Where's the sperm whale? A species distribution example analysis

Carsten F. Dormann[1] & Kristin Kaschner[2]

[1]Helmholtz Centre for Environmental Research-UFZ, Department of Computational Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany

[2]Evolutionary Biology & Ecology Lab, Institute of Biology I (Zoology), Albert-Ludwigs-University Freiburg, Germany

Abstract We analyse the coarse-scale distribution of sperm whale around Antarctica as an example study of a typical species distribution model. Following the outline and structure in chapter XX, we demonstrate each point with this data set, show results and their interpretation, compare two modelling techniques (GLM and Boosted Regression Trees), and discuss the steps of the analyses in the light of the ecology behind the target species. Data and R-code are provided in order for readers and teachers to be able to reproduce our analysis.

Introduction

In the following pages, we use some data collected over the course of almost 25 years during cetacean IDCR-DESS SOWER surveys conducted around Antarctica. Using only sperm whale sightings and some basic effort information, we use this data to run through a typical species distribution analysis. Every data set offers its own challenges, and while we can provide an example, we cannot provide a recipe that can be transferred one to one.

The structure is simple. We follow the general outline given in chapter XX (Modelling Species' Distributions), and present the code to run the analyses. All code and data are available from the books accompanying webpage.

The controversy about whale conservation vs. commercial whale hunting is largely founded in our uncertainty of the true population sizes of almost all large whale species. In fact, even our knowledge of the most relevant habitat preferences of these highly migratory animals with often circum-global ranges is restricted to relative small geographic areas (Hamazaki 2002; Laran and Gannier 2008; Torres et al. 2008), obviously due to the extreme difficulty of covering vast amounts of ocean surface and the relatively low detection probability of creatures living in three dimensions, but being only monitored in two. Most existing large-

scale data sets about cetacean occurrence are either limited to presence-only records such as those available through online data repositories (e.g. OBIS, www.iobis.org). These data are compiled from a variety of different sources, greatly varying in quality with respect to survey design and/or reliable species identification. In addition, they rarely meet the assumption of a representative sampling coverage in terms of available habitat. As a consequence predictions of large-scale species distribution based on these data are limited to very simple environmental envelope or niche models {Kaschner, 2006 #8194; Ready, 2010 #8195, www.aquamaps.org}.

In contrast, dedicated cetacean surveys providing presence-absence information tend to be limited to relatively small scales and short time periods, thus only providing small snapshots of a species occurrence and habitat usage in time and space. Some of the most comprehensive surveys are the IDCR-DESS SOWER circumpolar cruises, organized and funded and organized by the International Whaling Commission (IWC) and its member state and regularly conducted in Antarctic waters since the 1970s (Branch and Butterworth 2001; IWC 2001; Kasamatsu et al. 2000b). The primary focus of these surveys is the assessment of baleen whale abundance (Branch 2007; Branch and Butterworth 2001; Butterworth and DeDecker 1989; Corkeron et al. 1999; Goodall 1997; Kasamatsu et al. 2000a; Kasamatsu et al. 1988; Kato et al. 1995; Matsuoka et al. 2003), although some attempts have been made to use these data (in combination with others) to investigate habitat usage (Kasamatsu et al. 2000a) and to model cetacean distributions (Hedley et al. 2001), but the emphasis has been on mostly on mysticete species. Here, we will therefore focus on the sperm whale, the largest of the odontocete species, known to feed almost exclusively on deep-sea giant squid. As a consequence most regional studies investigating habitat usage of sperm whale elsewhere in the world have therefore identified slope as one of the primary predictors of species occurrence (Davis et al. 2002; Hamazaki 2002; Praca and Gannier 2008).

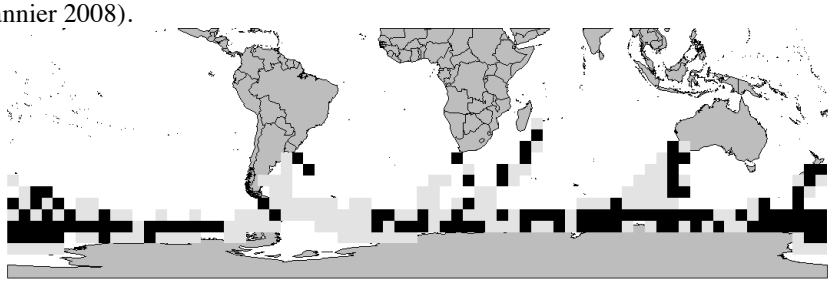


Fig. 1. Map of the southern hemisphere and the 5° grid cells visited during at least one of the three circumpolar IWC IDCR-DESS SOWER cruises conducted between 1978-2001 . Black squares indicate cells with at least one reported sperm whale sighting, grey squares no sightings.

An example analysis

If you are unfamiliar with R, please refer to one of the many documents and books available[1]

Loading, exploring and transforming the data

Response variable

Our response is binary, requiring no further considerations. If your data are heavily distorted, you may have to seek refuge in transformation. Note that the preferred option is always to model the data you have. Poisson and negative binomial should be able to take care of many skewed count data. Gamma distribution is an option for various strange distributions (Bolker 2008). If these "standard" options fail, consult a standard textbook about traditional transformations towards normality (Crawley 2002; Draper and Smith 1998; Neter et al. 1993; Quinn and Keough 2002; Sokal and Rohlf 1995; Underwood 1997; Zar 1996) or Bolker (2008) for an introduction to mixed distributions.

Explanatory environmental variables

Look at distribution of these variables and **transform** to maximise uniformity[2]:

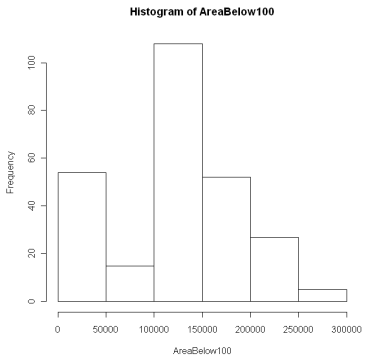
```
names(SW)
```

The first variable of interest is "AreaBelow100" (the total area below 100m of depth in each of the 5° cells). The others before will not be used as such.

```
hist(AreaBelow100); summary(AreaBelow100) # yielding the figure below
```

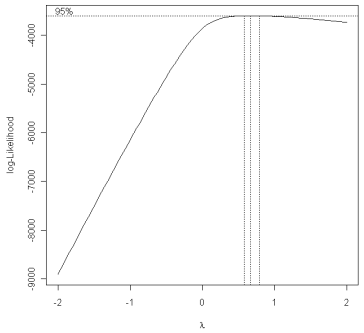
[1] Unter http://www.r-project.org/ check items "Books" and "Other" under the heading "Dokumentation". We particularly recommend "R for Beginners" by Emmanuel Paradis (http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf). For advanced R-users, check the "R Inferno" by Patrick Burns (http://www.burns-stat.com/pages/Tutor/R_inferno.pdf).

[2] We know already that we want to use BRT and GLM. Hence we can transform the variables as a first step. Using BRT alone would not require this step.

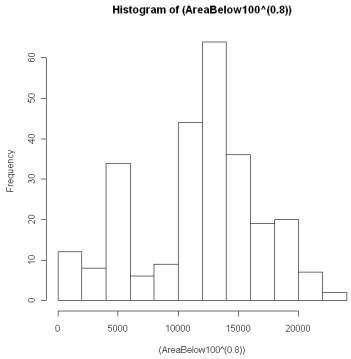


See what a Box-Cox-transformation would propose:

```
boxcox(lm(AreaBelow100 ~1)) #approx. 0.8
```




```
hist((AreaBelow100^(-1))) # does not look that much better:
```



We conclude that we stick to the untransformed data.

We now go through all variables in a similar way. This leads us to a diverse assortment of transformations (log, square-root and fourth-root), as well as several untransformed variables. Finally, we **standardize** all variables[3].

Collinearity

As a first step, we calculate a correlation matrix (round to 3 decimal places).

```
round(cor(SW.t[,-c(1:4)],use="complete.obs", method="kendall"), 3)
```

We can use a cluster representation to visualise collinearity.

```
require(Hmisc)
v <- as.formula(paste("~", names(SW.t)[-c(1,2)], collapse="+"))
plot(varclus(v, data=SW.t)) # yielding the figure below
```

---

[3] Here is the complete list of variables and transformations, stored in a new data set SW.t0.

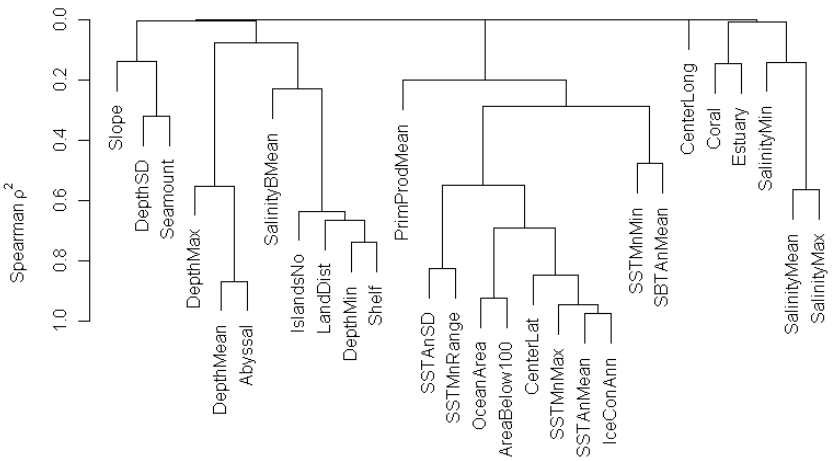
```
SW.t0 <- SW
SW.t0$SSTAnMean <- log(SW$SSTAnMean + 1.8) # mean annual sea surface temperature
SW.t0$SalinityMean <- log(SW$SalinityMean) # mean annual surface salinity
SW.t0$PrimProdMean <- log(SW$PrimProdMean)  # primary productivity (in mgC·m-²·day-1)
SW.t0$LandDist <- sqrt(SW$LandDist)    # distance to coast
SW.t0$Shelf <- log(SW$Shelf + 1)        # area < 200m depth
SW.t0$Slope <- sqrt(SW$Slope)           # area >200 – 4000m depth
SW.t0$Abyssal <- SW$Abyssal^0.25        # area > 4000m depth
SW.t0$Seamount <- log(SW$Seamount + 1) # number of seamounts
```

Other variables remain untransformed:

DepthSD	standard deviation of cell depth
DepthMean	mean depth of cell
SalinityBMean	mean annual salinity at sea bottom
DepthMin	minimum sea depth of a cell
SSTMnMin	mean annual minimum sea surface temperature
Coral	proportion of cell that is coral
Estuary	area of estuaries in a cell
SalinityMin	minimum monthly salinity

To standardise all explanatory variables to a mean=0 and sd=1 is simple:

```
SW.t <- SW.t0
SW.t[,-c(1:4)] <- scale(SW.t0[,-c(1:4)])
```



As we can see, several variables are unacceptably highly correlated (Spearman's rho^2 > 0.5). From those, we use the one with higher ecological relevance/interpretability (or, if this is not possible, highest importance in the randomForest model):

- DepthMean > DepthMax, Abyssal
- DepthMin > Shelf, IslandsNo, LandDist,
- SSTAnMean > IceConAnn, SSTMnMax, SSTAnSD, SSTMnRange, OceanArea, AreaBelow100
- SSTMnMin > SBTAnMean
- SalinityMean > SalinityMax

This leads to a reduced data set.

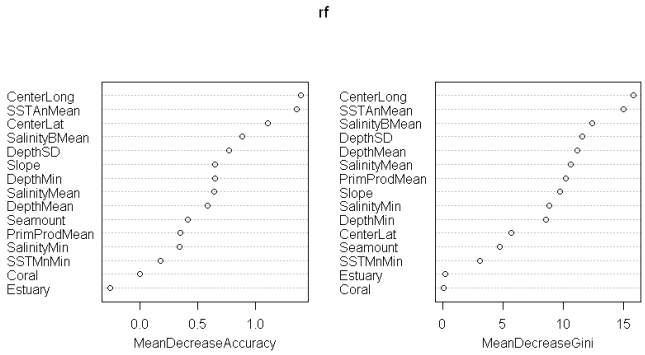
```
SW.t.red <- SW.t[ , c("PAsperm", "CsquareCode", "CenterLat", "Center-
Long", "Slope", "DepthSD", "Seamount", "DepthMean", "SalinityBMean",
"DepthMin", "PrimProdMean", "SSTAnMean", "SSTMnMin", "Coral", "Estu-
ary", "SalinityMin", "SalinityMean")]
```

Dimensional reduction

Our data set still comprises 13 predictors (not counting geographical location) for "only" 261 data points. In a GLM, using all 13 predictors, their first-order interactions and non-linear effects would lead to roughly 100 effects in the full model. Because we doubt that all variables are ecologically meaningful (e.g. proportion coral reefs or estuaries), we would have eliminated those before even entering the analysis. For the sake of demonstration, however, we pretend that all 13 variable could be reasonably correlated with sperm whale occurrence. To nevertheless re-

duce the number of predictors, we can use a univariate pre-scan, for example a GLM or GAM. Since some predictors may interact, this univariate pre-scan may not be unbiased. Instead, we use a machine-learning algorithm, called randomForest, to rank variables by importance (Breiman 2001; Hastie et al. 2008).

```
require(randomForest)
f <- as.formula(paste("as.factor(PAsperm)~", paste(names(SW.t.red)[-
c(1,2)], collapse="+"), sep=""))
rf <- randomForest(f, data=SW.t.red, na.action=na.omit, importance=T)
varImpPlot(rf)  # yielding the figure below
```

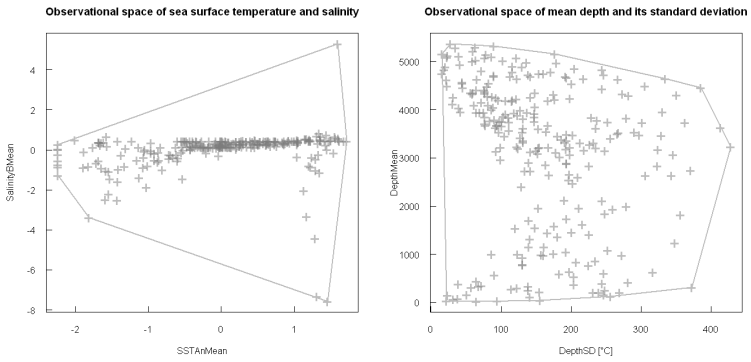


There are different ways to quantify importance of variables in randomForest[4]. Hastie et al. (2008, page 593) recommend using the Gini coefficient (right panel) for evaluating importance. Here, "SSTAnMean" "SalinityBMean" "DepthSD" "SalinityMean" "DepthMean" come out as the top five. It is somewhat arbitrary to set a threshold here, apart from the apparent "knee" after the first two and before the last two predictors.

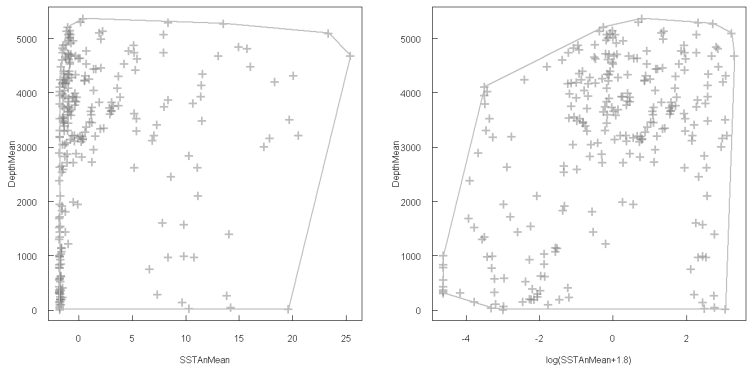
The final preparational step is to **visualize the parameter space**. Let us look at the two most important variables, SSTAnMean and SalinityBMean (untransformed). The two are uncorrelated (Kendall's $\tau = 0.377$), but clearly there are only few regions in the parameter space that is actually covered, even inside the convex hull!

```
pair <- SW.t.red[,c("SSTAnMean", "SalinityBMean")]
plot(pair, pch="+", cex=2, col=rgb(.5,.5,.5,.5), las=1, tcl=0.5,
main="Observational space of sea surface temperature and salinity")
polygon(pair[chull(pair),], lwd=2, border="grey")
```

[4] By default, mean decrease in accuracy and in Gini coefficient are returned. See help page for details: `?randomForest`



The majority of data points are from near Antarctica, hence sub-zero temperature. The trips to South Africa, South America and Australia yield the high-temperature values. Remember that we actually log-transformed SSTAnMean to spread the low values and shrink the high:



The effect is remarkable and should convince anyone skeptical of transformations of the explanatory variable. The coverage of the parameter space is much better now.
We continued for other combinations, but omit this here.

Modelling

Our statistical analysis comes in two flavors, Boosted Regression Trees and GLM.

Boosted Regression Trees

First, we include helper R-code to facilitate our analysis (from Elith et al. 2008).

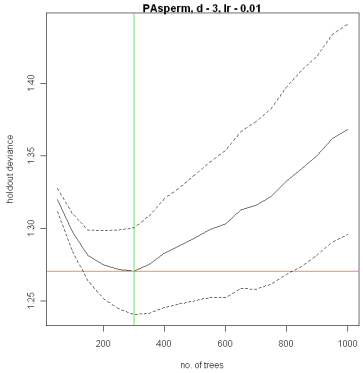
```
source("brtfunctions.r")
```

BRT offers various specifications; crucial ones are

- tree.complexity (don't use high values, e.g. > 5)
- learning.rate (the lower the slower; good values are 0.01 to 0.001)
- bag fraction (the proportion of data points used for fitting)

For defaults type: `fix(gbm.step)` and read the paper. Running the model itself is straight forward:

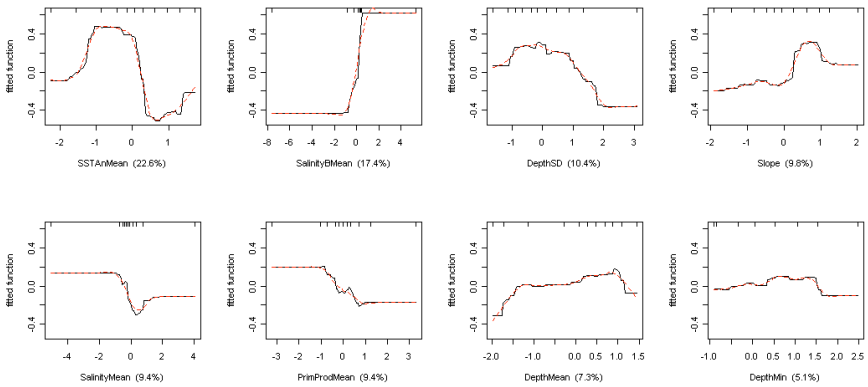
```
f.brt <- gbm.step(data=SW.t.red, gbm.x = 5:17, gbm.y = 1, family =
"bernoulli",  tree.complexity = 3, learning.rate = 0.01, bag.fraction
= 0.5, verbose=TRUE, silent=FALSE, plot.main=TRUE)
```



The graph returned depicts the development of residual variance in the validation data sets: the lower, the better. At some point there are too many trees, and the BRT overfits the data (the lines go up again). This point (indicated by a vertical green line) is the BRT-set used.

We can ask for an importance table similar to that for randomForest (here as percentage of all variance that is explained by a specific variable), and plot the partial plots, i.e. the functional relationships for each variable, averaged across the values of all other variables.

```
summary(f.brt)
gbm.plot(f.brt, smooth=T, n.plots=8, write.title = F)
```



GLM

For the GLM, we have to restrict the number of variables, based on a rule of thumb and the effective sample size (Harrell 2001).

```
table(SW.t.red$PAsperm)
# 0   1
#146 115
```

Since we have an effective sample size of 115, and we want to have at least 10 data points support per variable (events per variable, EPV=10), we allow only 10 effects into the GLM. We know (from the partial BRT plots) that variable effects are non-linear, hence we also include quadratic terms, as well as interactions. We start with a more complex, full model and reduce it to the desired complexity[5]:

```
source("COLL_allfunctions.r")
f <- formula.maker(SW.t.red[,c("PAsperm", "SSTAnMean", "Salin-
ityBMean", "DepthSD", "SalinityMean", "PrimProdMean")])
fm <- glm(f, data=SW.t.red, family=binomial)
anova(fm, test="Chisq")  # just a quick look at the first, overfitted
model
```

Since the function for SSTAnMean looked very non-linear in the BRT plots, lets add a third-order polynomial:

```
fm <- update(fm, .~. + I(SSTAnMean^3))
```

Model simplification

[5] The `formula.maker` function is a little convenience function for automatically formulating quadratic and/or interaction terms.

Obviously, this full model is unacceptably complex. We use an automatised stepwise backward selection procedure to simplify the model. Our criterion is, in contrast to what is implied by the name of the function, not the AIC, but the BIC[6].

```
fm.red <- stepAIC(fm, k=log(115))
anova(fm.red, test="Chisq")
```

From the first parameter-space plot above we know that the interaction of SSTAnMean and SalinityBMean cannot be supported by data. We thus manually delete this interaction:

```
fm.red2 <- update(fm.red, .~. - SSTAnMean:SalinityBMean)
anova(fm.red2, test="Chisq")
```

We cannot remove main effects when they are part of a significant interaction (marginality theorem). A quadratic effect without the main effect is "allowed", but strange. Let's put the main effect in, too:

```
fm.red3 <- update(fm.red2, .~.-I(DepthSD^2)+poly(DepthSD,2)) # not
significant
anova(fm.red3, test="Chisq")
```

We can now compare the two models: Was the quadratic effect worth its inclusion?

```
anova(fm.red3, fm.red2, test="Chisq")
```

No, models 3 and 2 are indistinguishable, so we use the simpler (3). Let's do the same for SSTAnMean:

```
fm.red4 <- update(fm.red2, .~.-I(SSTAnMean^3)-
SSTAnMean+poly(SSTAnMean,3))
anova(fm.red4, test="Chisq")
```

Was it worth it?

```
anova(fm.red4, fm.red2, test="Chisq")
```

No evidence! Still, for reasons of elegance and beauty, let's stick to the “proper” polynomial effect of SSTAnMean! (This is to show that there is always a level of arbitration in statistical modelling!)

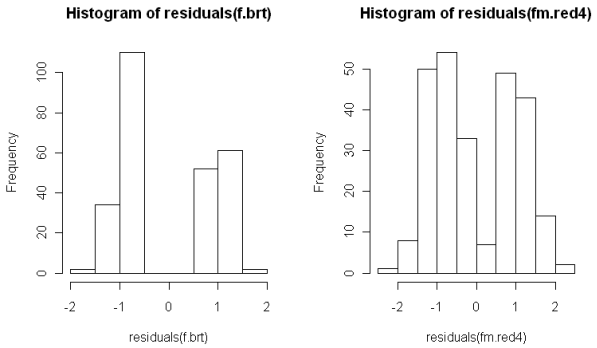
```
                         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       260     358.13
SalinityBMean             1    10.40       259     347.73 0.0012614 **
SalinityMean              1     0.42       258     347.32 0.5194294
PrimProdMean              1     2.73       257     344.59 0.0987485 .
I(DepthSD^2)              1    14.10       256     330.49 0.0001735 ***
poly(SSTAnMean, 3)        3    26.92       253     303.57 6.111e-06 ***
SalinityMean:PrimProdMean 1    18.77       252     284.80 1.477e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

[6] AIC (Akaike Information Criterion) penalises every parameter in the model with a factor of 2. The BIC penalises heavier, with a factor of $\log_e$(effective sample size), i.e. 4.7. This yields smaller models, particularly with many data points.

Model diagnostics

Now that we have a "final models", it's time to have a closer look at them. First, we want to get a feeling whether the error distribution has been handled properly:

```
par(mfrow=c(1,2))
hist(residuals(f.brt))
hist(residuals(fm.red4))
```



In both cases (BRT left and GLM right), residuals clump around two values (1/–1). This indicates that a substantial part of the data were estimated as absent when present (leading to positive values) or vice versa (negative values). The BRT is more "categorical" in its predictions, making the residuals larger. Note that these values are at the "link-scale", and hence a value of 1 represents a probability of

`plogis(1)` # 0.73.

But: are they good? Model fit is given by the reduction in deviance:

```
summary(fm.red4)
```

An intercept-only model ("null model") has a deviance of 64.3, while our final model reduces this to 57.6 - a moderate decrease. Expressing this as R^2 is tricky, there are different definitions for R^2s for GLMs. A pseudo-R^2 (also named D^2, because it is based on deviance) can easily be calculated as (358.1–284.8)/358.1 = 0.204, i.e. 20% explained deviance. In GLMs, explained deviance in a good model rarely exceeds 0.3, but 0.2 is not a yet a really good model. Often we use AUC to express the ability of a model to discriminate between 0s and 1s.

```
require(verification)
```

For the GLM, AUC is:

```
roc.area(obs=SW.t.red$PAsperm,                pred=predict(fm.red4,
type="response")) #0.78
```

For the BRT, we can calculate the same (but need to give some more information)[7]:

[7] However, BRT also returns this value when fitting (scroll in your R-window or repeat the BRT analysis and check): `training data ROC score = 0.883`

```
roc.area(obs=SW.t.red$PAsperm, pred=predict(f.brt, newdata=SW.t.red,
type="response", n.trees=200)) #0.883
```

Additionally, it gives the average cross-validation AUC-value, which is a much better indication of the model's predictive performance (and usually much lower):

```
cv ROC score = 0.677 ; se = 0.037
```

It is also extractable using:

```
mean(f.brt$cv.roc.matrix)
```

This drop in AUC from training to cross-validation is typical for weak models. Another thing to inspect binomial (and poisson) GLMs for is overdispersion. An estimate of dispersion is calculated by dividing the residual deviance by the residual degrees of freedom:

284.8/252 #1.13

High values (say, > 2 or so) indicate a problem[8].

To summarise model fits:

- Models are not spectacularly good, but not all that bad either.
- Predictive performance for models with an AUC around 0.8 is usually deemed "moderate".
- BRT: Annual mean sea surface temperature and salinity at the bottom are most relevant.
- GLM: Annual mean sea surface temperature and also bottom salinity are most relevant.

At least the two models are somewhat consistent!

Spatial Autocorrelation

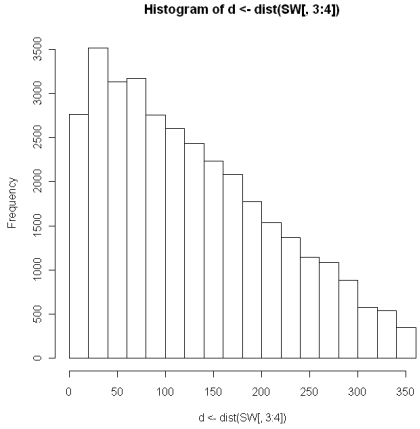
In a sense, testing for spatial autocorrelation is also part of model diagnostics. Spatial autocorrelation (SAC) in the model *residuals* is indicative of a violation of the assumption of data independence, and is all too often violated when analysing spatial data. We can investigate SAC by plotting a correlogram, which depicts similarity of residuals as a function of distance in space.

To get a feeling for the distances, we first make a histogram of Euclidean distances (which is wrong, because the earth is a sphere where -175° is close to +175° and no distance can be greater than 180*sqrt(2)). We are here, however, only interested in the spacing (grain size) and the min distance.

```
hist(d <- dist(SW[,3:4]))
min(d)
```

[8] By choosing the option "quasibinomial" as family (which yields no AIC and can hence not be used with stepAIC), we can fit an overdispersion parameter, which corrects the estimates of the parameter errors.

The shortest distance between two cells is 5° (as we know from the initial resolution), the histogram proposes steps of about 20 units. Now we plot a correlogram, acknowledging also the spherical nature of the data set (by setting latlon to TRUE; this leads to the transformation of degrees into kilometres).



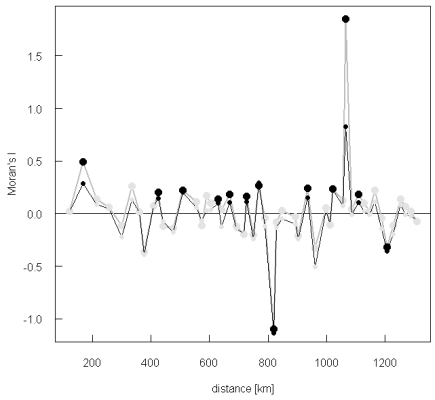

```
require(ncf)
cor.fglm <- correlog(x=SW$CenterLong, y=SW$CenterLat,
z=residuals(fm.red4), increment=20, resamp=999, latlon=T)
plot(cor.fglm)
abline(h=0)
```

This plot (not shown) is a bit cluttered and "long". Distance-classes over about 400 units[9] (i.e. 400 * 20 km = 8000 km) contain only few data points[10]. Moreover, it is unlikely that biological mechanisms lead to separation or aggregation of sperm whales at distances of more than, say, 1000 km. We can thus truncate the plot at 1000/20 = 50 units.

```
plot(cor.fglm$mean.of.class[1:50], cor.fglm$correlation[1:50],
type="n", lwd=2, col="grey", xlab="distance [km]", ylab="Moran's I",
las=1, tcl=0.5)
abline(h=0)
lines(cor.fglm$mean.of.class[1:50], cor.fglm$correlation[1:50],
lwd=2, col="grey")
points(cor.fglm$mean.of.class[1:50], cor.fglm$correlation[1:50],
pch=16, col=ifelse(cor.fglm$p<0.05, "black", "grey90"), cex=1.5)
```

[9] The unit is given by the original data, or, in this case, in km, because we used the option "`latlon=T`" to tell the function that we use degrees. It then automatically transforms data into kilometers (see `?correlog`). By setting "`increment=20`", we use 20 km distance classes. A bit confusing, admittedly.

[10] Type `cor.fglm$n` to get a listing of the number of comparisons per distance bin.



What we see is a significant positive spatial autocorrelation at 200km and again at around 500, 620-640, ... and a negative SAC at 820 km. We shall now try to account for this pattern. From experience, we would not expect to remove all of this pattern, but primarily the peak at 200 km, and perhaps those around 600-800 km. The method we use is called "spatial eigenvector mapping" or "principal coordinates of neighbourhood matrix" (Dray et al. 2006; Griffith and Peres-Neto 2006). First, we construct a list, which contains the neighbours' IDs for each data point. There are several ways to do so[11]. Note that coordinates must be given as long-lat, not lat-long which is consistent with plotting, but not with common usage)! Values for d1 and d2 must be given in kilometres.

```
require(spdep)
SW.nb <- dnearneigh(as.matrix(SW[,4:3]), d1=0, d2=2000, longlat=T)
summary(SW.nb)
ME.fit <- ME(fm.red4$formula, listw=nb2listw(SW.nb), data=SW.t.red,
alpha=0.5)
SW.t.red <- cbind(SW.t.red, fitted(ME.fit))
f <- as.formula(paste("PAsperm ~ ", fm.red4$formula[3], "+
vec1+vec3", sep=""))
fm.red4.me <- glm(f, data=SW.t.red, family=binomial)
anova(fm.red4.me)
```

So, the spatial dimension has almost as much relevance as the SSTAnMean-effect. Let us compare model coefficients to see whether these were affected by spatial autocorrelation:

```
summary(fm.red4)
summary(fm.red4.me)
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.2914     0.2100  -1.388 0.165280
```

[11] In R, check: `?knearneigh, ?dnearneigh, ?gabrielneigh`

```
SalinityBMean               1.1655     0.3400   3.428 0.000609 ***
SalinityMean               -0.5417     0.3218  -1.683 0.092301 .
PrimProdMean                0.3425     0.2519   1.360 0.173911
I(DepthSD^2)               -0.5261     0.1502  -3.504 0.000459 ***
poly(SSTAnMean, 3)1        -1.2922     4.1035  -0.315 0.752835
poly(SSTAnMean, 3)2       -11.8564     4.2140  -2.814 0.004899 **
poly(SSTAnMean, 3)3        11.9884     4.1264   2.905 0.003669 **
SalinityMean:PrimProdMean   0.9693     0.2560   3.787 0.000153 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 358.13  on 260  degrees of freedom
Residual deviance: 284.80  on 252  degrees of freedom
AIC: 302.80

> summary(fm.red4.me)
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.3676     0.2176  -1.689 0.091146 .
SalinityBMean               0.8078     0.3641   2.218 0.026528 *
SalinityMean               -0.2969     0.3401  -0.873 0.382675
PrimProdMean                0.4991     0.2724   1.832 0.066937 .
I(DepthSD^2)               -0.3755     0.1527  -2.459 0.013945 *
poly(SSTAnMean, 3)1        -1.2124     4.4624  -0.272 0.785866
poly(SSTAnMean, 3)2       -19.9111     4.9962  -3.985 6.74e-05 ***
poly(SSTAnMean, 3)3        14.2866     4.5395   3.147 0.001648 **
vec1                       10.5542     2.9731   3.550 0.000385 ***
vec3                       -7.1767     2.6968  -2.661 0.007786 **
SalinityMean:PrimProdMean   0.7996     0.2506   3.191 0.001417 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   Null deviance: 358.13  on 260  degrees of freedom
Residual deviance: 264.62  on 250  degrees of freedom
AIC: 286.62
```

Indeed! Parameter errors became smaller and parameters changed by tens of percent! Finally, let us investigate whether this new model has lower spatial autocorrelation in its residuals:

```
cor.fglm.me <- correlog(x=SW$CenterLong, y=SW$CenterLat,
z=residuals(fm.red4.me), increment=20, resamp=999, latlon=T) # takes
a while (1 min or so)
lines(cor.fglm.me$mean.of.class[1:50], cor.fglm.me$correlation[1:50],
lwd=1, col="black")
```

```
points(cor.fglm.me$mean.of.class[1:50],
cor.fglm.me$correlation[1:50], pch=16, col=ifelse(cor.fglm$p<0.05,
"black", "grey90"), cex=1)
```

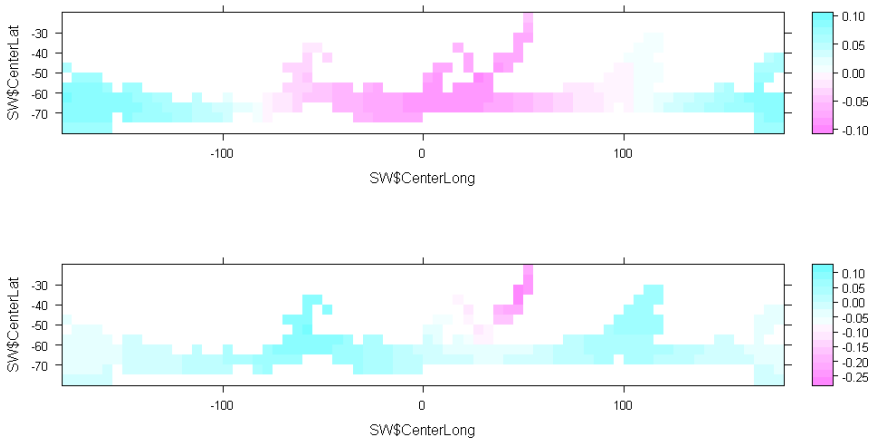
There is little difference overall (graph not shown), main effect at 200 and 600 km and at the funny peak at 1400 km. This indicates that our approach to adderss spatial autocorrelation has improved the model only marginally. Given that spatial eigenvectors are a very flexible method, it is unlikely that other approaches (reviewed in Carl et al. 2008; Dormann et al. 2007) would do a much better job. Just for fun, let us have a look what the eigenvector 1 looks like on a map:

```
require(lattice)
levelplot(SW.t.red$vec1 ~ SW$CenterLong+SW$CenterLat, aspect="iso")
```

This eigenvector thus codes for some effect that is high around Australia (+/- 180°, see the upper panel of the figure below) and low around South Africa (around 0°).

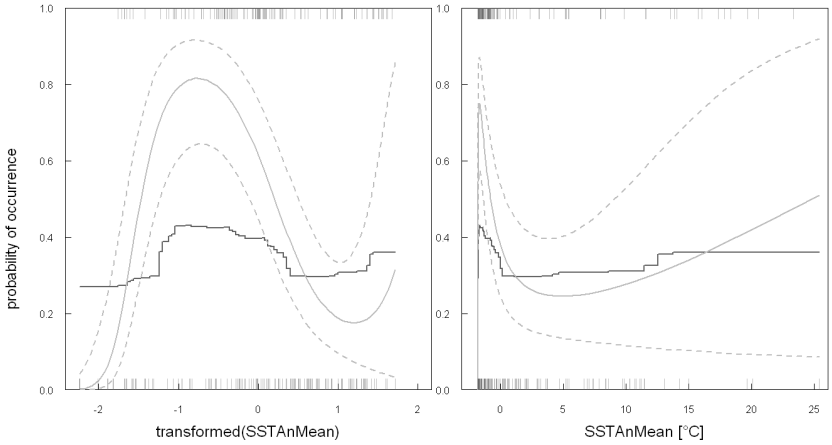
```
levelplot(SW.t.red$vec3 ~ SW$CenterLong+SW$CenterLat, aspect="iso")
```

Vec 3 in contrasts codes a much more fine-scaled pattern of unknown origin (representing only that part of the cruises which went up the eastern South African coast).



Interpretation

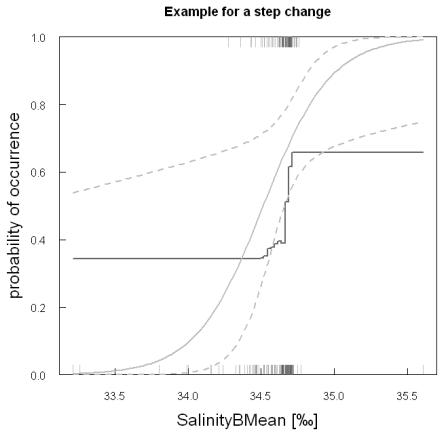
To be able to interpret the models, we have to make plots of the functional relationships they describe. To do so, we can also back-transform predictor variables (first un-standardise, then un-transform). The effect of SSTAnMean on occurrence probability for BRT (black line) and GLM (grey lines plus CI) are depicted here for both untransformed and transformed annual mean sea surface temperature:



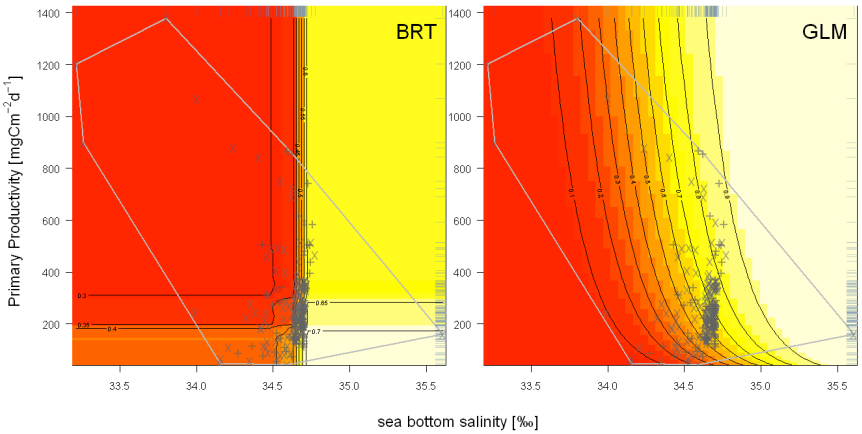
There are some obvious messages that we can take from these two graphs. Firstly, the BRT make step-like predictions, while the GLM fit a smooth function. Secondly, GLM seems to be more responsive, predicting greater changes in occurrence probability. This is, however, not a general feature of GLMs and possibly even an exception. Thirdly, the wide peak dominating the left panel (at the transformed scale) becomes a narrow spike on the untransformed axis (right panel). Here, the moderate but steady increase from values higher than approximately 3°C is more apparent (but also highly uncertain!). For all other model parameters, a similar plot is indicated and only omitted here due to spatial constraints.

We close this visual exploration with an illustration that variables in an interaction should not be studied on their own. The interaction in question (in the GLM) is that of SalinityBMean and PrimProdMean. Let us first look at a partial plot for SalinityBMean[12].

[12] The R-code for this and the following graph are too long to be given here, but are available as supplementary material along with the actual data.



We see a step-change in occurrence probability as we go from values lower than 0 to higher. The GLM is representing this step-change more gradually, due to the nature of the fitted logistic function. Since SalinityBMean interacts with PrimProdMean, we should really be plotting this as a 3-D graph. Of the many ways to do so, here is one:

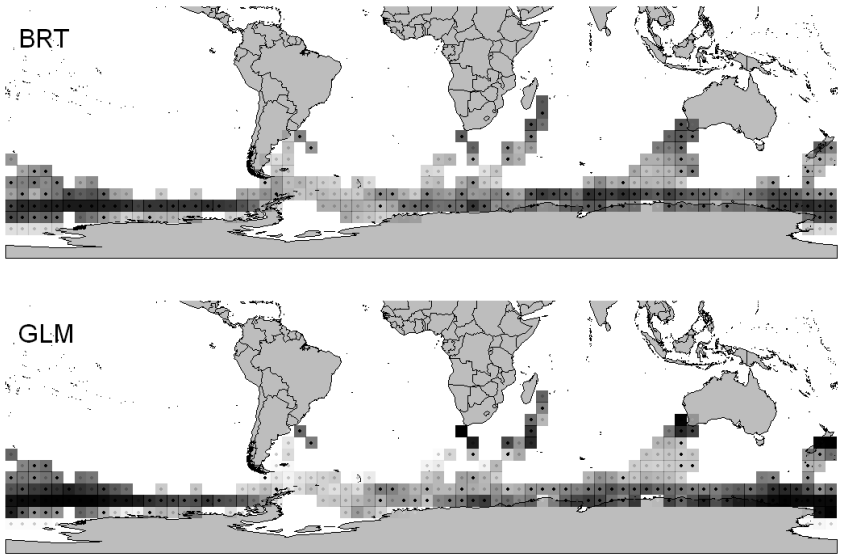


The dominant pattern is the transition from low probability values (red) to high values (in yellow) as we move from the left to the right[13]. This is the same as the step-change in the previous graph. However, we now also notice that this step-change is displayed (in the GLM) under very low productivity. For the BRT, the whole parameter space is divvied up into four rather discrete regions, also indicating an interaction and prominently displaying the step-changes induced by both salinity and productivity. And we notice, in this representation, that we can only

[13] This palette of colours is called heat colours. This is why red is colder than white. While the default and logical, it is not intuitive. It's easy to inverse, however.

make deductions (and, for that matter, predictions) about the region at the bottom centre, where the actual data points are, not even for the whole convex hull vector space spanned by the data points.

Finally, because we love being fooled by maps, we can plot the model predictions for sperm whale occurrence probabilities. Darker colours indicate higher occurrence probability, and observed presences (absences) are indicated by black (grey) points.



Thus, while the image/contour-plots indicate some disagreement between BRT and GLM, this is not detectable on the maps. Both predict high occurrence probabilities in the south Pacific and low values in the south Atlantic, with the Indian Ocean being in intermediate. One thing noticeable is that the GLM seems to be predicting higher values near ports visited by the survey vessels (Cape Town, ZA; Perth, AUS; Wellington, NZ).

Ecologically, our analysis will not have revealed much new. The problem of habitat preferences of whales is the seemingly unstructured nature of the oceans (although our analysis shows that temperature and salinity pattern are present and relevant), the scarcity of data on whale sightings and, in consequence, the spatial aggregation of the data before the analysis.

Statistically, our analysis shows nicely that very similar fits can be obtained from two different models. This remains a correlative analysis, and the potential mechanisms revealed should be interpreted as hypotheses requiring further investigation in the field.

Final comments

1. If you have many steps in your analysis, and you are uncertain if they may lead to spurious results, use a null model to find out. In its simplest form, simply shuffle the response variable (`sample(SW$PAsperm)` in our case) and re-run the analysis 1000 times. Obviously you then have to automatise the entire procedure.
2. In a more realistic setting, you may want to shuffle the response but maintain the spatial structure. This can be done, too: see Beale et al. (2008) for methods and R-code.
3. If you want to extrapolate beyond the geographic region or the parameter space, everything becomes rather uncertain. Note the 95% CI in the above plots, and how the wide towards the limits of the variable's range! The spatial eigenvectors (SV) are constructed in a way that their mean effect is 0. When you want to intra- or extrapolate, you have to make a "map" of the eigenvectors and use kriging to extrapolate in space to unobserved sites.

Acknowledgments We are indebted to C. Allison and the Secretariat of the International Whaling Commission for providing us with the IWC sighting data sets. CFD was funded by the Helmholtz Association (VH-NG-247).

Literature

Beale CM, Lennon JJ, Gimona A (2008) Opening the climate envelope reveals no macroscale associations with climate in European birds. Proceedings of the National Academy of Science USA 105:14908-14912

Bolker BM (2008) Ecological Models and Data in R. Princeton University Press, Princeton, NJ

Branch TA (2007) Abundance of Antarctic blue whales south of 60°S from three complete circumpolar sets of surveys. Journal of Cetacean Research and Management 9:253-262

Branch TA, Butterworth DS (2001) Estimates of abundance south of 60 degree S for cetacean species sighted frequently on the 1978/79 to 1997/98 IWC/IDCR-SOWER sighting surveys. Journal of Cetacean Research and Management 3:251-270

Breiman L (2001) Random Forests. Machine Learning 45:5-32

Butterworth DS, DeDecker JB (1989) Estimates of abundance for Antarctic blue, fin, sei sperm, humpback, killer and pilot whales from the 1978/79 to 1985/86 IWC/IDCR sighting survey cruises (SC/41/O20). International Whaling Commission - Scientific Committee Meeting, IWC

Carl G, Dormann CF, Kühn I (2008) A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data. Web Ecology 8:22-29

Corkeron PJ, Ensor P, Matsuoka K (1999) Observations of blue whales feeding in Antarctic waters. Polar Biology 22:213-215

Crawley MJ (2002) Statistical Computing: An Introduction to Data Analysis using S-Plus. Wiley, New York

Davis RW et al. (2002) Cetacean habitat in the northern oceanic Gulf of Mexico. Deep Sea Research (Part I): Oceanographic Research Papers 49:121-142

Dormann CF et al. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628

Draper NR, Smith H (1998) Applied Regression Analysis, 3rd edn. Wiley, New York

Dray S, Legendre P, Peres-Neto PR (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling 196:483-493

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal Of Animal Ecology 77:802-813

Goodall RNP (1997) Review of sightings of the hourglass dolphin, Lagenorhynchus cruciger, in the South American sector of the Antarctic and sub-Antarctic. Reports of the International Whaling Commission 47:1001-1013

Griffith DA, Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses in exploiting relative location information. Ecology 87:2603-2613

Hamazaki T (2002) Spatio-temporal prediction models of cetacean habitats in the mid-western North Atlantic ocean (from Cape Hatteras, North Carolina, U.S.A. to Nova Scotia, Canada). Marine Mammal Science 18:920-939

Harrell FE, Jr. (2001) Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York

Hastie T, Tibshirani R, Friedman JH (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Berlin

Hedley S et al. (2001) Modelling whale distribution: a preliminary analysis of data collected on the CCAMLR-IWC Krill Synoptic Survey, 2000. In: Paper SC/53/E9 presented to the 54th meeting of the Scientific Committee of the International Whaling Commission (London, 2001). URL: http://www.iwcoffice.org, London, UK

IWC (2001) IDCR-DESS SOWER Survey data set (1978-2001). IWC

Kasamatsu F, Ensor P, Joyce GG, Kimura N (2000a) Distribution of minke whales in the Bellingshausen and Amundsen Seas (60 degree W-120 degree W), with special reference to environmental/physiographic variables. Fisheries Oceanography 9:214-223

Kasamatsu F, Hembree D, Joyce G, Tsunoda LM, Rowlett R, Nakano T (1988) Distributions of cetacean sightings in the Antarctic: Results obtained from the IWC/IDCR minke whale assessment cruises 1978/79 - 1983/8. 4. Reports of the International Whaling Commission 38:449-487

Kasamatsu F, Matsuoka K, Hakamada T (2000b) Interspecific relationships in density among the whale community in the Antarctic. Polar Biology 23:466-473

Kato H, Miyashita T, Shimada H (1995) Segregation of the two sub-species of the blue whale in the Southern Hemisphere. Reports of the International Whaling Commission:273-283

Laran S, Gannier A (2008) Spatial and temporal prediction of fin whale distribution in the northwestern Mediterranean Sea. ICES Journal of Marine Science 65:1260-1269

Matsuoka K et al. (2003) Overview of minke whale sightings surveys conducted on IWC/IDCR and SOWER Antarctic cruises from 1978/79 to 2000/01. Journal of Cetacean Research and Management 58:173-201

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1993) Applied Linear Statistical Models, 4th edn. McGraw-Hill, Boston, MA

Praca E, Gannier A (2008) Ecological niches of three teuthophageous odontocetes in the northwestern Mediterranean Sea. Ocean Science 4:49-59

Quinn GP, Keough MJ (2002) Experimental Design and Data Analysis for Biologists. Cambridge Univ. Press, Cambridge

Sokal RR, Rohlf FJ (1995) Biometry, 3rd edn edn. Freeman, New York

Torres LG, Read AJ, Halpin PN (2008) Fine-scale habitat modeling of a top marine predator: Do prey data improve predictive capacity? Ecological Applications 18:1702-1717

Underwood AJ (1997) Experiments in Ecology: Their Logical Design and Interpretation using Analysis of Variance. Cambridge University Press, Cambridge

Zar JH (1996) Biostatistical Analysis, 3rd edn. Prentice Hall, Upper Saddle River

Exercise

The data show a clear relationship between sampling effort and probability of sperm whale sighting. This variable has not been used to correct the data.

As exercise we recommend repeating the entire analysis, but forcing TotalEffort to always be in the model (i.e. as a correcting covariate).

In several places, this will lead to problems that need to be overcome. For example, a stepwise model simplification with one variable always in the model needs to be done either by hand, or by scrutiny of the function stepAIC and its help page.

If you manage to replicate this analysis, and do the same with the sampling effort correction, you will have mastered more than most species distribution analyses published to date!

Good luck!