Spatial-functional statistics with applications to environment and health in Africa

Sophie Dabo-Niang

Professor of Applied Mathematics, University of Lille, Lab Painleve and INRIA-MODAL, Chair of EMS-CDC
sophie.dabo@univ-lille.fr

LIRIMA seminar, February 3th, 2021
FDA (Functional Data Analysis) object
Several objects = FDA sample
The raw sample is of the form:

\[ x_i(t_j), \quad t_j \in \mathcal{T}, \quad i = 1, \ldots, n \quad j = 1 \ldots, p \]

1. Transform \( x_i(t_j) \) as \( n \) observations \( \{X_i(t), t \in \mathcal{T}\} \), of a stochastic process \( \{X(t), t \in \mathcal{T}\} \):

2. \( \{X_i(t), t \in \mathcal{T}\} \), valued in a **functional space** \( (L^2(\mathcal{T}); \text{space of square integrable functions in } \mathcal{T}, \text{Hilbert, Banach space, ...}) \)

3. by a functional Basis expansion (Splines, Fourier Basis,...)
Outline

- Spatio-Functional Data
- Parametric modeling of spatial data and Application to Malaria automatic diagnostic
- Nonparametric spatial-functional estimation and Application to fisheries in West Africa
- Conclusion
Spatio-Functional Data
A random variable $X$ is called functional if it takes values in an functional space (of eventually infinite dimensional space) :

$$X = \{X_t : t \in T\}$$

- If $T \subseteq \mathbb{R}$ then $X$ is a curve
- If $T \subseteq \mathbb{R}^2$ then $X$ is an image
- If $T \subseteq \mathbb{R}^p \ (p > 2)$; $X$ is a more complexe structure

**Applications**: Environmental science, biometrics, chemometrics, mining, natural sciences, economics, medicine, ...
Application to fisheries in West African Coast

Scientific campaigns of CRODT, IRD (Senegal): mapping of fish stocks in order to avoid fishing some species, impact of climate change on stocks

- **Environmental Data**: Bottom Temperature (SBT), Surface temperature (SST), Bottom Salinity (SBS), Surface Salinity (SSS).

- **Biological Data**: different species
  - Dentex angolensis; Pagrus caeruleostictus; Galeoides decadactylus.
  - Pagellus bellottii; Octopus Vulgarus; Epinephelus-aeneus

- **Spatial Information**: latitude and longitude.
Spatial data

are collected from different spatial locations (area, country, county, town,...) on the earth, as in a variety of fields, including soil science, geology, oceanography, econometrics, epidemiology, environmental science,...

- temperature or precipitation in given locations and impact in neighborhoods
- images analysis
- disease spreading from one location to another
- migration of population from one region to another
- distribution of species in ecology
- soils properties, ...

**Spatial statistics** includes (statistical) techniques which study phenomenons observed on spatial sets: observations are **spatially** dependent.
Spatial Statistics and Africa

‘50 D. Krige (mining engineer from South Africa) : 1919-2013

Firstly mining and petroleum applications
Nowadays mostly data are spatial or spatio-temporal
Famous spatial prediction method (Kriging) by D. Krige, intensively used in Statistical and Machine Learning Theories and Applications.

‘60 Mathematical Foundation at MinesParisTech by G. Matheron (1930-2000)
Let $S$ a spatial set, $S \subset \mathbb{R}^N$, $N \geq 2$.

In particular, $N = 2$ and a location $s \in \mathbb{R}^2$ is (latitude, longitude).

Spatial data can be modeled as realizations of random fields: $Z = \{Z_s, s \in S\}$, collection of random variables indexed in $S$. 

\[ Z = \{Z_s, s \in S\} \]
Three types:

- **Lattices data**: spatial networks (neighbors locations are connected), generalization of time-series processes.
- **Point spatial processes**: Spatial location are random, generalization of time-point processes.
- **Geostatistical data**: data located at locations in $\mathbb{R}^N$, $N > 1$.

In about 20 years: most of usual statistical methods, technics, have been extended to FDA

**PCA, Change-point, Classification, Outlier detection, Regression,...**


**Review** See also the reviews on FDA of Wang et al. (2016), Martinez-Hernandez and Genton (2020)
More and More Spatial, Spatio-Functional software, scientific events, ressources

- Extensive number of ressources: published papers, scientific events, journals, societies, ...
- More than 100 free R, Python, Matlab,..., spatial, and FDA ressources
- Workshop on ”Recent FDA Advances”, AIMS, Senegal, April 2022, with most of researchers working on FDA
Parametric modeling of spatial data and Application to Malaria automatic diagnostic
Let $Z = \{ Z_s \in \mathbb{R}, s \in \mathbb{R}^N \}$ be a real *gaussian* spatial process that we suppose *second order or intrinsic stationary*. 

$Z = (Z_s, s \in \mathbb{R}^N)$ is observed on $n$ sites $s_1, \ldots, s_n$.

- Prediction of $Z_{s_0}$ (or $g(Z_{s_0})$) using $Z_{s_1}, Z_{s_2}, \ldots, Z_{s_n}$

- With the help of a covariance (variogram) model to account the spatial dependency
Second order stationary is not always satisfied.

One way to establish the second order stationarity hypothesis is to consider the $h-$ increments process from $Z$: intrinsic process

**Definition**

- $Z$ is an **intrinsic** process, if for every $h$:
  - $Z$ is of constant mean
  - the increment process $I^{(h)} = \left\{ I_s^{(h)} = Z_{s+h} - Z_s \right\}$ is a second order stationary process in $s$:
    
    $$
    E(I_s^{(h)}) = 0 \quad \text{Var}(Z_{s+h} - Z_s) = 2\gamma(h),
    $$

where $2\gamma$ is called **variogram** function.
Variogram

Figure 1 – Experimental and theoretical variograms. The nugget is the variance between very close observations. The range indicates the distance after which observations are independent.
Prediction (Kriging) $Z_{s_0}$ is to find the best linear predictor (BLUP) of $Z_{s_0}$ based on observations $Z = (Z_{s_1}, \ldots, Z_{s_n})$ and a variogram model $\gamma(.)$:

$$
\hat{Z}_{s_0} = \sum_{i=1}^{n} \lambda_i Z_{s_i}
$$

$$
E(\hat{Z}_{s_0}) = E(Z_{s_0}) \text{ and } E(\hat{Z}_{s_0} - Z_{s_0})^2 \text{ minimum}
$$
Isotropic admissible variograms

Examples of isotropic variograms associated with a covariance

\[ C(h) = \sigma^2 - \gamma(h), \text{ where } \sigma^2 \text{ is the associated stationary variance.} \]

- **nugget** \( \gamma(h) = C \)
- **exponential** \( \gamma(h) = C(1 - \exp(-\|h\|/a)) \quad C > 0 \quad a > 0 \)
- **spherical** \( \gamma(h) = \begin{cases} 
C\left(\frac{3\|h\|}{2a} + \frac{1}{2}\frac{\|h\|^3}{a^3}\right) & \text{if } 0 \leq \|h\| \leq a \\
C & \text{if } \|h\| \geq a 
\end{cases} \)
- **gaussian** \( \gamma(h) = C(1 - \exp(-\|h\|^2/a)) \quad C > 0 \quad a > 0 \)
- **power** \( \gamma(h) = C\|h\|^\alpha, \quad \alpha < 2 \)
Detection and segmentation of erythrocytes in multispectral label-free blood smear images for automatic cell counting: (Doumoun, Dabo-Niang and Zoueu; 2020)

- Kriging+ Sampling: Reduction of 66% of acquisition time

Figure 2 – Segmentation process illustrated with a 435 nm image; (a) sample image, (b) adaptive thresholding, (c) post-processing and (d) counting
Mapping malaria risk of exposure in Ivory Coast

Spatio-temporal malaria risk of exposure mapping and rapid diagnosis of malaria (Doumoun, Dabo-Niang and Zoueu; 2020)

- Population dynamics of anopheles mosquitoes: innovative entomological lidar + bistatic optical sensors + trapping methodologies to study population dynamics of anopheles mosquitoes.

- Spatial Information: different urban and rural areas in Ivory Coast.

- Climat and environment: weather data

- Spatio-temporal big data from physical traps, optical sensors, weather stations
Nonparametric spatial-functional estimation and Application to fisheries in West Africa
Non-parametric spatial prediction for real data

Preliminary step: estimation of a regression function $r(\cdot)$

$$Y = r(X) + \epsilon$$

$Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, $\epsilon$, the error term.

→ with the help of spatial observations $(Z_{s_i}, s_i \in \mathbb{Z}^N), N \geq 1,$

$Z_{s_i} = (X_{s_i}, Y_{s_i})$, same law as $(X, Y)$. 
Let the spatial process \((Z_{s_i} = (X_{s_i}, Y_{s_i}) \in \mathbb{R}^d \times \mathbb{R}, s_i \in \mathbb{Z}^N)\), observed on some discret set

\[ \mathcal{I}_n = \{s_i \in \mathbb{N}^N, 1 \leq s_k \leq n_k, k = 1, \ldots, N\} \]

- the \(Z_{s_i}\) have unknown density, \(f_{X,Y}\) w.r.t to Lebesgue measure on \(\mathbb{R}^{d+1}\);
- \(X_{s_i}\) of density \(f\);
- the regression function is

\[
r(x) = \begin{cases} \varphi(x)/f(x) & \text{if } f(x) \neq 0; \\ \mathbb{E}(Y) & \text{otherwise,} \end{cases}
\]

where \(\varphi(x) = \int_\mathbb{R} yf_{X,Y}(x, y)dy\) with \(y \in \mathbb{R}\) and \(x \in \mathbb{R}^d\).
Classical Kernel estimation of $f(\cdot)$ and $r(\cdot)$

With observations $\{(X_{s_i}, Y_{s_i}), s_i \in \mathcal{I}_n\}$, we have for $x \in \mathbb{R}^d$ :

$$r_n(x) = \begin{cases} 
\varphi_n(x)/f_n(x) & \text{if } f_n(x) \neq 0; \\
\frac{1}{\hat{n}} \sum_{s_i \in \mathcal{I}_n} Y_{s_i} & \text{otherwise}, 
\end{cases}$$

with

$$\varphi_n(x) = \frac{1}{\hat{n}h^d_n} \sum_{s_k \in \mathcal{I}_n} Y_{s_k} K\left(\frac{x - X_{s_k}}{h_n}\right)$$

and

$$f_n(x) = \frac{1}{\hat{n}h^d_n} \sum_{s_k \in \mathcal{I}_n} K\left(\frac{x - X_{s_k}}{h_n}\right)$$

- $K$ a kernel on $\mathbb{R}^d$, with compact support, valued in $\mathbb{R}$
- $\lim_{n \to \infty} h_n = 0$
- $\hat{n} := n_1 \times \ldots \times n_N$ (sample size)
Drawback of this kernel regression estimator

- Do not take into account the spatial information in the data
- Strict stationarity assumption
- Estimate the regression function in the same as if the processes were independent
More flexible estimates

Based on product of 2 kernels:

one on observations
the other on the locations (sites)

Kernel Spatial prediction and local stationarity

- Let \( \{Z_{s_i} = (X_{s_i}, Y_{s_i}) \in \mathbb{R}^d \times \mathbb{R}, s_i \in \mathbb{N}^N\} \) be a spatial process observable in

\[ \mathcal{I}_n = \{ s_i \in \mathbb{N}^N : 1 \leq s_r \leq n_r \text{ for } r = 1, \ldots, N \} \]

- Regression model:

\[ Y_{s_i} = r(X_{s_i}) + \varepsilon_{s_i}, \quad s_i \in \mathbb{N}^N \quad \text{with} \quad r(\cdot) = E(Y_{s_i} | X_{s_i} = \cdot) \]

\( \{\varepsilon_{s_i}\} \) is centred, \( \alpha \)-mixing process

\( \rightarrow \) Goal: prediction of \( \{Y_{s_i}, s_i \in \mathbb{N}^N\} \) at an unobserved site \( s_0 \in \mathcal{I}_n \), using:

- \( X_{s_0} \) and
- \( (X_{s_i}, Y_{s_i})_{s_i \in \mathcal{O}_n}, \) with \( s_0 \notin \mathcal{O}_n \subset \mathcal{I}_n \)

under the assumption: \( (X_{s_i}, Y_{s_i})_{s_i \in \mathcal{O}_n} \) are locally identically distributed around \( s_0 \) (Klemelä, 2008) instead of the classical stationarity condition.
Let $\tilde{s}_j = \frac{s_j}{n}$ (→ normalized).

For fixed $X_{s_j} = x \in \mathbb{R}^d$, observed at $s_j$ (denoted $x_{s_j}$):

$$f_n(x_{s_j}) = \frac{1}{\hat{n} b_n^d \rho_n^N} \sum_{s_i \in \mathcal{I}_n} K_1 \left( \frac{x_{s_j} - X_{s_i}}{b_n} \right) K_2 \left( \frac{\|\tilde{s}_j - \tilde{s}_i\|}{\rho_n} \right)$$

where

- $K_1$ et $K_2$: kernels on $\mathbb{R}^d$ and $\mathbb{R}$, valued in $\mathbb{R}$

- $\lim_{n \to \infty} b_n = 0$ and $\lim_{n \to \infty} \rho_n = 0$

- $n = (n_1, \ldots, n_N) \to \infty$ means that $\min_i n_i \to \infty$, $i = 1, \ldots, N$
Estimator of $r(\cdot)$

$$r_n(x_{s_j}) = \begin{cases} 
\varphi_n(x_{s_j}) / f_n(x_{s_j}) & \text{if } f_n(x_{s_j}) \neq 0; \\
\frac{1}{\hat{n}} \sum_{s_i \in \mathcal{I}_n} Y_{s_i} & \text{else,}
\end{cases}$$

with

$$\varphi_n(x_{s_j}) = \frac{1}{\hat{n} b_n^d \rho_n^N} \sum_{s_i \in \mathcal{I}_n} Y_{s_i} K_1 \left( \frac{x_{s_j} - X_{s_i}}{b_n} \right) K_2 \left( \frac{\|\tilde{s}_j - \tilde{s}_i\|}{\rho_n} \right)$$
Kernel spatial predictor

Let $s_0$ a location where $Y_{s_0}$ is not available, the kernel prediction is

$$
\hat{Y}_{s_0}^{NW} = \frac{\sum_{s_i \in O_n} Y_{s_i} K_1 \left( \frac{X_{s_0} - X_{s_i}}{h_n} \right) K_2 \left( \rho_n^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)}{\sum_{s_i \in O_n} K_1 \left( \frac{X_{s_0} - X_{s_i}}{h_n} \right) K_2 \left( \rho_n^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)},
$$

- $K_1$ and $K_2$ are two kernels and $\frac{s_i}{n} = \left( \frac{s_{i,1}}{n_1}, \ldots, \frac{s_{i,N}}{n_N} \right)$.
- $h_n$, $\rho_n$ are non random bandwidths.

$\hookrightarrow$ The fact that $h_n$ is deterministic, does not allow to capture some heterogeneity.
Explore $k$–NN kernel method (Collomb, 1980) to deal with an eventual heterogeneity:

$$
\hat{Y}_{s_0} = \frac{\sum_{s_i \in O_n} Y_{s_i} K_1 \left( \frac{X_{s_0} - X_{s_i}}{H_{n,X_{s_0}}} \right) K_2 \left( h_{n,s_0}^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)}{\sum_{s_i \in O_n} K_1 \left( \frac{X_{s_0} - X_{s_i}}{H_{n,X_{s_0}}} \right) K_2 \left( h_{n,s_0}^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)},
$$

$H_{n,X_{s_0}}$ : random bandwidth on the covariate:

$$
H_{n,X_{s_0}} = \min \left\{ h \in \mathbb{R}^+ \mid \sum_{s_i \in O_n} \mathbb{I}(\|X_{s_i} - X_{s_0}\| < h) = k(n) \right\}
$$

$h_{n,s_0}$ : bandwidth on locations:

$$
h_{n,s_0} = \min \left\{ h \in \mathbb{R}^+ \mid \sum_{s_i \in O_n} \mathbb{I} \left( \left\| \frac{s_i - s_0}{n} \right\| < h \right) = k_{1_n} \right\}
$$

The smoothing parameters $k_{1_n}$ and $k(n)$ are positive integers.
Data from coastal demersal surveys off Senegalese coasts performed by the scientific team of the CRODT in cold and hot seasons, in North, Center and South areas, from 2001 to 2015.

→ At 495 stations:

- **Environmental data**: sea bottom temperature (SBT), sea surface temperature (SST), sea bottom salinity (SBS) and sea surface salinity (SSS).

- **Spatial information**: latitude and longitude.

- **Biological data**: three species
  - **Dentex**: Dentex angolensis
  - **Pagrus**: Pagrus caeruleostictus
  - **Thiekem**: Galeoides decadactylus

→ **Goal**: predict the presence of species
The following methods are compared:

<table>
<thead>
<tr>
<th>Method</th>
<th>Covariates</th>
<th>Spatial param</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic k-NN SVM</td>
<td>SBT, SST, SBS, SSS, <em>latitude</em> and <em>longitude</em></td>
<td>No</td>
</tr>
<tr>
<td>Logit model</td>
<td>SBT, SBS, <em>latitude</em> and <em>longitude</em></td>
<td>No</td>
</tr>
<tr>
<td>Kernel k-NN</td>
<td>SBT, SST, SBS, SSS</td>
<td>Yes</td>
</tr>
</tbody>
</table>

For the three species:

- Dataset is split into: training and testing samples, with respective sizes equal to 80% and 20% of the whole sample size.
- Tuning parameters of each method are selected through cross validation method or AIC (Logit model).
- Correct classification rate (CCR) is computed for each method.
<table>
<thead>
<tr>
<th>Specie</th>
<th>Method</th>
<th>CCR for :</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Observed</td>
<td>Non observed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.776 0.811 0.733</td>
<td>0.816 0.887 0.733</td>
<td>0.714 0.774 0.644</td>
<td></td>
</tr>
<tr>
<td><strong>Dentex</strong></td>
<td>Basic k-NN</td>
<td>0.776</td>
<td>0.811</td>
<td>0.733</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.816</td>
<td><strong>0.887</strong></td>
<td>0.733</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Logit model</td>
<td>0.714</td>
<td>0.774</td>
<td>0.644</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kernel</td>
<td>0.745</td>
<td>0.736</td>
<td>0.756</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Kernel k-NN</strong></td>
<td><strong>0.847</strong></td>
<td>0.868</td>
<td><strong>0.822</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Pagrus</strong></td>
<td>Basic k-NN</td>
<td>0.735</td>
<td>0.516</td>
<td>0.836</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.786</td>
<td>0.452</td>
<td>0.940</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Logit model</td>
<td>0.704</td>
<td>0.258</td>
<td>0.910</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kernel</td>
<td><strong>0.827</strong></td>
<td><strong>0.581</strong></td>
<td><strong>0.940</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Kernel k-NN</strong></td>
<td>0.796</td>
<td>0.581</td>
<td>0.896</td>
<td></td>
</tr>
<tr>
<td><strong>Thiekem</strong></td>
<td>Basic k-NN</td>
<td>0.857</td>
<td>0.400</td>
<td>0.974</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.908</td>
<td>0.750</td>
<td>0.950</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Logit model</td>
<td>0.878</td>
<td>0.750</td>
<td>0.910</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kernel</td>
<td>0.898</td>
<td>0.700</td>
<td>0.949</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Kernel k-NN</strong></td>
<td><strong>0.959</strong></td>
<td><strong>0.850</strong></td>
<td><strong>0.987</strong></td>
<td></td>
</tr>
</tbody>
</table>
Application to prediction

Prediction of SBT, SST, SBS and SSS in several positions:

Sea bottom temperature

Sea surface temperature

Sea bottom salinity

Sea surface salinity
Use the previous predictions of SBT, SST, SBS and SSS to predict the probability to observe each specie:

**Dentex**

**Pagrus**

**Thiekem**
Spatio-functional kernel regression and prediction

- \( X = (X_s(.), \ s \in \mathbb{R}^N) \), a measurable spatial process \( N \geq 1 \), defined on some probability space \( (\Omega, \ A, \mathbb{P}) \)

- \( X_s \) is valued in a space \( (\mathcal{E}, \ d) \) of eventually infinite dimension

- \( d(., .) \) is some measure of proximity, e.g. a metric or a semi-metric

- \( \mathcal{E} \) is a space of functions, typically the space of squared integrable functions defined on some finite interval \( \mathcal{T} = [0, T] \).

- \( X \) is observed at a set of locations \( \mathcal{I} \subseteq \mathbb{R}^N \) of cardinal \( n \), \( \mathcal{I} = \{s_1, \ldots, s_n\}, \ s_i \in \mathbb{R}^N, \ i = 1 \ldots n \) and a set of depth points \( \mathcal{J} = \{t_1, \ldots, t_p\} \)

- In practice, we have a set of \( n \times d \) discrete observations, \( S = \{x_{s_i}(t_j), \ s_i \in \mathcal{I}, t_j \in \mathcal{J}\} \).
Smoothing the data

- **First FDA step**: the discrete data \( \{x_{si}(t_j), s_i \in \mathcal{I}, t_j \in \mathcal{J} \} \) are converted into curves \( \{X_{si}(t), s_i \in \mathcal{I}, t \in \mathcal{T} \} \) by using smoothing methods (e.g. Splines).

- \( \{X_{si}(t), s_i \in \mathcal{I}, t \in \mathcal{T} \} \) are valued in \( \mathcal{E} \) of infinite dimension.

- **Reduce the infinite dimension**: expand each \( X_{si}(.) \) in terms of basis functions (here Fourier basis).

- The number of basis functions chosen by using cross-validation.
Temperature and Pression curves
Spatio-functional kernel regression

Kernel-based procedure for estimating the nonlinear functional regression model:

\[ Y_{si} = r(X_{si}) + \epsilon_{si}, \quad i = 1, \ldots, n \]

- \( Y_{si} \): scalar response \( \in \mathbb{R} \)
- \( X_{si} \): functional predictor (e.g. curves) \( \in (\mathcal{E}, d(., .)) \) where \( d(., .) \) is a semi-metric
  \( \alpha \)-mixing process (general case of weakly dependence)
- \( \{\epsilon_s\} \): stationary residual process
  independent of the process \( \{X_s\} \)
Kernel regression for functional data

Let \( X_{s_i} \in (E, d) \) (dimension \( \infty \)) and \( Y_{s_i} \in \mathbb{R} \), the kernel estimator:

\[
\hat{r}(x) = \frac{\sum_{i=1}^{n} Y_{s_i} K \left( \frac{d(x, X_{s_i})}{h} \right)}{\sum_{j=1}^{n} K \left( \frac{d(x, X_{s_j})}{h} \right)}, \quad x \in (E, d)
\]

using observations \((X_{s_i}, Y_{s_i})_{i=1,...,n}\), where \( d(\cdot, \cdot) \) is a semi-metric.

→ Investigated first by Ferraty and Vieu (2000) and then

for independent data by e.g. Ferraty and Vieu (2002);
Dabo-Niang and Rhomari (2003); . . .

for dependent data by e.g. Masry (2005); Ternynck (2014); ,
Account the spatial feature

Use the spatial locations of the sample to improve estimation of $r(\cdot)$ to a prediction goal:

$$\hat{Y}_{s_0} = \frac{\sum_{s_i \in \mathcal{O}_n} Y_{s_i} K_1 \left( \frac{d(x, X_{s_i})}{b_n} \right) K_2 \left( \rho_n^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)}{\sum_{s_i \in \mathcal{O}_n} K_1 \left( \frac{d(x, X_{s_i})}{b_n} \right) K_2 \left( \rho_n^{-1} \left\| \frac{s_0 - s_i}{n} \right\| \right)}, \quad (1)$$
Application to fish species: *Dentex angolensis*

### Table 1 – ER and CCR with functional covariate salinity

<table>
<thead>
<tr>
<th>Functional Methods</th>
<th>ER</th>
<th>CCR → Y=1,0</th>
<th>TCC → Y=1</th>
<th>TCC → Y=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFC</td>
<td>0.34</td>
<td>0.44</td>
<td>0.83</td>
<td>0.37</td>
</tr>
<tr>
<td>FGLM</td>
<td>0.44</td>
<td>0.59</td>
<td><strong>0.98</strong></td>
<td>0.14</td>
</tr>
<tr>
<td>FGSAM</td>
<td>0.36</td>
<td>0.64</td>
<td>0.90</td>
<td>0.34</td>
</tr>
<tr>
<td>KNNF</td>
<td>0.43</td>
<td>0.53</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$DD^G$</td>
<td>0.27</td>
<td>0.57</td>
<td>0.75</td>
<td>0.37</td>
</tr>
<tr>
<td>MFSD</td>
<td><strong>0.09</strong></td>
<td><strong>0.75</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.70</strong></td>
</tr>
</tbody>
</table>
Another alternative: reduce the dimension

- Expansion of $X(.)$ using a set of basis functions (for instance PCA basis) $\phi_k(\cdot), \ k = 1, \ldots, K_x$.

- Observed curve $X_{s_i}$ from $X(.)$ is expressed

$$X_{s_i}(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t)$$

- Do kernel regression estimate with the coefficients $c_{ik}$ as covariates

- Functional kernel problem is then reduced to regress a finite number of coefficients $c_{ik}$
Other ongoing projects in Africa


- Sea Surface temperature Analysis: Spatio-functional Model Based-Clustering (Dabo-Niang et al. (2021)). F. Adjibode, M. Ndiaye, ..., Benin


Other ongoing projects in Africa: Cancer cell characterisation

**Rapid Cancer detection**: by FDA modeling and Bio-MEMS

C. Agonkoui, **Benin**

---

**Fig. 1: Overview of HiPCIC.**

- **a)** WP1: MEMS device for high-throughput single-cell characterization using autonomous adaptive sensing.
- **b)** WP2: Multi-parameter phenotyping to classify biophysical cell properties according to known metastatic potential of cancer cell lines.
- **c)** WP3: Predicting the metastatic potential (low, mid & high) of cancer cell lines & patient samples.
Other ongoing projects in Africa: Cancer cell characterisation
Hot topics to investigate

- Non-parametric space-time modeling
- Multivariate Spatio-functional data modeling
- Spatial Survival nonparametric analysis
- Spatial nonparametric analysis of extreme data
- See the recent reviews on Spatial FDA:
  - Martinez-Hernandez and Genton (2020)
  - Mateu and Giraldo (2021)
- ...
Thank you for your attention