# 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]

[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 REF.

Previous research [REF has addressed the challenge of modelling species distribution in fisheries using various statistical methods, including time series analysis (e.g., ARIMA models) REF, model-based geostatistics (e.g., SPDE approach, GRFs and kriging)REF, and regression models (e.g., GLMs) to model parametrically temporal, spatial and other complex structures REF.

This study addresses the challenge of variable selection by employing *distance correlation (DC)* to investigate the relationships between environmental data (*functional data*) and other sources of information, such as sale prices at landing, atio of euros per total catches, calendar variables, and the scalar response (LPUE).

## LPUE study

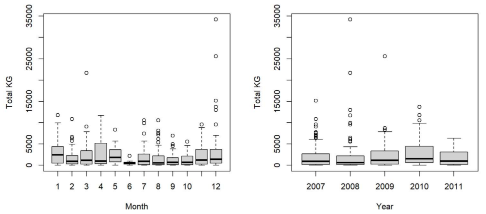
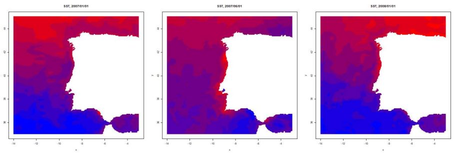


*Figure 1: LPUE distribution by month and year*

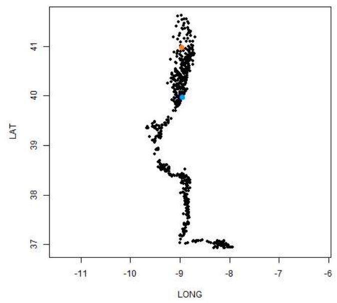
## Ocean Monitoring

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



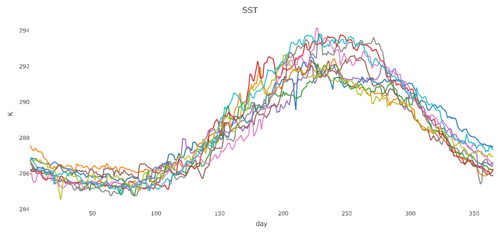
SST

*Figure 2: Figure: Daily sea surface temperature levels (2007-2020 28028 locations)*

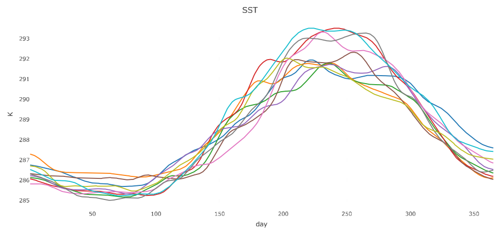
## Functional data analysis (FDA)

This study proposes an approach based on functional data analysis (FDA). 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: $\hat{\mathcal{X}}(t) = S\mathcal{Y}(t)$



### 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

Functional Additive Model with variable selection [Febero *et al.*, 2019] using $\mathcal{DC}$. The main idea behind this iterative procedure is that the residuals of the regression can capture information not collected in previous steps. FAM model: $LPUE_{h,s} = m_1(LONG)_h + \ldots + s(SST(t))_{h,s} + \ldots \quad + \varepsilon_{h,s}$ , where $m$ and $s$ are smooth functions and, $\varepsilon_{h,s} \overset{iid}{\sim} N(0,\sigma)$

*Nature of predictors*

| Type | | | |
|---|---|---|---|
| Escalar | $LONG_h$ | $LAT_h$ | $TIME_s$ |
| Escalar | $RATE_{h,s}$ | $SST_{h,s}$ | $S\bar{S}T_{h,s}$ |
| Functional | $SST_{h,s}(t_1)$ | $SST_{h,s}(t_2)$ | $SST^{'}(t)$ |
| Functional | $CHL_{h,s}(t_1)$ | $CHL_{h,s}(t_2)$ | $CHL^{'}(t)$ |

We present a step-wise method to select the optimal features to include in the model based on the calculation of the distance correlation ($\mathcal{DC}$) [Szekely et *al*, 2007] between each feature and the response variable. $\mathcal{DC}$ 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 $Y$ and $\mathcal{DC}$ satisfies $0 \leq \mathcal{DC} \leq 1$.

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

### % of times each variable enters the model

Show 6 entries    Search:

| | Percent |
|---|---|
| s(rate) | 100 |
| foclast | 56.1 |
| 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

| 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

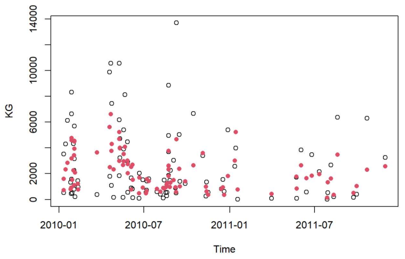


*Figure 3: Figure: Observed (black) and predicted values (red)*

## Conclusion

The proposed functional approachhas demonstrated promising results when applied to a real dataset LPUE of juvenile sardine along the northern Portuguese coast in 2007-2011. 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. and Oviedo de la Fuente, M. Variable selection in Functional Additive Regression Models. Computational Statistics, 34, 469–487, 2019. Rodríguez-Climent, S., Angélico, M. M., Marques, V., Oliveira, P., Wise, L., and Silva, A. Essential habitat for sardine juveniles in Iberian waters. *Scientia Marina*, 81(3), 351-360, 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.