracy (MDA) by dividing the standard error by the misclassification rate. The number of covariates randomly selected at each node is defined by *m* (commonly defined as the square root of the number of covariates). The contribution of covariates can also be obtained with the Gini Index. Each time a node split occurs based on a particular variable, the Gini impurity criterion for the two descendent nodes is less than the parent node. The Gini index is calculated by summing the Gini decreases for each individual variable over all trees in the forest.

As with other machine learning techniques, the most commonly used evaluation metric for presence-absence data is the area under the curve (AUC) of a receiver operating characteristic (ROC) plot, where the best-

performing model is selected with the Boyce Index or the Youden Index. The Kappa statistic is used for multi-class models, as is the true skill statistic (TSS). Iterative cross-validation is performed internally as a quasi-independent validation of model fit. Predictions are also subject to model parameter selection. For instance, the two user-defined parameters, the number of trees and the number of randomly selected variables to split the nodes, should be optimized to improve predictive accuracy.

APPLICATIONS OF RF AS SDM

Owing to its unique approach to modelling, the use of RF in SDM has proven robust and stable. The vegetation community has successfully used it at a variety of scales (Prasad et al. 2006, Rehfeldt et al. 2006, Evans and Cushman 2009, Marmion et al. 2009a, Hegel et al. 2010). RF has since been applied to a variety of terrestrial animal taxa at continental scales (Marmion et al. 2009, Howard et al. 2014). For marine taxa, several studies have successfully applied RF to seabird survey (Oppel and Huettmann 2010, Oppel et al. 2012, Renner et al. 2013, Liske et al. 2014) as well as tracking data (Scales et al. 2015). These studies should constitute excellent background material for applications with cetaceans.

An area of focus in SDM has been the assessment of the performance of RF relative to various other modelling techniques (Marmion et al. 2009a, b). RF also has been used to assess the relative performance of models trained on abundance data and those trained on presence-absence data (Howard et al. 2014). RF is included in the suite of techniques for ensemble forecasting of species distributions [along with Generalized Additive Models (GAM), Maximum Entropy (MaxEnt), and Boosted Regression Trees (BRT)] that are implemented in the extremely popular BIOMOD platform (Araujo and New 2007, Thuiller et al. 2009, Thuiller 2014).

APPLICABILITY OF RF AS SDM WITH CETACEANS

To the best of my knowledge, RF has not been used directly as a SDM approach with cetaceans, but this is only a matter of time since the methodology is well established and is well suited for cetacean data sets, either as abundance or as presence-absence. Because seabirds have great similarities with cetaceans in terms of ecology and data collection techniques, the studies discussed in the previous section (Oppel and Huettmann 2010, Oppel et al. 2012, Renner et al. 2013, Liske et al. 2014, Scales et al. 2015) provide excellent background material for applications of RF as SDM with cetaceans. I also note that a recent SDM study used GAM to generate habitat-based cetacean density predictions for a large number of cetacean species in waters of the U.S. Atlantic and Gulf of Mexico, and then implemented RF to resolve ambiguity in models containing similar species due to difficulties in field identification (Roberts et al. 2016).

ADVANTAGES AND DISADVANTAGES

Advantages

- RF is ideal for modelling ordinal and categorical data, including presence-absence (Marmion et al. 2009a, b, Hegel et al. 2010). Additionally, it is among the techniques well suited to deal with the zero-inflated, overdispersed data typical of line-transect abundance surveys (together with negative binomial generalized linear modelling and Hurdle modelling) (Lieske et al. 2014).
- RF makes few assumptions about the distribution of variables, is robust to over-fitting, and is widely recognized to produce predictions that typically outperform traditional regression-based approaches (Breiman 2001, Liaw and Wiener 2002, Prasad et al. 2006, Marmion et al. 2009a, b, Hegel et al. 2010).
- RF predictions can easily be projected into new variable space, making it an appropriate algorithm for projective modelling such as climate change (Rehfeldt et al. 2006).
- In contrast to other dimensionality reduction techniques like those based on projection (e.g. principal component analysis) or compression (e.g. using information theory), RF does not alter the original representation of the variables, but merely selects a subset of them, thus maintaining their interpretability (Saeys et al. 2007).

Disadvantages

- When used as a classifier, ancillary data are necessary for training the RF algorithm and validating classes, which can be problematic for small datasets and which also requires time-consuming visual classification.
- While providing highly accurate predictions, and despite maintaining the original representation of the covariates, RF models can be difficult to interpret (Renner et al. 2013).

- RF can be sensitive to the number of covariates and to the number of trees comprising the classifier (Oppel and Huettmann 2010). Multi-collinearity and imbalance between classes are additional factors that can affect model performance.
- The two major limitations of the ROC as an evaluation metric is that it is only suited for discrete data and few strategies exist for validating more than two classes (presence-absence). For multi-class models the Kappa statistic has been criticized because it is not truly chance-constrained, although a weighting function has been implemented to account for near agreement and adjust for expectation in the frequency of observations.

SOFTWARE

RF methods are well developed in the R statistical language (R Development Core Team 2016). The online Comprehensive R Archive Network maintains "task views" that compile information about libraries ("packages") for popular subjects, including machine learning (http://cran.r-project.org/web/views/MachineLearning.html). The following information is quoted directly from the RF section of the Machine Learning Task View:

"The reference implementation of the random forest algorithm for regression and classification is available in package *randomForest*. Package *ipred* has bagging for regression, classification and survival analysis as well as bundling, a combination of multiple models via ensemble learning. In addition, a random forest variant for response variables measured at arbitrary scales based on conditional inference trees is implemented in package *party*. *randomForestSRC* implements a unified treatment of Breiman's random forests for survival, regression and classification problems. Quantile regression forests *quantregForest* allow to regress quantiles of a numeric response on exploratory variables via a random forest approach. For binary data, *LogicForest* is a forest of logic regression trees (package *LogicReg*). The *varSelRF* and *Boruta* packages focus on variable selection by means for random forest algorithms. In addition, packages *ranger* and *Rborist* offer R interfaces to fast C++ implementations of random forests."

Additional notes:

- Package *randomForest* provides information on variable importance, which is determined by how much prediction error increases when testing data for that variable is permuted while all others are left unchanged (Liaw and Wiener 2002).
- Package *party* uses a RF implementation based on a conditional inference framework (Hothorn et al. 2006a, b, Strobl et al. 2009), which can be useful for accounting for a high degree of correlation between covariates and the potential for biased variable selection. This package offers the *cforest* classifier, which was used by Roberts et al. (2016) to resolve ambiguity among related cetacean species and thus generate separate model predictions.
- Finally, package *biomod2* includes an implementation of RF as part of its ensemble ecological niche modelling tools (Thuiller et al. 2009, Thuiller 2014).

REFERENCES

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22: 42–47. doi:10.1016/j.tree.2006.09.010

Breiman, L., 2001. Random forests. Machine Learning 45: 5–32.

Cutler, D.R., Edwards, T.C. Jr, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

Evans J.S., Cushman S.A. 2009. Gradient modeling of conifer species using random forests. Landsc Ecol 24: 678–683.

Hegel, T.M., Cushman, S.A., Evans, J., Huettmann, F., 2010. Current State of the Art for Statistical Modelling of Species Distributions, in: Cushman, S.A., Huettmann, F. (Eds.), Spatial Complexity, Informatics, 273 And Wildlife Conservation. Springer, Tokyo, pp. 273–311. doi:10.1007/978-4-431-87771-4_16

Hothorn, T., Hornik, K. and Zeileis, A. 2006. party: A Laboratory for Recursive Part(y)tioning.

Hothorn, T., Hornik, K. and Zeileis, A. 2006. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 15: 651–674.

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods in Ecology and Evolution 5: 506–513. doi:10.1111/2041-210X.12184

Liaw, A., Wiener, M. 2002. Classification and Regression by *randomForest*. R News 2/3.

Lieske, D.J., Fifield, D.A., Gjerdrum, C., 2014. Maps, models, and marine vulnerability: Assessing the community distribution of seabirds at-sea. Biological Conservation 172: 15–28. doi:10.1016/j.biocon.2014.02.010

McIntyre, T., Ansorge, I., Bornemann, H., Plötz, J., Tosh, C., Bester, M., 2011. Elephant seal dive behaviour is influenced by ocean temperature: implications for climate change impacts on an ocean predator. Marine Ecology Progress Series 441: 257–272. doi:10.3354/meps09383

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009a. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15: 59–69. doi:10.1111/j.1472-4642.2008.00491.x

Marmion, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009b. The performance of state-of-the-art modelling techniques depends on geographical distribution of species. Ecological Modelling 220: 3512–3520. doi:10.1016/j.ecolmodel.2008.10.019

Prasad, A.M., Iverson, L.R. and Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9: 181–199.

Oppel, S., Huettmann, F., 2010. Using a random forest model and public data to predict the distribution of prey for marine wildlife management. In: Cushman, S.A., Huettmann, F. (Eds.), Spatial Complexity, Informatics, and Wildlife Conservation. Springer, New York, pp. 151–163.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A.F., Miller, P.I., Louzao, M., 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156: 94–104. doi:10.1016/j.biocon.2011.11.013

Photopoulou, T., Lovell, P., Fedak, M.A., Thomas, L., Matthiopoulos, J., 2015. Efficient abstracting of dive profiles using a broken-stick model. Methods in Ecology and Evolution 6: 278–288. doi:10.1111/2041-210X.12328

R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rehfeldt G.E., Crookston N.L., Warwell M.V., Evans J.S. 2006. Empirical analyses of plant-climate relationships for the western United States. Int J Plant Sci 167: 1123–1150.

Renner, M., Parrish, J.K., Piatt, J.F., Kuletz, K.J., Edwards, A.E., Hunt, G.L., Jr, 2013. Modeled distribution and abundance of a pelagic seabird reveal trends in relation to fisheries. Marine Ecology Progress Series 484: 259–277. doi:10.3354/meps10347

Roberts, J.J., Best, B.D., Mannocci, L., Fujioka, E., Halpin, P.N., Palka, D.L., Garrison, L.P., Mullin, K.D., Cole, T.V.N., Khan, C.B., McLellan, W.A., Pabst, D.A., Lockhart, G.G., 2016. Habitat-based cetacean density models for the U.S. Atlantic and Gulf of Mexico. Sci. Rep. 1–12. doi:10.1038/srep22615

Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23: 2507–2517. doi:10.1093/bioinformatics/btm344

Scales, K.L., Miller, P.I., Ingram, S.N., Hazen, E.L., Bograd, S.J., Phillips, R.A. 2015. Identifying predictable foraging habitats for a wide-ranging marine predator using ensemble ecological niche models. Diversity and Distributions 1-13. doi:10.1111/ddi.12389

Strobl, C., Hothorn, T. & Zeileis, A. 2009. Party on!. R Journal 1: 14–17.

Strobl, C., Malley, J. & Tutz, G. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods 14: 323–348.

Thuiller, W. 2014. Editorial commentary on 'BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change'. Glob Change Biol. 20: 3591–3592. doi:10.1111/gcb.12728

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32: 369–373. doi:10.1111/j.1600-0587.2008.05742.x

Thums, M., Bradshaw, C. J. A. and Hindell, M. A. 2008. A validated approach for supervised dive classification in diving vertebrates. J. Exp. Mar. Bio. Ecol. 363: 75–83.

Appendix 6

Some initial thought on framework of guideline for species distribution models (SDMs) applied to cetaceans

HIROTO MURASE

National Research Institute of Far Seas Fisheries, Japan Fisheries Research and Education Agency (FRA), 2-12-4 Fukuura, Kanazawa, Yokohama, Kanagawa 236-8648, Japan

ABSTRACT

Some initial thought on framework of guideline for species distribution models (SDMs) applied to cetaceans is presented in this paper. Ten iterative steps in development and evaluation of models proposed by Jakeman *et al.* (2006) are used as a templated for this purpose. Written statement of these steps will help reviewers.

INTRODUCTION

Application of species distribution models (SDMs) to cetaceans started in the late 1990s and the number of studies has been increasing in recent years (Murase et al., 2015). Statistical models (including traditional regression models and machine learning) relating occurrence and/or abundance of species to its environment at a certain time period is termed SDMs. Development of a guideline is required for proper assessment of the outcomes of modelling because a lot of choices are available for scientists in the course of construction of SDMs such as statistical models, covariates, model selections and evaluations and these choices confound reviewers. This is especially true in the context of the Scientific Committee of the International Whaling Commission (IWC/SC), because the results are expected to be used as an element of management advices. The Sub-Committee on the Revised Management Procedure (RMP) of the IWC/SC is currently trying to develop a guideline for model-based abundance estimation methods, mainly focusing on a generalized additive model (GAM) which can be treated as a type of SDMs. The Working Group on Ecosystem Modelling (EM) of the IWC/SC also recognized the necessity for the development of a guideline on the techniques and underlying assumptions of SDMs based on up-to-date and comprehensive knowledge (IWC, 2015). This paper provides some of initial thought on framework of a guideline for SDMs.

FREMEWRK OF A GUIDELINE

Although there is a number of guidelines for modelling in the context of environmental/ecological studies, ten iterative steps in development and evaluation of models proposed by Jakeman *et al.* (2006) can be considered as a good starting point for development of a guideline for SDMs applied to cetaceans. Written statement of these steps will help reviewers. The following are general comments on each steps for the purpose of the review of SDMs.

First step: Definition of purposes for modelling

In border context, there are at least three purposes of development of SDMs applied to cetaceans: (1) estimation of spatial abundance (2) estimation of spatial distribution and (3) investigation on ecological questions (e.g. habitat requirement). These purposes might involve interpolation (estimation within a target [survey] area) and extrapolation (prediction outside of a target area and future projection). However, these two words can be defined differently: interpretation can be regarded as estimation within line transect strip (length of transect times effective search width) while extrapolation can be regarded as estimation within a target area. These three are not mutually exclusive and some of statistical models can address these at once. Nevertheless, distinction of main purposes for modelling is important because they affect details of subsequent model development and evaluation steps. Former two purposes are more related to in-depth assessment and management of stocks while the third purpose is more related to ecological questions.

Second step: Specification of the modelling context

According to the Jakeman *et al.* (2006), following 9 points should be considered at this step: (1) the specific questions and issues that the model is to address, (2) the interest group, including the clients or end-users of the model, (3) the outputs required, (4) the forcing variables (drivers), (5) the accuracy expected or hoped for, (6) temporal and spatial scope, scales and resolution, (7) the time frame to complete the model as fixed, (8) the effort and resources available for modelling and operating the model and (9) flexibility. Jakeman *et al.* (2006) considered that the crucial point at this step is (6).

Specific questions and issues that the model is to address
Following are some examples of specific questions and issues that the SDMs to address:

- Spatial abundance estimation for the purpose of RMP
- Investigation on reasons of change in spatial abundance and distribution for the purpose of IA
- Identification of distribution area of whales to reduce ship strikes in a certain area
- Investigation on habitat requirements for ecological study

Interest group, including the clients or end-users of the model
In general, interest groups of SDMs could consist of managers, scientists, fishermen, conservation groups and general public though specific combination would be varied from case to case. In the context of IWC/SC primary interest groups are managers and scientists.

Outputs required
Primary outputs from SDMs are estimated maps of probability of occurrence and/or abundance. Point estimate can also be obtained in the case of abundance. Importance (rank) of environmental variables affecting occurrence/abundance can also be obtained from models.

Forcing variables (drivers)
This is not applicable to SDMs as forcing variables are not used in the models in general.

Accuracy expected or hoped for
Acceptable level of accuracy should be determined by discussion before conducting modelling but it might be changes on the course of analysis.

Temporal and spatial scope, scales and resolution
Specification of temporal and spatial scope, scales and resolution is closely tied with the purpose of modelling. For example, if one aims to estimate spatial abundance and distribution of a stock in a particular season, broader spatial area should be covered by the modelling. For instance, in the case of the IWC SOWER CPIII, it took approximately 40 days by two vessels to cover an area in a 30° longitude sector from ice edge to 60°S. In such a case, temporal scale of environmental data (e.g. temperature) for modelling might be restricted to month or seasonal mean data. In contrast, if one aims to estimate spatial abundance and distribution in a local area (e.g. bay), it can be covered by a few days. In that case, environmental data with high temporal resolution might be used but only a fraction of a stock might be studied. Specification might be limited by available environmental data. For example, sea surface temperature derived from satellite data is commonly used as an environmental data in modelling. However, both temporal (e.g. observed period) and spatial (e.g. grid size and cloud cover) coverages are limited.

Time frame to complete the model as fixed, for example, by when it must be ready to help a decision
Time frame should be determined by discussion before conducting modelling but it might be changed on the course of analysis.

Effort and resources available for modelling and operating the model
Identification of effort and resources available for modelling is important to set time frame and required budget. Consideration of operation of the model might be necessary if constructed models are applied to new environmental data (e.g. temperature) continuously.

Flexibility; for example, can the model be quickly reconfigured to explore a new scenario proposed by a management group
Flexibility of models used as SDMs should be described although most of them are reasonably flexible for reconfiguration.

Third step: Conceptualisation of the system, specification of data and other prior knowledge
Reasonable hypothesis about relationship between explanatory variables (usually environmental variables) and response variable (presence/absence or abundance) should be provided. It is directly related to selection of explanatory variables for an initial model.

Response variable are one from the followings:

- Presence/absence: Typically collected by sighting survey (either dedicated or opportunistic) as sighting effort data are required.
- Presence only: Typically collected by satellite tags as the data provide only location of cetaceans.
- Abundance: Typically collected by dedicated sighting survey which records distance and angle of sightings, and school size to calculate effective half width and mean school size.

Type of response variable has strong influence on selection of model features and families.

A variety of explanatory variables has been used. The details of data should be provided. The followings are some of the examples:

- *In-situ* environmental data: Environmental data record during field surveys such as water temperature obtained by CTD and prey density obtained by echosounder are used in SDMs. Interpolation and/or extrapolation of data for target area are necessary as these data are recorded along track lines in most cases.
- Satellite data: Environmental data obtained satellite are commonly used in SDMs as the data have wide coverage both temporally and spatially. Types of data include such as SST, SSH, sea surface chlorophyll-a concentrations (chl-a) and sea ice concentrations. Interpolation and/or extrapolation of SST and chl-a data might be necessary in cases of missing values due to cloud cover. Secondary data products such as thermal fronts calculated using satellite data are also available for some regions.
- Terrain data: Digital bottom depth data and variables calculated using the data (e.g. slope) are used in SDMs. Distance from terrain futures such as coastline are also used.
- Ocean model data: Output from ocean model data (e.g. Regional Ocean Modeling System [ROMS]) are used in SDMs.
- Climatological data: Climatological data (e.g. World Ocean Atlas published by NOAA) are used in SDMs.

At this stage, considerations on spatial autocorrelation of response variable and collinearity among explanatory variables are also required especially for regression models.

Forth step: Selection of model features and families
Although a number of statistical models can be used as SDMs, selection of families (i.e. specific statistical models) is limited by of features (e.g. types of variables and linear/nonlinear functions). Description of reasons why a particular model is selected is inevitable. It is preferable to use several models and compare the results. An alternative choice could be ensemble modelling if the primary objective is estimation of spatial abundance and distribution. However, major drawback of ensemble modelling is that it cannot be utilized for ecological inferences.

Fifth step: Choice of how model structure and parameter values are to be found
Choice of model structure (i.e. relation between variables) can be inferred from prior scientific knowledge. However, the choice could be limited by availability of explanatory variables for SDMs. Methods to estimate of parameter values are specific to each statistical model.

Sixth step: Choice of estimation performance criteria and technique
Each statistical model has unique methods for parameter estimation performance criteria and technique, and it should be described fully.

Seventh step: Identification of model structure and parameters
In many cases, this step just consists of dropping or adding of particular parameters to reduce or increase model complexity based on fifth and sixth steps.

Eighth step: Conditional verification including diagnostic checking
There are generally two forms of verifications: quantitative and qualitative verifications. Qualitative (conceptual) verification is verification between real system and conceptual model based on qualitative information such as expert knowledge. Quantitative (model) verification is verification between conceptual and quantitative model based quantitative criteria such as goodness fit and test on residuals.

Ninth step: Quantification of uncertainty
Uncertainty associated with abundance is estimated such as bootstrapping in GAM. However, uncertainty associated with probability of occurrence has rarely exploited so far.

Tenth step: Model evaluation or testing (other models, algorithms, comparisons with alternatives)
Model evaluation using test data have been conducted for probability of occurrence based on AUC. However, it has rarely conducted for abundance. Comparison of results among different statistical models is recommended to evaluate them. Point estimate (e.g. abundance) comparison is relatively easy but ecological inference might be difficult if different models show different results (e.g. shape of response form).

REFERENCES

Jakeman, A.J., Letcher, R.A. and Norton, J.P. 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Modell. Softw. 21: 602-14.

Murase, H., Friedlaender, A., Kelly, N., Palacios, D.M. and Palka, D. 2015. A preliminary review of species distribution models (SDMs) applied to baleen whales. Paper SC/66a/EM3 presented to the 66a IWC Scientific Committee, May 2015. (unpublished). 18pp.