Development of Generalized Additive Models (GAMs) in EnviEFH

1. Introduction

The US Congress defined Essential Fish Habitats (EFH) as 'those waters and substrate necessary to fish for spawning, breeding, feeding, or growth to maturity', a definition that includes the physical, chemical and biological properties of marine areas and the associated sediment and biological assemblages that sustain fish populations throughout their full life cycle (DOC, 1997).

Generalized Additive Models (GAMs) are a powerful tool for modelling fisheries data with other data that characterise species EFH areas. GAMs (Hastie *et al*., 2001) are straightforward extensions of additive modelling with differences on (i) the way the response variable is linked with the explanatory variables, and (ii) the distribution function of the data (Zuur *et al*., In press).

The general GAM formula is:

$$g(\mu_i) = \mu + \sum_{j=1}^{p} f_j(\mathbf{x}_i)$$

where *g* is the differentiable and monotonic link function, $\mu i = E(Yi)$ is the expectation of the response, $\sum_{j=1}^{p} f_j(\mathbf{x}_i)$ is a function called additive predictor where *fj* is a smooth function (such as a spline or a loess smoother). The degree of smoothness achieved is balanced against the deviance by a tuning constant, often chosen by cross-validation, so that estimation is by the method of maximum 'penalized' likelihood rather than of maximum likelihood. This gives GAMs a partially non-parametric aspect (Maunder and Punt, 2004).

Here, the GAM development process in EnviEFH is explained on step-by-step basis.

2. Data selection

A GAM model for EFH is generally described as:
Response variable = s(explanatory var.1) + s(explanatory var.2) … + s(explanatory var.i), where s is a smoother.

Different types of fisheries data can be used as response variables (e.g. biomass index, sonar data etc). The selection of the proper explanatory data based on those parameters that describe more efficient an EFH, according to information on species life history. As explanatory data, environmental parameters are used includin environmental satellite and model data, such as sea surface temperature, chlorophyll-a concentration, salinity, altimetry, photosynthetically active radiation, substrate types, bathymetry, etc.

3. Exploration

The exploration process is very important because the next step of the analysis require the data to comply with several assumptions before any valid conclusions can be made (Zuur *et al*., In press). Exploration routines that provide a clear graphical idea of each dataset are described below (Fig. 1):

Boxplots is a tool for identifying outliers. A boxplot visualises the mean and spread for a univariate variable. Normally, the midpoint of a boxplot is the median, but it can also be the mean. The 25% and 75% quartiles define the hinges and the difference between the hinges is called the spread. Lines are drawn from each hinge to 1.5 times the spread or to the most extreme value of the spread, whichever is the smaller. Any point outside these values is normally outlier.

Dotplots or Cleveland dotplots (Cleveland, 1985) are useful to identify outliers and homogeneity. Homogeneity means that the variance of the data does not change along the gradient.

QQ-plots or Quantile-Quantile plots are graphical tools used to determine whether the data follow a particular distribution.

Coplots are conditional scatterplots that show relationship between x and y, for different values of a third variable z. Coplots are useful for detecting interactions between the explanatory variables.

Pairplots are multiple pair-wise scatterplots in one graph and can be used to detect relationships between variables and to detect collinearity.

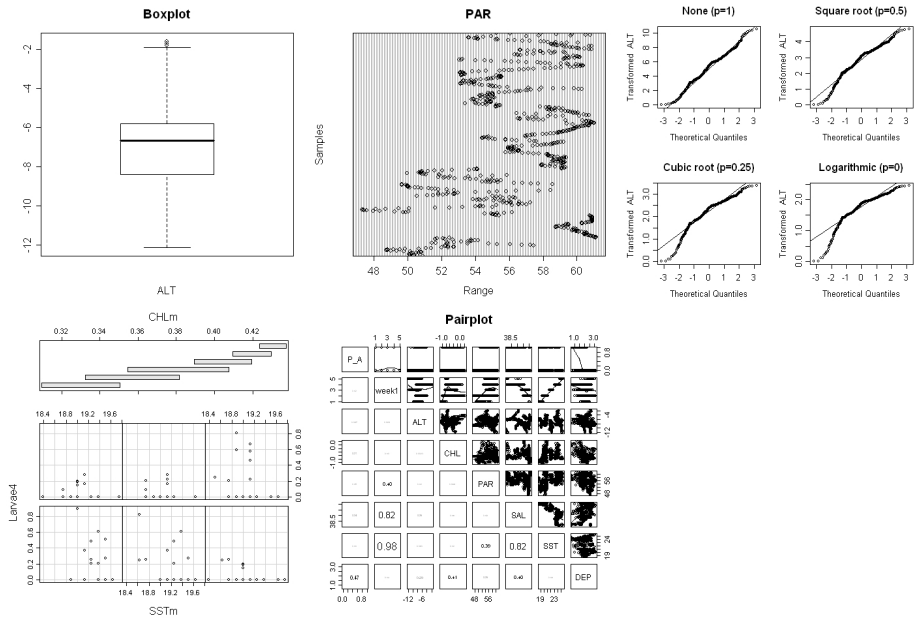

Fig. 1: From left to right: boxplot, dotplot, QQ-plot, coplot and pairplot data exploration routines.

The spread of the variables and the outliers are factors that can affect a GAM model. In some cases, data transformation is required because it provides a better data fit in the model. Both response and explanatory variables are transformed, and

different types of transformations are applied to different variables within the same dataset.
Inclusion of explanatory variables that are themselves correlated, the so-called problem of 'collinearity' is avoided. This can make the model fitting process numerically unstable or lead to problems similar to those of over-fitting (Maunder and Punt, 2004).

4. Create a model

The next step is the composition of a primary submodel. This contains the response variable, the explanatory un-correlated variables, the proper distribution and a link factor. Spline is used as smoother in EFH. In the first submodel the same degrees of freedom are used between all the explanatory factors. For count data, Poisson distribution with log link is suggested. If the data are presence-absence, the binomial distribution with logit link should be used. Table 1 shows distributions and related links for some commonly used models.

Distribution	link
Normal	Identity
Binomial	Logit
Gamma	Reciprocal
Gamma	Log
Poisson	Log
Inverse Gaussian	μ^{-2}

Table 1: Distributions and related links for commonly used models (Hastie and Tibshirani, 1990).

The first submodel is a 'control' model that provides the 'best' fitted model through a selection process.

5. Model Selection

In a model's numerical output different factors provide several information about it (Fig. 2). P-values for smoothing terms show the significance of the terms in the model. Different nested models are possible to be compared with ANOVA test. A non significant variable can take part in a model. An alternative way to compare different models (not necessarily nested) is the Akaike information criterion (AIC, Akaike, 1973; Burnham and Anderson, 2002). Lower AIC characterize a better fitted model. Another value that provides crucial information about the model is the deviance explained. At last, over-dispersion (not over 1) must be considered and sometimes, it must be corrected by using a quasi-distribution.

	Df Npar	Df Npar	Chisq	P(Chi)
(Intercept)	1			
s(ALT, 5)	1	4	15.0322	**0.0046**
s(PAR, 5)	1	4	12.5302	**0.0138**
s(DEP, 3)	1	2	31.0024	**1.851e-07**

Dispersion parameter = 1
Deviance = 504.93
n (null degrees of freedom) = 718
df.residual (residual degrees of freedom) = 701
df (n-df.residual) = 17

Overdispersion (Deviance/df.residual) = 0.72

AIC according to formula: -2log(Likelihood) + 2*df = 540.93

Fig. 2: A GAM numerical output.

After the selection of the significant smoothing terms, we have to choose the best combination of the degrees of freedom for the explanatory variables. Step-wise search is a way that gives the 'best' model. Any final model must be validated. We have to verify the assumptions of homogeneity and normality and check for potential influential observations. If the fitted values against the residuals (Fig. 3) show a clear spatial pattern then the model is not valid. In this case, another model selection is required, with the use of different terms and perhaps transformations of the initial data.

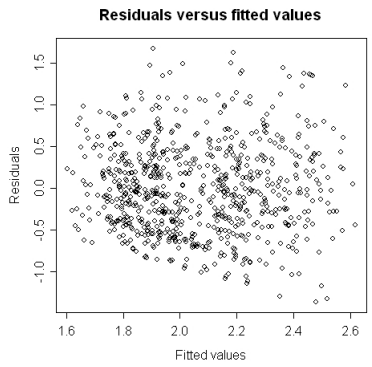

Fig. 3: Fitted values against the residuals.

6. Ranges

The partial plots (Fig. 4) for the explanatory variables are the model output that provides environmental ranges for EFH mapping. We get the range, from each plot, that has a positive effect on the fitted values (e.g. range of environmental parameter that is over the zero axis).

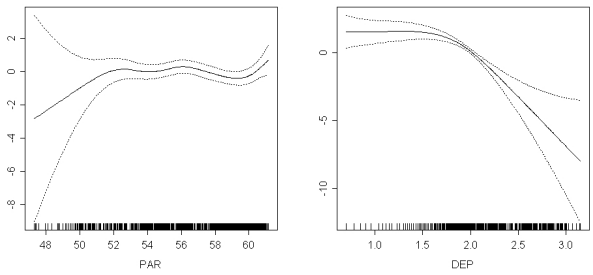

Fig. 4: Partial plots of two explanatory variables.

7. EFH maps

By applying those ranges on GIS grids, we map areas where these ranges are simultaneously met and imply potential essential fish habitats (Fig. 5). Environmental ranges extracted from a specific surveyed area (e.g. North Aegean Sea in Eastern Mediterranean) is applied to satellite data that cover the whole Mediterranean basin, thus providing EFH maps for the whole Mediterranean Sea.

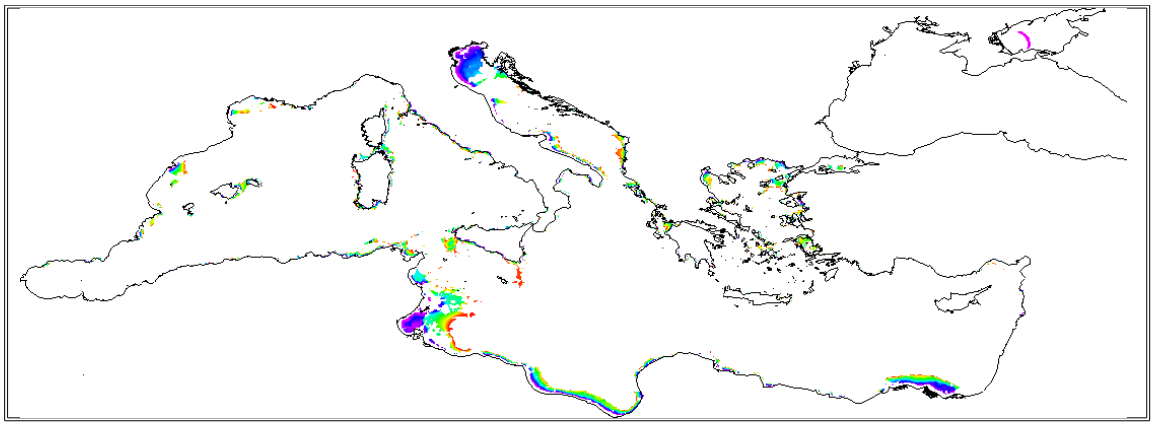

Fig. 5: EFH map in Mediterranean Sea.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the 2nd International Symposium on Information Theory. Publishing House of the Hungarian Academy of Sciences, Budapest, pp. 268–281.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach, 2nd ed. Springer-Verlag, New York.

Cleveland W. S., 1985. *The Elements of Graphing Data*, Monterey, CA: Wadsworth.

DOC, 1997. Department of Commerce. Magnuson–Stevens Act Provisions: Essential Fish Habitat (EFH). Federal Register, vol. 62, issue 244, pp. 66531–66559.

Hastie T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, New York.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer- Verlag, New York.

Maunder M. N. and A. E. Punt, 2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research, 70, 141–159.

Zuur, A.F., Ieno, E.N. and Smith, G.M., In press, March 2007. *Analysing Ecological Data*. Springer. 700p. *Series: Statistics for Biology and Health.*