Functional Data Analysis for Predicting Landed Fish Abundance per unit effort

Manuel Oviedo de la Fuente [1]

manuel.oviedo@udc.es

Raquel Menezes [2]

Alexandra A. Silva [3]

[1] CITIC, A Coruña University, Spain

[2] CMAT, Minho University, Portugal

[3] Portuguese Institute for the Sea and Atmosphere (IPMA), Portugal

Introduction

Predicting the abundance of landed fish per unit effort (LPUE) is a critical challenge in competitive fish markets (Hilborn & Walters, 1992). Previous research (Maunder & Punt, 2004) has addressed the challenge of modelling species distribution in fisheries using various statistical methods, including time series analysis (e.g., ARIMA models) (Box *et al.,* 2015), model-based geostatistics (e.g., SPDE approach, GRFs and kriging) (Lindgren *et al.,* 2011), and regression models (e.g., GLMs) to model complex structures.

This study addresses the challenge of variable selection by employing *distance correlation ($\mathcal{DC}$)* (Szekely *et al.,* 2007) to investigate the relationships between environmental data (*functional data*) and other sources of information, such as sale prices, ratio of euros per total catches, calendar variables, and the scalar response, LPUE.

LPUE study

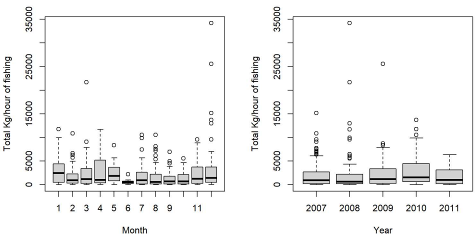
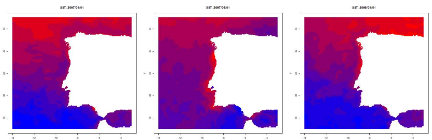


Figure 1: LPUE distribution by month and year

Ocean Monitoring

We use sensor data monitoring, such as chlorophyll-a concentrations (CHL), intensity of ocean currents , Sea Surface Temperature (SST), wind speed and wind direction curves (WS, WD) measured daily during 10 yrs.

SST in 2007/01/01, 2007/07/01 and 2008/01/01



LPUE captures (black) and 2 SST locations

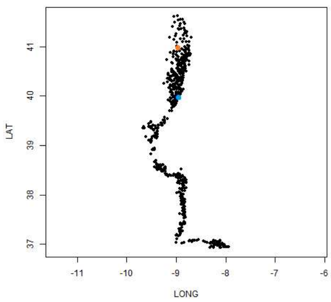
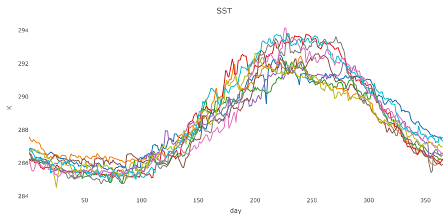


Figure: Daily sea surface temperature levels (2007-2020 28028 locations)

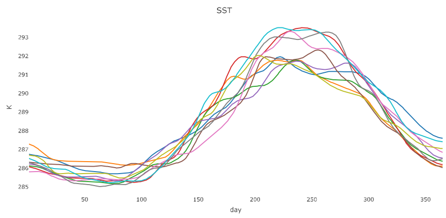
Functional data analysis (FDA)

This study proposes an approach based on FDA (Ramsay et al. 2005). FDA is a branch of statistics that focuses on the analysis of data consisting of curves or anything else that varies along a continuum.

SST raw curves (1 location, 10 years): $\mathcal{X}(t)$



SST smoothing curves: $\hat{\mathcal{X}}(t) = \mathcal{S}\mathcal{Y}(t)$



Dimension reduction using functional principal components (FPC): $\mathcal{X}(t) = \mu(t) + \sum_k c_k v_k(t),\ c_k$

```
## [1] TRUE
##
##      - SUMMARY:   create.pc.basis  object   -
##
## -With 2  components are explained  69.23 %
##  of the variability of explicative variables.
##
## -Variability for each component (%):
##    PC1    PC2
## 50.27 18.96
```

Functional Additive Model (FAM) with variable selection (VS) using Distance Correlation $\mathcal{DC}$

We present a step-wise method to select the features based on the calculation of the $\mathcal{DC}$ between each feature and the response variable (or model residals). $\mathcal{DC}$ is defined for X and Y random vector variables in arbitrary finite dimension spaces, $\mathcal{DC}(X,Y) = 0$ characterizes the independence of X and Y. For simplicity we use the FAM model as it allows for non-linear estimates of the effects. The covariates can be used from the `last` observed value, its mean or the principal component `PC` representation of the curve of the values recorded in the last 30 days (e.g.), etc.

$$LPUE_{h,s} = s_1(LONG)_h + \ldots + s_j(SST(t))_{h,s} + \ldots \quad + \varepsilon_{h,s}$$

where s_j is a smooth function and, $\varepsilon_{h,s} \overset{iid}{\sim} N(0,\sigma)$

Results: Train data $h = 1,\ldots,399,\ \ B = 100$ replications, $b = h+1,\ldots,h+B$

From more than 50 covariates were selected: ratio of euros per total catches (`taxa`), ocean current (`foclast`) and v-component (south to north flow) of wind direction (fvlast), bathymetry (`bath`) and longitude (`long`) and functional PC of the derivative of: wind intensity curve (`dint`) and smoothed chlorophyll (`schl`).

% of times each variable enters the model

Show 6 entries Search:

	Percent
rate	100
foclast	56.1
time index	43.9
sstlast	26.5
ssst	20.4
bath	17.3

Showing 1 to 6 of 28 entries

Previous 1 2 3 4 5 Next

Table 1: Example of a model fit

Component	Term	Estimate	Std Error	t-value	p-value
A. parametric coefficients	(Intercept)	7.011	0.076	92.095	0.0000 ***
Component	**Term**	**edf**	**Ref. df**	**F-value**	**p-value**
B. smooth terms	s(taxa)	6.013	6.659	5.598	0.0000 ***
	s(bath)	1.000	1.000	5.450	0.0203 *
	s(foclast)	1.000	1.000	7.740	0.0058 **
	s(fvlast)	1.000	1.000	2.387	0.1235
	s(long)	1.966	2.493	3.080	0.0437 *
	s(dint.PC1)	1.000	1.000	6.919	0.0090 **
	s(dint.PC2)	1.863	2.377	2.723	0.0567 .
	s(dint.PC3)	1.000	1.000	3.459	0.0640 .
	s(dint.PC4)	1.000	1.000	0.789	0.3752
	s(schl.PC1)	1.000	1.000	1.928	0.1660
	s(schl.PC2)	1.000	1.000	3.681	0.0561 .
	s(schl.PC3)	1.000	1.000	1.243	0.2659
	s(schl.PC4)	4.928	5.835	2.446	0.0311 *

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Adjusted R-squared: 0.240, Deviance explained 0.300
GCV : 1.901, Scale est: 1.745, N: 301

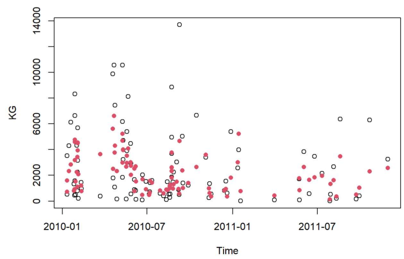


Figure 2: Observed (black) and predicted values (red)

Conclusion

The proposed functional approach has demonstrated promising results when applied to a real dataset LPUE of juvenile sardine along the northern Portuguese coast. These findings present decision makers with a valuable tool to advance marine sustainability and conservation efforts by enhancing our understanding of the factors influencing LPUE.

References

- Febrero-Bande, M, González-Manteiga, W, Oviedo de la Fuente, M. Variable selection in Functional Additive Regression Models. *Comput. Stat.*, 34, 469–87, 2019.
- Lindgren, F, Rue, H, Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields.J. R. Stat. Soc. Ser. B Methodol, 73(4), 423–98.
- Rodríguez-Climent, S, Angélico, M, Marques, V, Oliveira, P, Wise, L, Silva, A. Essential habitat for sardine juveniles in Iberian waters. *Sci. Mar.*, 81(3), 351–60, 2017.
- Székely, G J, Rizzo, M L, Bakirov, N K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* 35(6): 2769–2794.

Acknowledgements

This research/work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2020-14 and Centro de Investigación del Sistema universitario de Galicia ED431G 2019/01), all of them through the ERDF. Authors also acknowledge the FCT Foundation for funding their research through projects PTDC/MAT-STA/28243/2017, UIDB/00013/2020 and UIDP/00013/2020 and MAR2020 for funding the SARDINHA2020 project (MAR-01.04.02-FEAMP-0009).