Linear predictive functional model on environmental data: case of chlorophyll-a oceanographic profiles

Séverine Bayle$^1$, Pascal Monestiez$^1$, David Nerini$^2$

$^1$INRA, UR 546 Biostatistics and Spatial Processes (BioSP), F-84914 AVIGNON. $^2$Mediterranean Institute of Oceanography (MIO) - UMR 7294, Pytheas Institute (OSU), Aix-Marseille University, Campus de Luminy, Case 901, 13288 MARSEILLE Cedex 09.

7th Days of functional statistics, Montpellier, June 28-29, 2012
1. Introduction
2. Methodology
3. Results
4. Conclusion
Context and purpose of the study

Physical data (profiles) collected within the framework of ANR project IPSOS-SEAL between October 2009 and January 2010 in Southern Ocean around Kerguelen islands:

- Chlorophyll-a (Chl-a) : CTD-Fluo and Argos devices
- Brightness : TDR + GPS devices
Capturing elephant seals for installing devices
Capturing elephant seals for installing devices

Introduction
Capturing elephant seals for installing devices
Capturing elephant seals for installing devices
Capturing elephant seals for installing devices
Capturing elephant seals for installing devices
Capturing elephant seals for installing devices
Elephant seal dataset
Context and purpose of the study

- Primary productivity: production of vegetal matter
- Photosynthesis: permitted through the oceanic phytoplankton content in Chl-a
  → Vital link between living and inorganic stocks of carbon
Context and purpose of the study

- Primary productivity: production of vegetal matter
- Photosynthesis: permitted through the oceanic phytoplankton content in Chl-a
  - Vital link between living and inorganic stocks of carbon
- Measurement of Chl-a concentration throughout the water column in Southern Ocean is used as an indicator of the amount of phytoplankton and allows to know the distribution of primary productivity
Introduction

Context and purpose of the study

- Primary productivity: production of vegetal matter
- Photosynthesis: permitted through the oceanic phytoplankton content in Chl-a
  - Vital link between living and inorganic stocks of carbon
- Measurement of Chl-a concentration throughout the water column in Southern Ocean is used as an indicator of the amount of phytoplankton and allows to know the distribution of primary productivity
- Few Chl-a data profiles recorded: devices which record fluorescence are energy-intensive
Primary productivity: production of vegetal matter
Photosynthesis: permitted through the oceanic phytoplankton content in Chl-a
→ Vital link between living and inorganic stocks of carbon
Measurement of Chl-a concentration throughout the water column in Southern Ocean is used as an indicator of the amount of phytoplankton and allows to know the distribution of primary productivity
Few Chl-a data profiles recorded: devices which record fluorescence are energy-intensive
But a lot of brightness data profiles
Context and purpose of the study

- Primary productivity: production of vegetal matter
- Photosynthesis: permitted through the oceanic phytoplankton content in Chl-a
  → Vital link between living and inorganic stocks of carbon
- Measurement of Chl-a concentration throughout the water column in Southern Ocean is used as an indicator of the amount of phytoplankton and allows to know the distribution of primary productivity
- Few Chl-a data profiles recorded: devices which record fluorescence are energy-intensive
- But a lot of brightness data profiles
- Idea: reconstruct Chl-a profiles from brightness profiles
Context and purpose of the study

In order to calibrate relationships between 2 kinds of data profiles, only data profiles collected during day were kept.
In order to calibrate relationships between 2 kinds of data profiles, only data profiles collected during day were kept.

To be more accurate in estimation and smoothing of profiles, only Chl-a data profiles which have 18 observations recorded every 10 meters between -5 et -175 meters were kept (407 profiles selected)
Context and purpose of the study

- In order to calibrate relationships between 2 kinds of data profiles, only data profiles collected during day were kept.
- To be more accurate in estimation and smoothing of profiles, only Chl-a data profiles which have 18 observations recorded every 10 meters between -5 and -175 meters were kept (407 profiles selected).
- Selection of Chl-a and brightness data profiles collected at the same time: 208 profiles altogether.
In order to calibrate relationships between 2 kinds of data profiles, only data profiles collected during day were kept.

To be more accurate in estimation and smoothing of profiles, only Chl-a data profiles which have 18 observations recorded every 10 meters between -5 and -175 meters were kept (407 profiles selected).

Selection of Chl-a and brightness data profiles collected at the same time: 208 profiles altogether.

Reconstruction of one Chl-a data profile is made for each 208 pairs.
Functional data analysis

- Chl-a and brightness functional profiles can be considered as curves

\[ z_{ci}(t) = y_i(t) + \epsilon_i(t), \quad z_{bi}(s) = x_i(s) + \epsilon_i(s) \]
Functional data analysis

- Chl-a and brightness functional profiles can be considered as curves

\[ z_{ci}(t) = y_i(t) + \epsilon_i(t), \quad z_{bi}(s) = x_i(s) + \epsilon_i(s) \]

- Modeling these functional profiles needs definition of basis functions \( \phi_k, k = 1, \ldots, K \)
Functional data analysis

- Chl-a and brightness functional profiles can be considered as curves
  \[ z_{ci}(t) = y_i(t) + \epsilon_i(t), \quad z_{bi}(s) = x_i(s) + \epsilon_i(s) \]

- Modeling these functional profiles needs definition of basis functions \( \phi_k, k = 1, \ldots, K \)
- Functional profiles are defined as linear combinations of these basis functions:
  \[ y_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t), \quad x_i(s) = \sum_{k=1}^{K} d_{ik} \phi_k(s) \]
  - \( c_1, \ldots, c_K \) and \( d_1, \ldots, d_K \) : expansion coefficients
  - \( \phi_1, \phi_2, \ldots, \phi_K \) : basis functions
Functional data analysis

- Reconstruct functional profiles $y$ and $x$ using data $(t, z_{ci})$ and $(s, z_{bi}), i = 1, \ldots, n$
Reconstruct functional profiles \( y \) and \( x \) using data \((t, z_{ci})\) and \((s, z_{bi}), i = 1, \ldots, n\)

Utilisation of 10 splines of order 4

\[
\frac{1}{n} \sum_{i=1}^{n} (x(t_i) - y(t_i))^2 + \lambda \int (x''(u))^2 \, du
\]

\( \lambda \) : Trade-off between smoothness of the curve and sum of squared deviations between model and data
Functional data analysis

- Reconstruct functional profiles $y$ and $x$ using data $(t, z_{ci})$ and $(s, z_{bi}), i = 1, \ldots, n$
- Utilisation of 10 splines of order 4

$$1/n \sum_{i=1}^{n} (x(t_i) - y(t_i))^2 + \lambda \int (x''(u))^2 du$$

- $\lambda$: Trade-off between smoothness of the curve and sum of squared deviations between model and data
Methodology

Functional data analysis

- Reconstruct functional profiles $y$ and $x$ using data $(t, z_{ci})$ and $(s, z_{bi})$, $i = 1, \ldots, n$
- Utilisation of 10 splines of order 4

\[ 1/n \sum_{i=1}^{n} (x(t_i) - y(t_i))^2 + \lambda \int (x''(u))^2 du \]

- $\lambda$: Trade-off between smoothness of the curve and sum of squared deviations between model and data
- We work now with splines coefficients $c_k$ and $d_k$
Number of basis functions = number of knots + order of splines
We consider a fully functional linear model

Assumption: relationship between derivative of brightness function and Chl-a function

\[ y(t) = \alpha(t) + \int \beta(s, t)x(s)ds + \epsilon(t) \]

- \( y(t) \): Chl-a profile reconstructed (or predicted)
- \( t \) and \( s \): Depths
- \( x(s) \): Derivative of brightness function
- \( \alpha(t) \): Univariate coefficient (functional intercept)
- \( \beta(s, t) \): Bivariate coefficient
- \( \epsilon(t) \): Functional error
**Functional linear model**

- We consider a fully functional linear model
- Assumption: relationship between derivative of brightness function and Chl-a function

\[ y(t) = \alpha(t) + \int \beta(s, t)x(s)ds + \epsilon(t) \]

- \( y(t) \): Chl-a profile reconstructed (or predicted)
- \( t \) and \( s \): Depths
- \( x(s) \): Derivative of brightness function
- \( \alpha(t) \): Univariate coefficient (functional intercept)
- \( \beta(s, t) \): Bivariate coefficient
- \( \epsilon(t) \): Functional error

- FDA Package on R
Chl-a functional profiles well predicted...
...But some problems remain!
Cross validation

- Is 10 basis functions the optimal number to use?
Cross validation

- Is 10 basis functions the optimal number to use?
- Check by cross validation
  → Computation for each marine mammal
  → One profile is withdrawn (validation set), and others profiles represent training set
  → Calculation of mean square error
  → Repetition choosing another validation set which has not yet been used for the validation of the model
  → Mean of all mean square errors is calculated to estimate prediction error
Cross validation

- Is 10 basis functions the optimal number to use?
- Check by cross validation

→ Computation for each marine mammal
→ One profile is withdrawn (validation set), and others profiles represent training set
→ Calculation of mean square error
→ Repetition choosing another validation set which has not yet been used for the validation of the model
→ Mean of all mean square errors is calculated to estimate prediction error

- 5 basis functions are enough to minimize prediction error
Comparison of $R^2$ between the use of 5 and 10 basis functions

Calculation of $R^2$ between measured Chl-a profiles and predicted Chl-a profiles:

$$R_i^2 = \frac{||y_i - \bar{y}_i||^2 - ||\hat{y}_i - y_i||^2}{||y_i - \bar{y}_i||^2}$$
Comparison of $R^2$ between the use of 5 and 10 basis functions

Calculation of $R^2$ between measured Chl-a profiles and predicted Chl-a profiles:

\[
R^2_i = \frac{\|y_i - \bar{y}_i\|^2 - \|\hat{y}_i - y_i\|^2}{\|y_i - \bar{y}_i\|^2}
\]

<table>
<thead>
<tr>
<th>Number of basis functions</th>
<th>Elephant seal</th>
<th>Mean $R^2$</th>
<th>Median $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1st</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>0.77</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>0.70</td>
<td>0.85</td>
</tr>
<tr>
<td>10</td>
<td>1st</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>0.70</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Results

Characterization of fine scale variations (one day)

- Only one Chl-a functional profile
- 21 profiles predicted from 21 brightness functional profiles

→ Highlighting of fine-scale structures
Discussion

- Method well suited to predict Chl-a profiles
Discussion

- Method well suited to predict Chl-a profiles
- Applicable under similar conditions. Make sure that brightness profiles are recorded during day
Discussion

- Method well suited to predict Chl-a profiles
- Applicable under similar conditions. Make sure that brightness profiles are recorded during day
- Difficulty of choice of number of basis functions: cross validation seems to indicate a few number
Discussion

- Method well suited to predict Chl-a profiles
- Applicable under similar conditions. Make sure that brightness profiles are recorded during day
- Difficulty of choice of number of basis functions: cross validation seems to indicate a few number
- Chl-a data required pre-treatment (data day), this has a significant influence on the adjustment
Prospect

- Promotion of many historical records of brightness profiles over a large geographic coverage which will enable monitoring of phytoplankton production
Promotion of many historical records of brightness profiles over a large geographic coverage which will enable monitoring of phytoplankton production

Methodological development (function \textit{linmod} on R) to integrate several explanatory variables in the model

→ Interpolation using kriging
Prospect

- Promotion of many historical records of brightness profiles over a large geographic coverage which will enable monitoring of phytoplankton production
- Methodological development (function `linmod` on R) to integrate several explanatory variables in the model
- How to account for Chl-a profiles registered by night?
  → Interpolation using kriging
References

Thanks for your attention!