2.3. Distribution Modelling

Sophie Mormède[1], Jean Olivier Irisson[2, 3] & Ben Raymond[4, 5, 6]

[1] *NIWA, Wellington, New Zealand*
[2] *UPMC Université Paris 06, Observatoire océanologique, Villefranche/mer, France*
[3] *CNRS, Observatoire océanologique, Villefranche/mer, France*
[4] *Australian Antarctic Division, Department of the Environment, Kingston, Australia*
[5] *Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Hobart, Australia*
[6] *Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Australia*

1. Introduction

One of the aims of this Atlas is to characterise the spatial distribution of each species of interest. In many cases, this information is conveyed by a map of the locations where each species was recorded to be present. This gives a solid and conservative overview of their distribution, but can be misleading in two ways. Firstly, it may not be possible to indicate locations that were appropriately sampled but where the species was not found to be present. Second, the spatial distribution of sampling effort is typically uneven, and so heavily-sampled regions can appear very crowded while scarcely-visited or unsurveyed regions appear blank, when the reality might be a more homogeneous distribution. When sufficient data are available, modelling can help to provide a less biased estimate of the distribution, including inferences about the likely distributions in unsurveyed areas (Gutt *et al.* 2012). Two broad types of analysis were carried out: single-species distribution models, and multi-species distribution models (i.e. community assemblages).

Several modelling techniques were used, and are described below. Although the details differ between techniques, in general these approaches all use regression techniques to find relationships between the species observations and various physical and chemical environmental variables (see Table 1), thereby characterising the dependence of the biological patterns on environmental conditions. The validity of such a relationship depends on a number of factors and assumptions, including (a) that the same relationship between the biota and the environment holds across the entire spatial domain, and (b) that the environmental variables — at their available spatial and temporal resolution — adequately describe the environmental factors and processes that are relevant to the species of interest. Studies such as this on large spatial scales are reliant on remote-sensed or modelled environmental data, in order to obtain full coverage across the region of interest. In many cases this precludes the use of otherwise-useful predictor data (e.g. ship-based estimates of prey densities from acoustic measurements) in favour of synoptic data that might be less directly ecologically relevant to the target species (e.g. satellite-derived chlorophyll-*a* concentration as a proxy for primary productivity).

Once fitted, a relationship between environment and biology is typically used to make predictions of biological patterns across the entire spatial domain of interest, often including geographic areas that have not been surveyed. This process involves interpolation (i.e. making predictions for data that are within the range of the training data) and may also involve extrapolation (i.e. predictions for data that lie outside of the coverage of the training data). Interpolation is generally safer than extrapolation, because the latter relies on the model to behave sensibly for input data outside of the range used to estimate the model parameters. It is important to recognise that interpolation and extrapolation can be viewed in either geographic or environmental space. Consider a point that geographically lies outside of the spatial bounds of the training data, but which has environmental conditions within the range encompassed by the training data. Prediction at this point therefore involves interpolation in environmental space, but extrapolation in geographic space. This can generally be expected to be a reasonable action, provided that the assumptions (a and b, above) are met. Extrapolation in environmental space may require more care in order to ensure that the predictions are reasonable, because it requires not only numerically sensible behaviour from the model beyond the training data range, but also requires that the ecological implications and assumptions of the model are reasonable for the new regions of the environmental domain.

2. Single-species distribution models

Two methods were used for single-species distribution models: boosted regression trees (BRT) and MaxEnt. The BRT approach requires both presence and absence locations for a species, whereas MaxEnt requires only presence locations. Absence is rarely explicitly recorded in the databases from which the Atlas data were aggregated. Where possible, absences were inferred from related presence data. For example, absences of a specific plankton species were inferred from records of similar plankton species, assuming that if the species of interest had been caught in those locations, it would have been recorded, and that the method used was suitable to catch that species. Where absences were not available and could not be inferred, the MaxEnt method was used.

For both techniques, the preferred habitat of a single species was estimated by regressing the species observations against the environmental variables. Modelling was only conducted for species with sufficient data, and where the fitted relationships between species presence and environmental variables were biologically sensible. For example, few benthic species were successfully modelled, because they typically respond to highly local environmental factors, proxies for which were not available at the appropriate resolution (or with full circumpolar coverage).

2.1. Boosted regression trees

The BRT modelling technique has been shown to be useful for distribution modelling of Southern Ocean pelagic zooplankton species (Pinkerton *et al.* 2010). It is an ensemble method, meaning that predictions from a large number of relatively simple models (in this case, binary regression trees) are combined to give a single overall prediction (Hastie *et al.* 2001, Friedman & Meulman 2003). The manner in which this combining is performed allows the BRT to automatically fit complex, non-linear relationships and interactions between variables. Only two-way interactions (i.e. interactions between pairs of variables) were considered in each simple tree, but by combining many of those trees, the overall model accounts for more complex interactions. All models were presence-absence (no abundance information was used) and so used a binomial distribution with logit link. Critical parameters for these models are the learning rate (how much each simple tree contributes to the final model) and the number of simple trees in the ensemble. For maximum predictive power, the learning rate should very low and the number of trees, therefore, very high (Elith *et al.* 2008); but this has a cost in terms of computation speed. A suitable choice of learning rate and number of trees was assessed through 10-fold cross validation. The dataset was divided into ten subsets, the model was built on nine of them and used to make predictions for the remaining one. By comparing those model predictions with the actual data, an estimate of the residual deviance (i.e. the model error) was obtained. This procedure was repeated ten times, withholding a different subset of data each time. The number of trees and learning rate was chosen to minimise the residual deviance. We considered models with at least 1000 trees and a learning rate smaller than 5×10^{-4}.

The area under the receiver operating characteristic curve (AUC) was calculated (Swets 1988) to assess the performance of the models. The AUC value is the area under a plot of the fraction of true positives (i.e. the chance of correctly identifying presence) against the fraction of false positives (i.e. the chance of predicting presence when the species is actually absent). An ideal model will predict 100% true positives and 0% false positives, giving an AUC value of 1. A model with no discriminatory power will generate equal fractions of true and false positives, giving an AUC value of 0.5. The AUC value can also be interpreted as the probability that the model will predict a higher probability of presence for a randomly chosen presence sample than for a randomly chosen absence sample. A model with an AUC value greater than 0.7 is considered “useful” (Swets 1988), although of course this depends on the particular application.

As well as assessing the overall model fit, the relative influence of each predictor variable can be calculated as the weighted average of the number of times that the variable appears in the trees that make up the model. These values are expressed as percentages and sum to 100% regardless of the absolute value of the variance explained: that is, they only quantify a *relative* influence. The shape of the effect of each variable on the presence of the species can be visualised through partial dependence plots. A partial dependence plot shows the marginal effect of the variable of interest on the logit of the probability of presence (“marginal” in this context meaning that the effects of the other predictor variables are integrated out). When the value on the y-axis is high, the conditions on the x-axis are favourable.

In some instances, the data were biased towards a few specific locations, and hence had a high number of records with very similar values for the environmental variables. This could happen when specific areas were intensively sampled by a specific scientific program, for example, or when a single station was recorded multiple times in the database with no formal way to identify that it was a single station. In these instances, the data was binned to the same resolution as the environmental data used in the model (0.1° longitude by 0.1° latitude), and each record was weighted by the inverse of the number of records in that bin.

Predictions were not made outside of the environmental coverage of the training data (i.e. no “environmental extrapolation”). These areas appear in grey in the maps.

The environmental variables chosen in each model were determined by potential biological relevance for the specific organism modelled, with regards to potential correlation between available variables. Initial model runs determined which variables were of importance, and those with little or no importance were removed from the model. 200 bootstraps were carried out on both the model and the predictions to determine the confidence interval around the environmental effects and predictions.

Table 1 List of abiotic layers selected for establishing the prediction maps.

Parameter	Source	Description and processing notes
Depth	Smith & Sandwell (1997) http://topex.ucsd.edu/WWW_html/mar_topo.html-Source data version: V13.1 (Sep 4, 2010)	Data from satellite altimetry and ship depth soundings, subsampled from original 1-minute to 0.05-degree resolution and interpolated to 0.1-degree grid using bilinear interpolation.
Slope	Derived from Smith & Sandwell V13.1 bathymetry data (above).	Bathymetric slope calculated on 0.1-degree gridded depth data (above), using the equation by Burrough & McDonell (1998, p. 190). See http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=How%20Slope%20works.
Geomorphology	Mapping based on GEBCO contours, ETOPO2, seismic lines	Mapped from bathymetric analysis, with features cross-checked from seismic lines. And classified at a scale of 1: 1–2 million
Distance to shelf break	Derived from geomorphic features map.	Distance calculated from the coastline to the upper slope as defined in Table 1 of the geomorphic features.
Chlorophyll-a summer	Feldman & McClain (2010)	Near-surface chl-a summer mean from MODIS Aqua. Data span the 2002/03 to 2009/10 austral summer seasons. Data interpolated from original 9km resolution to 0.1-degree grid using bilinear interpolation.
Sea ice	Derived from AMSR-E satellite estimates of daily sea ice concentration at 6.25km resolution (Spreen *et al.* 2008) http://iup.physik.uni-bremen.de:8084/amsredata/asi_daygrid_swath/l1a/s6250/	Concentration data from 1-Jan-2003 to 31-Dec-2009 used. The fraction of time each pixel was covered by sea ice of at least 85% concentration was calculated for each pixel in the original (polar stereographic) grid. Data then regridded to 0.1-degree grid using triangle-based linear interpolation.
Southern Ocean fronts	Sokolov & Rintoul (2009)	Data provided as mean positions (line features) from satellite altimetry. Distance to the polar front also calculated, using the minimum distance from each pixel in the 0.1–degree grid to the middle branch of the polar front.
Distance to nearest seabird breeding colony	Calculated from the Inventory of Antarctic seabird breeding sites, collated by Eric Woehler http://data.aad.gov.au/aadc/biodiversity/display_collection.cfm?collection_id=61	
Salinity (winter) 0/50/200/500m	World Ocean Atlas 2009 (Antonov *et al.* 2010)	Data regridded to 0.1-degree grid using bilinear interpolation
Salinity (summer) 0/50/200/500m	See salinity (winter)	
NOx (winter) 0/50/200/500 m	World Ocean Atlas 2009 (Garcia *et al.* 2010b)	See salinity (winter)
NOx (summer) 0/50/200/500m	See NOx (winter)	
Oxygen (winter) 0/50/200/500m	World Ocean Atlas 2009 (Garcia *et al.* 2010a)	See salinity (winter)
Oxygen (summer) 0/50/200/500m	See oxygen (winter)	
Temperature (winter) 0/50/200/500m	World Ocean Atlas 2009 (Locarnini *et al.* 2010)	See salinity (winter)
Temperature (summer) 0/50/200/500m	See temperature (winter)	
Sea surface temperature (SST) summer	Feldman & McClain (2010)	Data from MODIS Aqua. Climatology spans the 2002/03 to 2009/10 austral summer seasons. Data interpolated from original 9km resolution to 0.1-degree grid using bilinear interpolation.
Seafloor temperature	Clarke *et al.* (2009)	Original data derived from World Ocean Atlas 2005 data and provided on a 1-degree grid. Isolated missing pixels (i.e. single pixels of missing data with no surrounding missing pixels) were filled using bilinear interpolation, and then data were regridded from 0.1-degree grid using nearest neighbour interpolation.
Last glacial ice sheet maximum grounding line	Modified from Anderson *et al.* (2002)	The location of the LGM grounding line was based on the work of Anderson *et al*. (2002), but modified to account for the position of the shelf break as identified on the geomorphic map
Granulometry	McCoy (1991)	Derived from sediment types.
Biogenic component in sediment	See Granulometry	Siliceous vs calcareous.

2.2. MaxEnt

Where absences were not available, the MaxEnt method (Phillips *et al.* 2006, Elith *et al.* 2011) was used for single-species distribution modelling. MaxEnt is probably the best-known method for presence-only species distribution modelling, although we note that this is an active area of research (see e.g. Warton & Shepherd 2010, Royle *et al.* 2012, Hastie & Fithian 2013). MaxEnt differs from a presence-absence model in that it utilises the presence locations along with "background" data from across the landscape or region of interest. The MaxEnt approach compares the environmental characteristics of the presence locations with those of the background samples (i.e. the environmental conditions which are potentially available to the species of interest). In doing so, it identifies the environmental characteristics that are being "selected" by the species in question, relative to those that are available. MaxEnt does not give estimates of true probability of presence, but rather relative probability of presence, or "habitat suitability".

MaxEnt assumes by default that the presence samples have been drawn randomly from across the species range. Sample selection bias (i.e. where the presences have not been sampled uniformly, perhaps because some areas have been more intensively surveyed than others) can be problematic for presence-only models (Phillips *et al.* 2009). While in principle it is possible to minimise the effects of this bias (e.g. by introducing a matching bias to the background samples; Phillips *et al.* 2009), we did not do so here. The sampling bias in high-latitude Southern Ocean species observations is not merely spatial, but also has temporal, seasonal, and methodological components. Adequately estimating and accounting for these biases is an ongoing area of research.

3. Species assemblages

Species assemblages were determined by clustering analyses, either applied directly to the observations, using generalised dissimilarity modelling (GDM), or applying clustering analyses to the results from single-species predictions. Assemblage analyses were carried out for pertinent groupings of species, e.g. all euphausiids together.

3.1. Generalised dissimilarity modelling

GDM is a technique that models variation in species turnover between sites as a function of environment (Ferrier *et al.* 2007). This uses a modified form of matrix regression to model the relationship between biological and environmental distances between pair-wise combinations of sites. Predictions can be made for any pair of sites (i.e. not confined to those used in fitting the model), provided that values for all environmental predictors are specified for both sites. Thus, as with most modelling methods, it is possible to fit the model to the available observations and then use that fitted model to make predictions across the full region of interest (circumpolar Southern Ocean, in this case).

Given a pair of sites, a fitted GDM model will provide a prediction of the biological dissimilarity between those two sites. The predicted dissimilarities between all pairs of sites in a region can then be used as an input for more conventional dissimilarity-based community ecology analysis methods such as clustering or multidimensional scaling. Here, our region of interest is sufficiently large that we cannot work directly with the all-pairs dissimilarity matrix due to computational considerations (the matrix is too large to fit in memory on most systems). Instead, we transform the environmental data in each point of the grid with the same nonlinear transformation used internally by the fitted GDM model (function *gdm.transform* in the code distributed by Ferrier *et al.*). The Manhattan distance d_{ij} between this transformed data for a pair of sites *i* and *j* is monotonically related to the predicted dissimilarity D_{ij} usually computed by GDM (function *gdm.predict*):$D_{ij} = 1 — e^{-d_{ij}}$

Thus, we can obtain similar results by clustering the transformed data (using a clustering algorithm with the Manhattan distance) as would be obtained by clustering on the basis of the predicted dissimilarities from the GDM model. If a hierarchical UPGMA clustering algorithm was used on the transformed data, the two sets of results would be identical. However, with this transformed data we used the non-hierarchical clustering algorithm *clara* (Kaufman & Rousseeuw 2008), which does not require all pairwise dissimilarities and is therefore computationally feasible. Using this method is not guaranteed to produce identical results to hierarchical clustering of the

predicted dissimilarities from the fitted GDM; however, in practice we have found the results to be sufficiently similar for applied use.

The optimum number of clusters (between 3 and 10) was computed automatically by maximising average bandwidth. The data were binned at the same 0.1-degree resolution as the environmental data (see Table 1). Any species recorded within each bin at least once was assumed present in that bin.

3.2. Clustering of single-species distributions

The computational demands of GDM limit its use to relatively small data sets (usually 2500 sites or less). For species groups with large amounts of data, and where most important individual species were modelled using BRT (e.g. euphausiids), the results from BRT were classified using a two step clustering method: a non-hierarchical cluster using *clara*, as above, to 200 clusters, followed by a hierarchical clustering to 12 groups. This method is inspired from the gradient forest methodology (Leaper *et al.* 2011, Ellis *et al.* 2012), but controls and optimises the prediction of each species individually, and applies no weight between the different species before clustering. Direct hierarchical clustering was not suitable because of its high computational demands.

Acknowledgments

Part of this project was funded by the New Zealand MBIE project C01X1001 ("Protecting Ross Sea Ecosystems"), and NIWA travel funds. This is CAML contribution # 92.

References

Anderson, J.B., Shipp, S.S., Lowe, A.L., Wellner, J.S., Mosola, A.B., 2002. The Antarctic ice sheet during the last glacial maximum and its subsequent retreat history: a review. *Quaternary Science Reviews*, **21**, 49–70.

Antonov, J.I., Seidov, D., Boyer, T.P., Locarnini, R.A., Mishonov, A.V., Garcia, H.E., 2010. *World Ocean Atlas 2009, Volume 2: Salinity*. Washington, DC: U.S. Government Printing Office, 184 pp.

Burrough, P.A., McDonell, R.A., 1998. *Principles of Geographical Information Systems*. New York: Oxford University Press, 333 pp.

Clarke, A., Griffiths, H.J., Barnes, D.K.A., Meredith, M.P., Grant, S.M., 2009. Spatial variation in seabed temperatures in the Southern Ocean: implications for benthic ecology and biogeography. *Journal of Geophysical Research*, **114**, G03003. doi:10.1029/2008JG000886.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813. doi:10.1111/j.1365-2656.2008.01390.x.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57. doi:10.1111/j.1472-4642.2010.00725.x.

Ellis, N., Smith, S.J., Pitcher, C.R., 2012. Gradient forests: calculating importance gradients on physical predictors. *Ecology*, **93**, 156–168. doi:10.1890/11-0252.1.

Feldman, G.C., McClain, C.R., 2010. *Ocean Color Web, MODIS Aqua Reprocessing, NASA Goddard Space Flight Center*. Eds. Kuring, N., Bailey, S.W. http://oceancolor.gsfc.nasa.gov/.

Ferrier, S., Manion, G., Elith, J., Richardson, K., 2007. Using generalised dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264. doi:10.1111/j.1472-4642.2007.00341.x.

Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, **22**, 1365–1381. doi:10.1002/sim.1501.

Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., 2010a. *World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation*. Washington, DC: U.S. Government Printing Office, 344 pp.

Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Zweng, M.M., Baranova, O.K., Johnson, D.R., 2010b. *World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate)*. Washington DC: U.S. Government Printing Office, 398 pp.

Gutt, J., Zurell, D., Bracegridle, T.J., Cheung, W., Clark, M.S., Convey, P., Danis, B., David, B., Broyer, C.D., Prisco, G.d., Griffiths, H., Laffont, R., Peck, L.S., Pierrat, B., Riddle, M.J., Saucède, T., Turner, J., Verde, C., Wang, Z., Grimm, V., 2012. Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. *Polar Research*, **31**, 11091. doi:10.3402/polar.v31i0.11091.

Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 1–4. doi:10.1111/j.1600-0587.2013.00321.x.

Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. New York: Springer, 745 pp. doi:10.1007/978-0-387-84858-7.

Kaufman, L., Rousseeuw, P.J., 2008. *Finding Groups in Data: An Introduction to Cluster Analysis*: John Wiley & Sons, Inc., 342 pp. doi:10.1002/9780470316801.

Leaper, R., Hill, N.A., Edgar, G.J., Ellis, N., Lawrence, E., Pitcher, C.R., Barrett, N.S., Thomson, R., 2011. Predictions of beta diversity for reef macroalgae across southeastern Australia. *Ecosphere*, **2**, art73. doi:10.1890/ES11-00089.1.

Locarnini, R.A., Mishonov, A.V., Antonov, J.I., Boyer, T.P., Garcia, H.E., 2010. *World Ocean Atlas 2009, Volume 1: Temperature*. Washington, DC: U.S. Government Printing Office, 184 pp.

McCoy, F.W., 1991. Southern Ocean sediments: circum-Antarctic to 30°S. In: Hayes, D.E. (ed.). *Marine Geological and Geophysical Atlas of the Circum-Antarctic to 30°S. Antarctic Research Series*, Volume **54**. Washington, DC: AGU, 37–46. doi:10.1029/AR054p0037.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259. doi:10.1016/j.ecolmodel.2005.03.026.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197. doi:10.1890/07-2153.1.

Pinkerton, M.H., Smith, A.N.H., Raymond, B., Hosie, G.W., Sharp, B., Leathwick, J.R., Bradford-Grieve, J.M., 2010. Spatial and seasonal distribution of adult *Oithona similis* in the Southern Ocean: predictions using boosted regression trees *Deep-Sea Research Part I: Oceanographic Research Papers*, **57**, 469–485. doi:10.1016/j.dsr.2009.12.010.

Royle, J.A., Chandler, R.B., Yackulic, C., Nichols, J.D., 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554. doi:10.1111/j.2041-210X.2011.00182.x.

Smith, W.H.F., Sandwell, D.T., 1997. Global seafloor topography from satellite altimetry and ship depth soundings. *Science*, **277**, 1957–1962. doi:10.1126/science.277.5334.1956.

Sokolov, S., Rintoul, S.R., 2009. Circumpolar structure and distribution of the Antarctic Circumpolar Current fronts: 1. Mean circumpolar paths. *Journal of Geophysical Research*, **114**, C11018. doi:10.1029/2008JC005108.

Spreen, G., Kaleschke, L., Heygster, G., 2008. Sea ice remote sensing using AMSR-E 89 GHz channels. *Journal of Geophysical Research*, **113**, C02S03. doi:10.1029/2005JC003384.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293. doi:10.1126/science.3287615.

Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, **4**, 1383–1402. doi:10.1214/10-AOAS331.