Statistical methods of data science

An introduction to Functional Data Analysis

Alessandra Menafoglio[1*]

[1]MOX, Department of Mathematics, Politecnico di Milano

*alessandra.menafoglio@polimi.it

6. Spatial Statistics for functional data

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

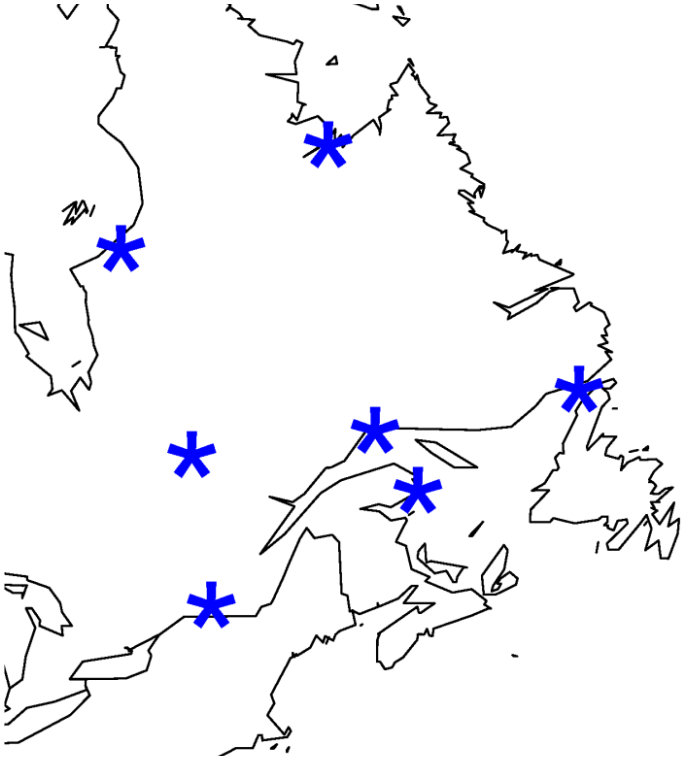
6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Premiss: object data with spatial dependence

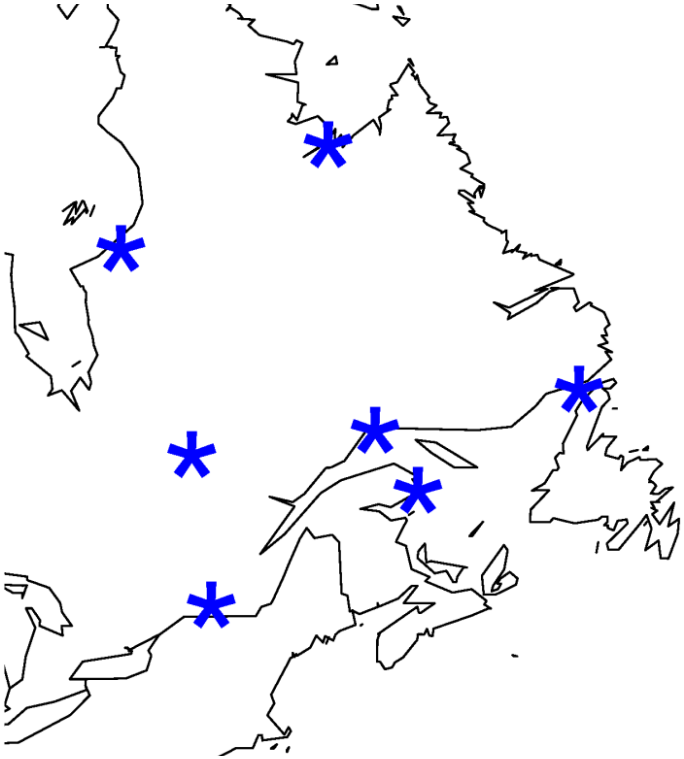
Spatially dependent data



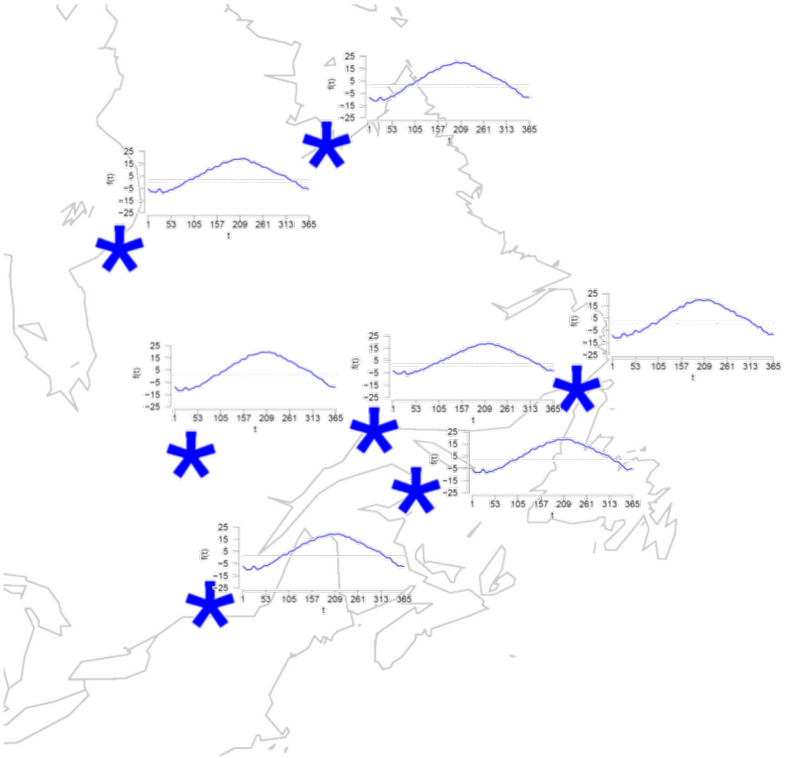
Spatial location of meteorological stations in Quebec (Canada)

Premiss: object data with spatial dependence

Spatially dependent functional data



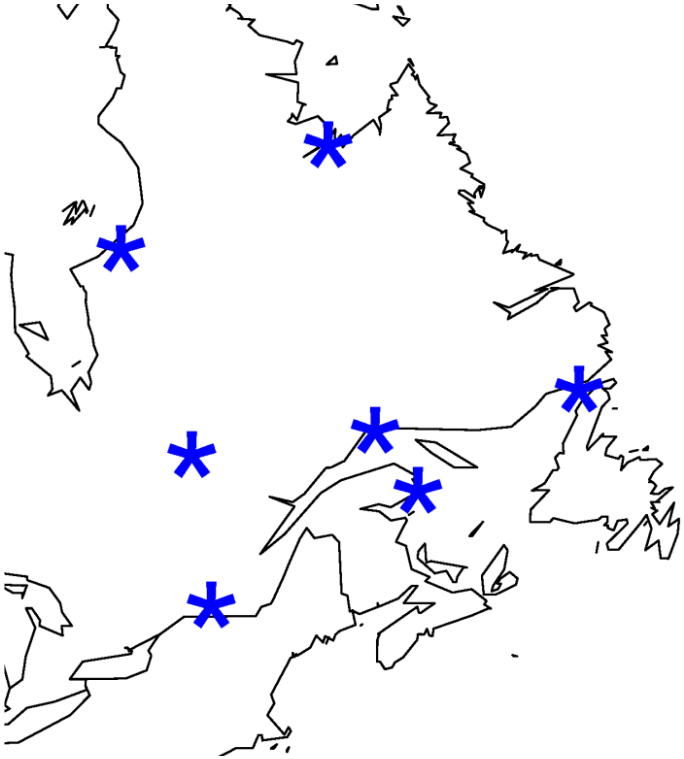
Spatial location of meteorological stations in Quebec (Canada)



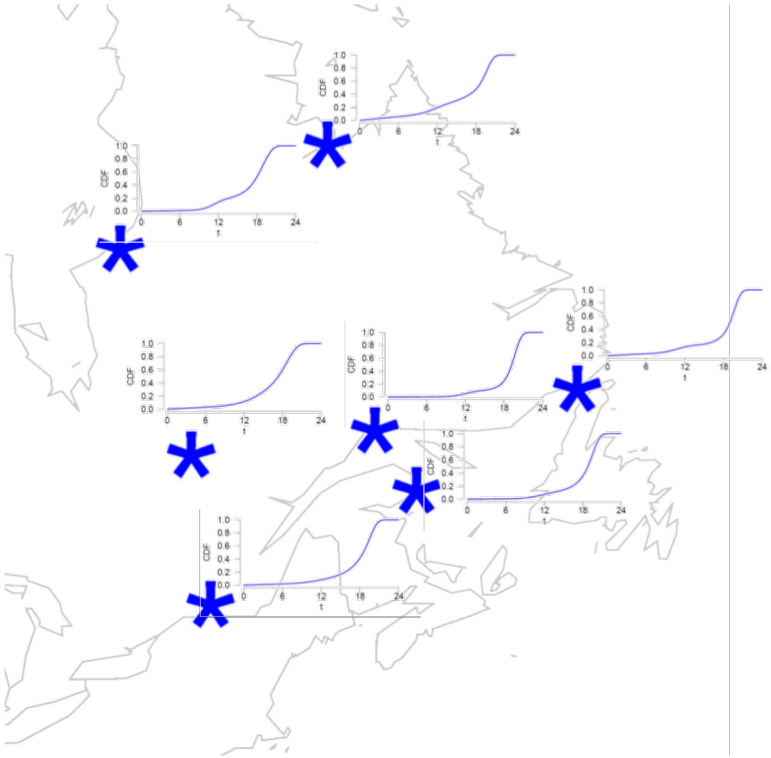
Daily mean temperatures along one year

Premiss: object data with spatial dependence

Spatially dependent distributional data



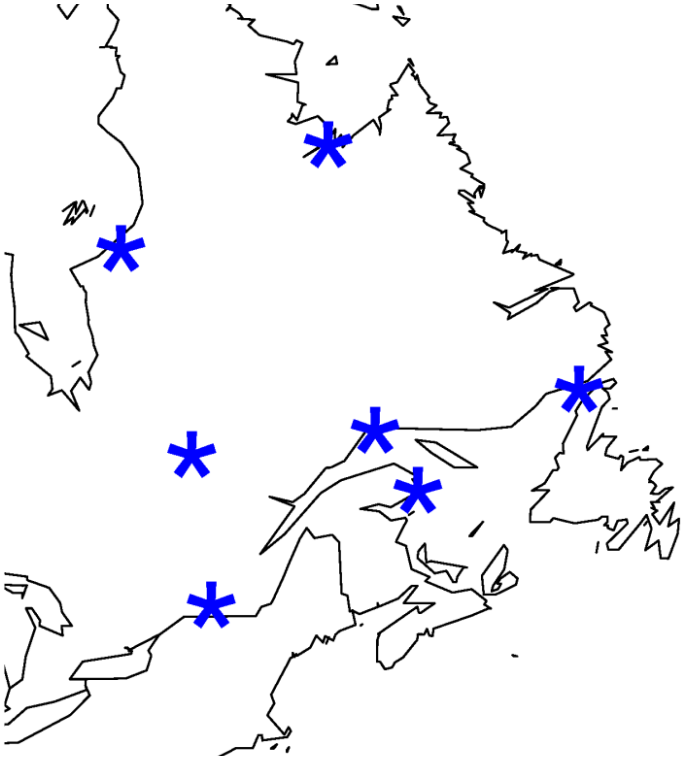
Spatial location of meteorological stations in Quebec (Canada)



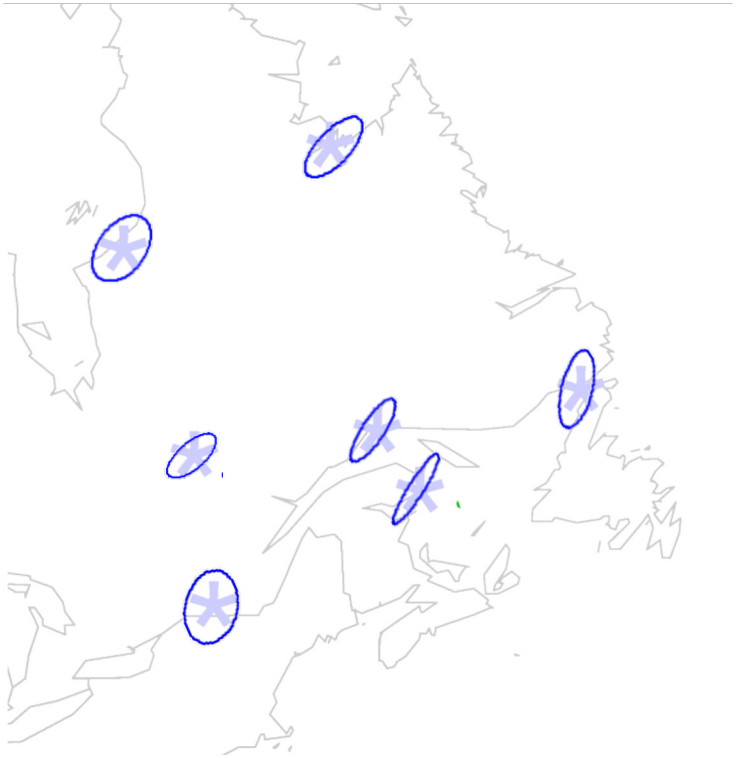
SCS precipitation distributions

Premiss: object data with spatial dependence

Spatially dependent manifold data

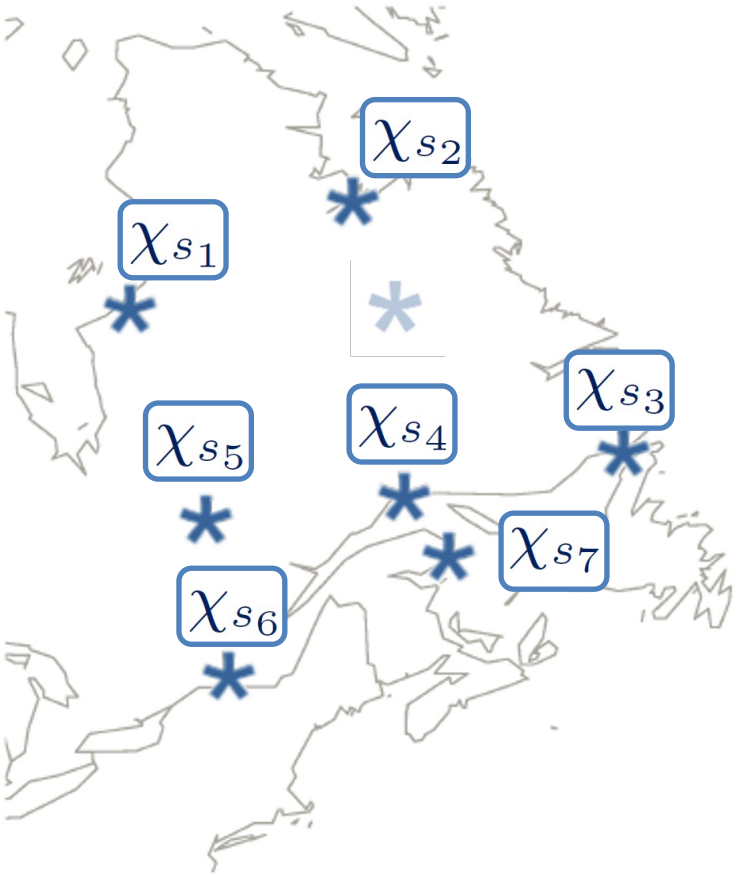


Spatial location of meteorological stations in Quebec (Canada)



S=Cov(Temperature, Precipitation) covariance matrices in January

Spatially dependent object data



Aim:

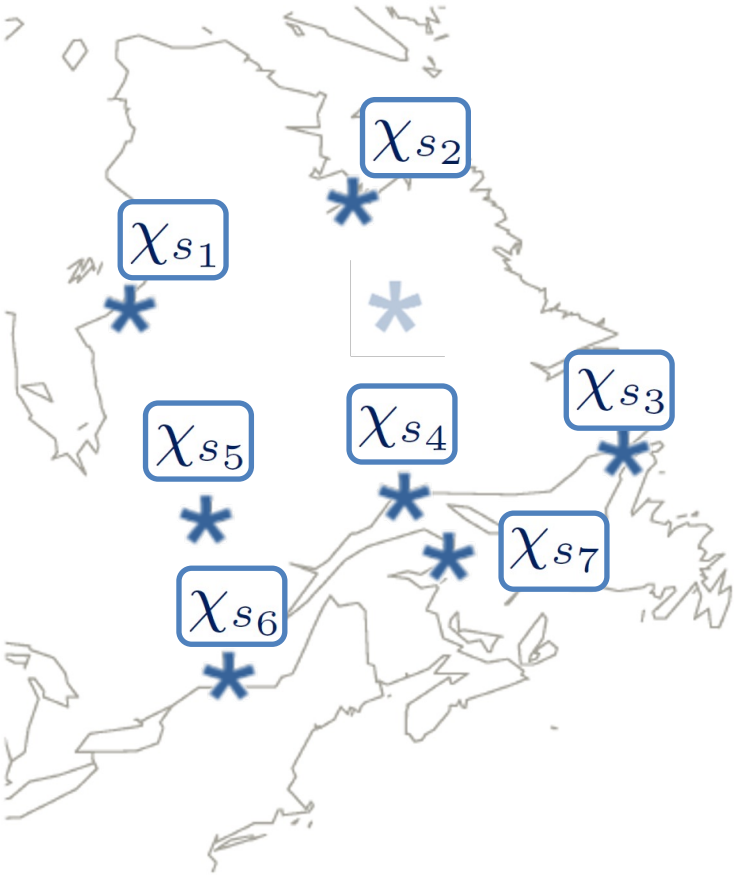
1) Estimate the **structure of spatial dependence**

First Law of Geography

"*Everything is related to everything else, but near things are more related than distant things*."

Waldo Tobler (1970)

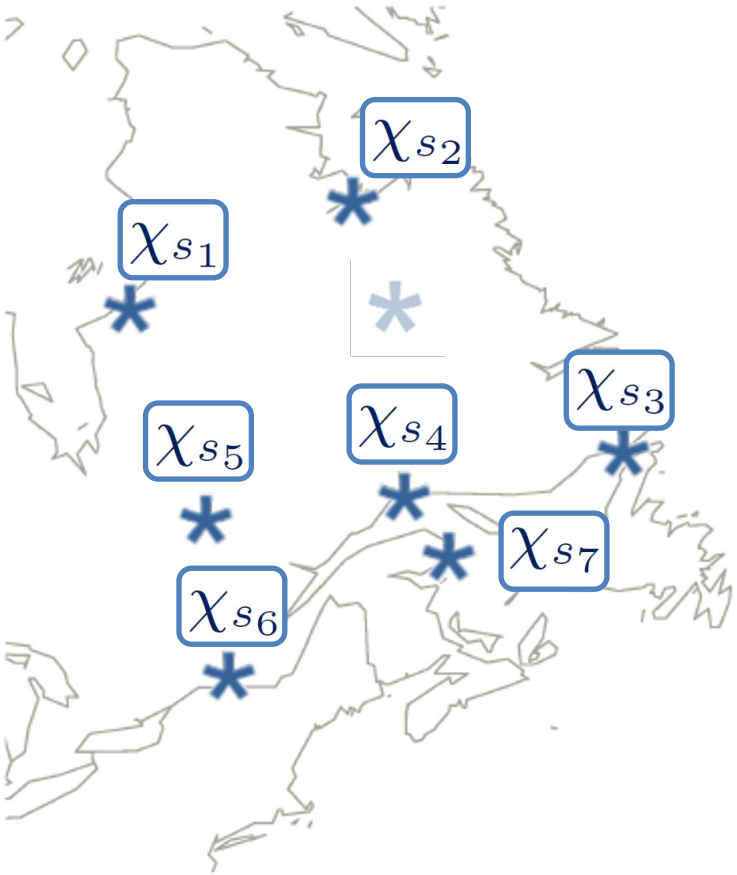
Spatially dependent object data



Aim:

1) Estimate the **structure of spatial dependence**
2) **Prediction** of an observation at an unsampled location s_0

Spatially dependent object data

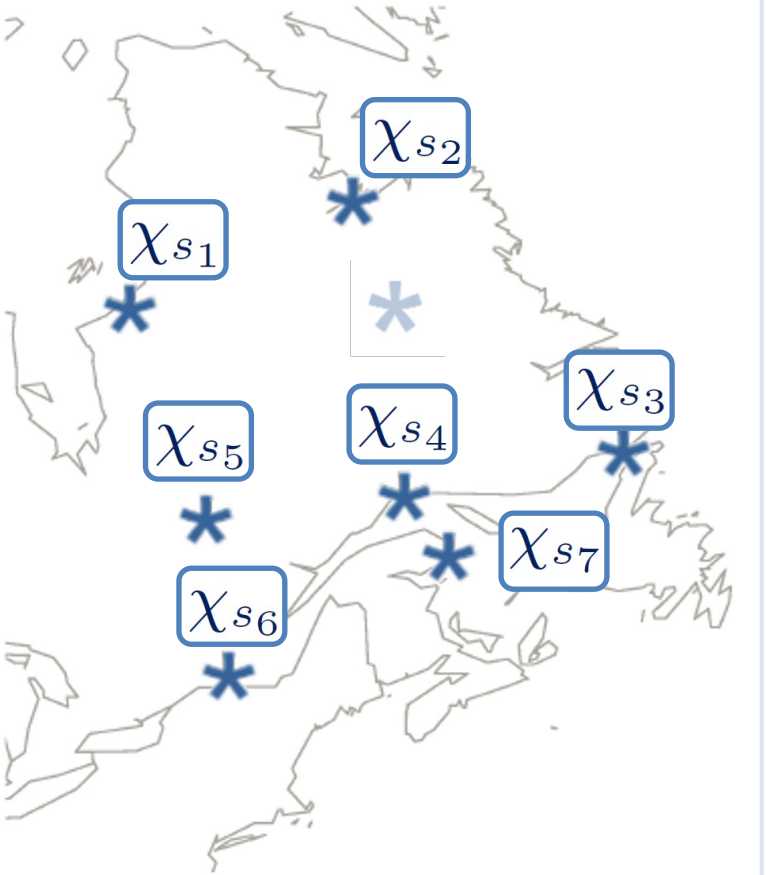


Aim:

1) Estimate the **structure of spatial dependence**
2) **Prediction** of an observation at an unsampled location
3) Uncertainty quantification

Geostatistics for Object data

Spatially dependent object data



Aim:

1) Estimate the **structure of spatial dependence**
2) **Prediction** of an observation at an unsampled location
3) Uncertainty quantification

Methods: Object Oriented Geostatistics

a) **Object-oriented** approach:

$$\chi_{s_1}, \ldots, \chi_{s_n} \in H$$

Space *H* should capture the **"appropriate geometry"** for the data objects (e.g., a Hilbert space)

b) **Geostatistical** approach (**kriging**) to treat the **spatial dependence**

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Definitions and assumptions

Data

$s_1, \ldots, s_n$ spatial locations in $D \subseteq \mathbb{R}^d$, *d=2,3*

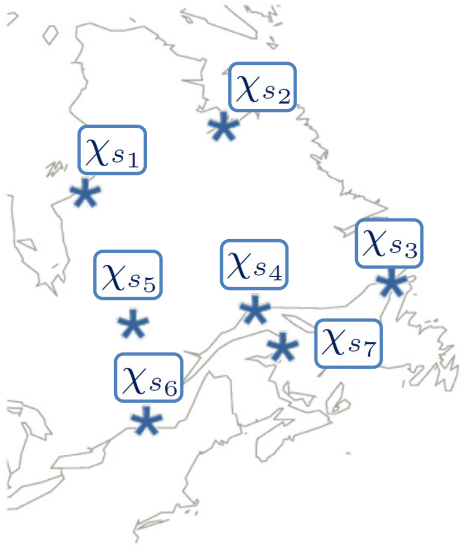
$\chi_{s_1}, \cdots, \chi_{s_n}$ scalar observations at the sampling locations

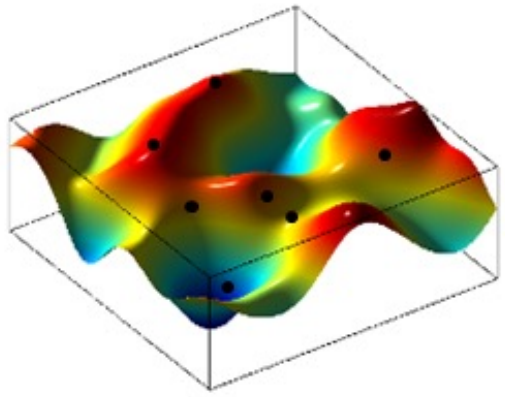
Model

Observations are assumed to be a partial observation of a random field $\{\chi_s, s \in D\}$

Assumptions

- $\mathbb{E}[\chi_s] < \infty$ for all $s \in D$
- $Var(\chi_s) < \infty$ for all $s \in D$
- *Stationarity*





Stationarity

- Assumption of homogeneity in space
- Different degrees of stationarity:
  - Strong stationarity: assumption on all the finite-dimensional laws
  - Second-order stationarity: assumption on the first two moments
  - Intrinsic stationarity: assumption on the first moment, and weak assumption on the second moment

- We focus on second-order stationarity

Second Order Stationarity

Process $\{\chi_s, s \in D\}$ is second-order stationary if:

i. $\mathbb{E}[\chi_{\boldsymbol{s}}] = m$ for all $s \in D$

ii. $Cov(\chi_{s_1}, \chi_{s_2}) = C(s_1 - s_2)$ for all $s_1, s_2 \in D$

Isotropy

- Besides assuming stationarity, one can also assume isotropy

Isotropy

A second-order stationary field $\{\chi_s, s \in D\}$ is isotropic if

$$Cov(\chi_{\boldsymbol{s}_1}, \chi_{\boldsymbol{s}_2}) = C(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$$

for all $s_1, s_2 \in D$.

- Isotropy is an assumption of directional homogeneity
- Isotropy can be relaxed in most methods, but it greatly simplifies the notation and the estimation procedures

The covariogram

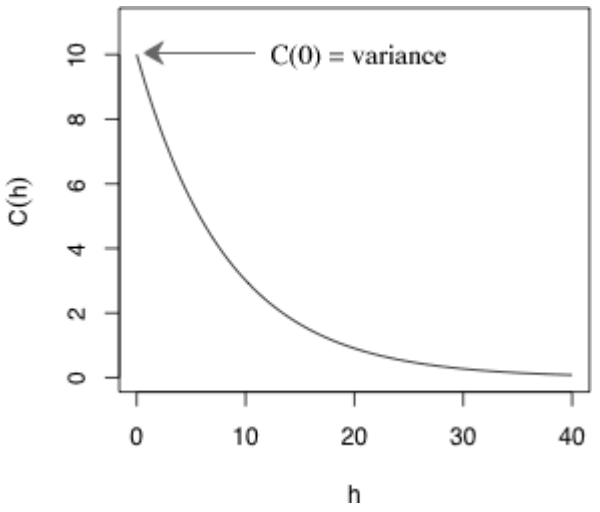
Function C in the definition of second-order stationarity is called **covariogram**

- It describes the second-order properties of the field, i.e., the spatial dependence
- Properties of the covariogram:
  - Positive semi-definite function

$$\sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i - \boldsymbol{s}_j) \geq 0,$$
$$\forall\ \lambda_i, \lambda_j \in \mathbb{R};\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D;$$

  - Symmetry w.r.t. the vector **0**
  - Boundedness $|C(\boldsymbol{h})| \leq C(\mathbf{0})$

Under isotropy



First Law of Geography
"*Everything is related to everything else, but near things are more related than distant things.*"

Waldo Tobler (1970)

The variogram

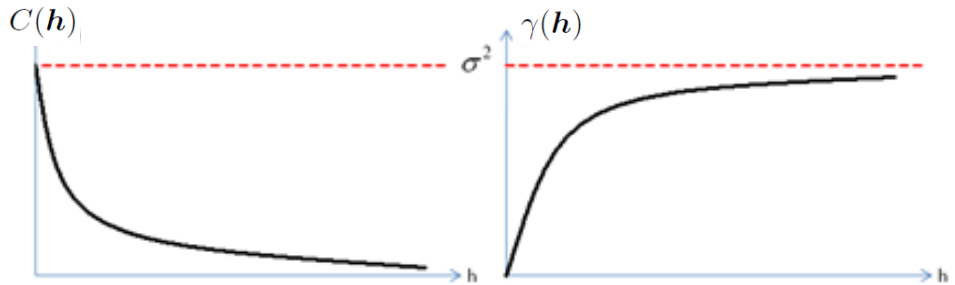
- The covariogram is closely related with an alternative (more general) measure of spatial dependence, the **variogram**
- Under second-order stationarity, the variogram is defined as the function

$$2\gamma(\boldsymbol{h}) = Var(\chi_{\boldsymbol{s}+\boldsymbol{h}} - \chi_{\boldsymbol{s}}) = \mathbb{E}[(\chi_{\boldsymbol{s}+\boldsymbol{h}} - \chi_{\boldsymbol{s}})^2]$$

- Function $\gamma(\boldsymbol{h})$ is called semivariogram
- The covariogram and the variogram are related via the identity

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}), \quad \boldsymbol{h} \in \mathbb{R}^d.$$

Under isotropy



Covariogram Semivariogram

The variogram

- Properties of the variogram $2\gamma(\boldsymbol{h}) = Var(\chi_{\boldsymbol{s}+\boldsymbol{h}} - \chi_{\boldsymbol{s}}) = \mathbb{E}[(\chi_{\boldsymbol{s}+\boldsymbol{h}} - \chi_{\boldsymbol{s}})^2]$
    1. Conditionally negative semi-definite function

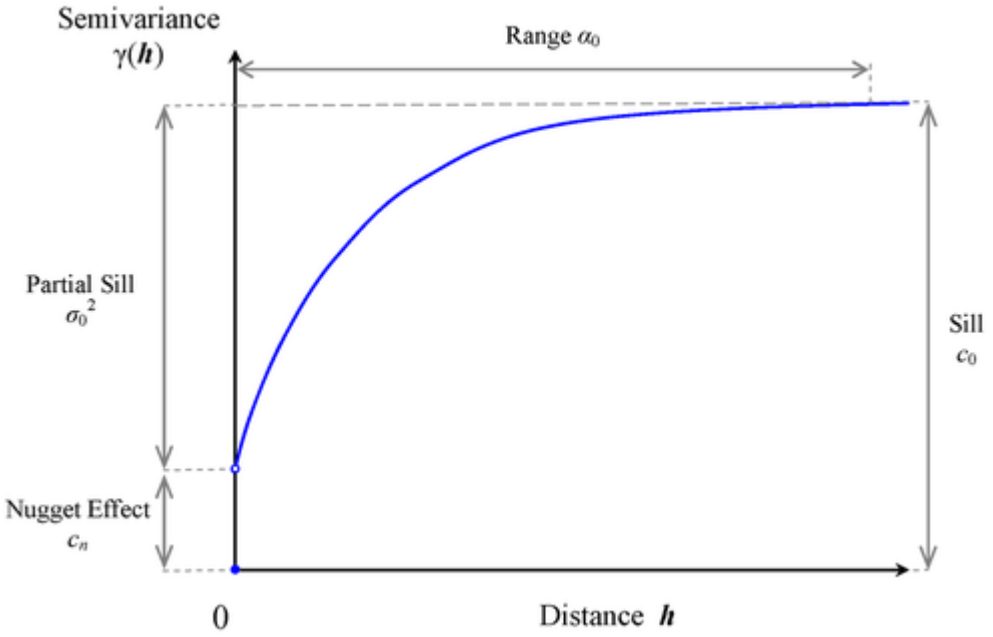
$$\sum_i \sum_j \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0, \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j, \in D,$$
$$\forall\, \lambda_i, \lambda_j \quad \text{s.t.} \quad \sum_i \lambda_i = 0;$$

    2. Symmetry w.r.t. the vector **0**
    3. Non-negativity
    4. Zero at the origin
    5. Sub-quadratic growth
    6. …
- A real-valued function fulfilling all the properties of a variogram is called «valid model»

Structural properties of the variogram

Structural Properties

- Sill/partial sill: related with the variance of the process
- Range: related with the degree of dependence (~range of influence)
- Nugget: related with possible process discontinuities or measurement error



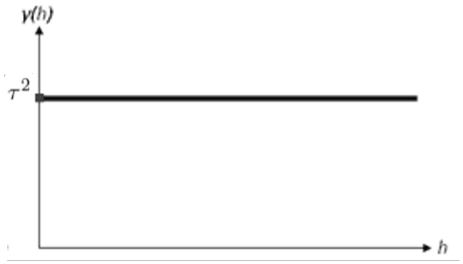
Examples of isotropic valid models

Examples: Valid Models

- *Pure Nugget* (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(h) = \begin{cases} \tau^2, & h > 0, \\ 0, & h = 0, \end{cases}$$

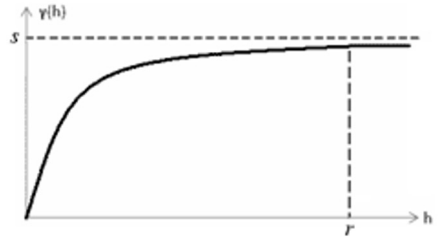
with $\tau \in \mathbb{R}$.



- *Exponential model* (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-h/a}), & h > 0, \\ 0, & h = 0, \end{cases}$$

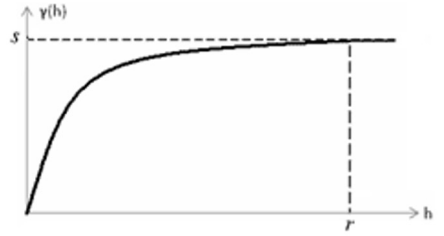
where $a, \sigma \in \mathbb{R}$.



- *Spherical model* (valid in $\mathbb{R}^d$, $d = 1, 2, 3$):

$$\gamma(h) = \begin{cases} 0, & h = 0, \\ \sigma^2\{\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3\}, & 0 < h < a, , \\ \sigma^2, & h \geq a, \end{cases}$$

with $a, \sigma \in \mathbb{R}$.



Examples of isotropic valid models

Valid models can be used as building blocks for more complex structures

(I) Sum of valid models: if γ_1 and γ_2 are valid models, $\gamma = \{[0,\infty) \ni h \mapsto \gamma_1(h) + \gamma_2(h)\}$ is a valid model;

(II) Product by a constant: if γ_1 is a valid model and $c \in \mathbb{R}^+$ is a real positive constant, $\gamma = \{[0,+\infty) \ni h \mapsto c \cdot \gamma_1(h)\}$ is a valid model.

Structural properties of the variogram

Another important feature: behavior near zero

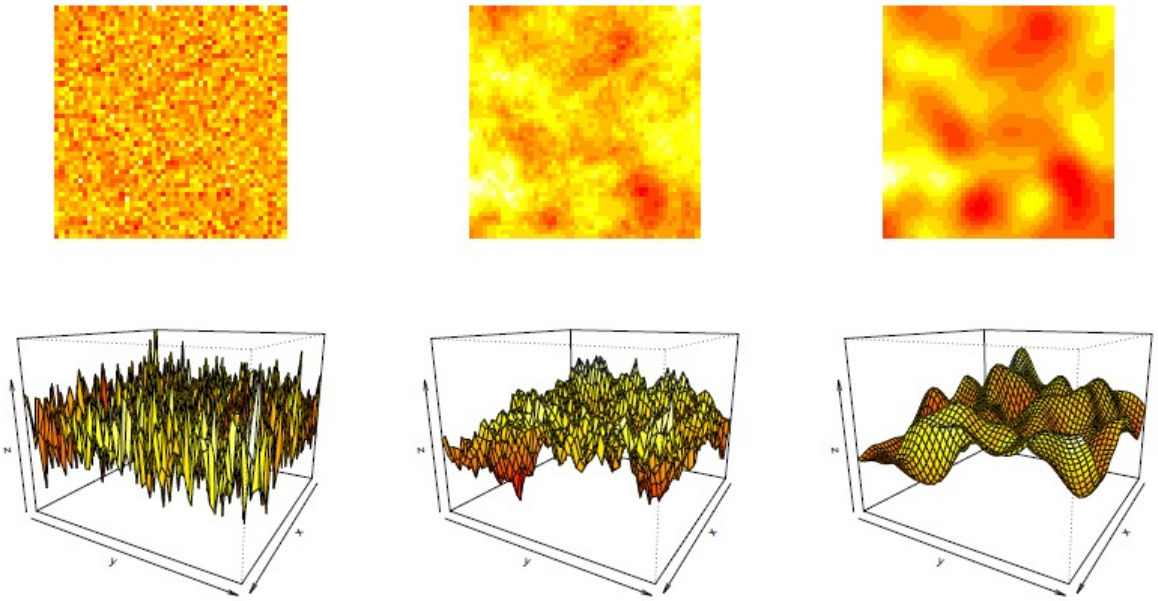


Figure 1: Example of realization from pure nugget (left panels), spherical (central panels; range: 0.3) and Gneiting (right panels; range: 0.3) models, with the same sill.

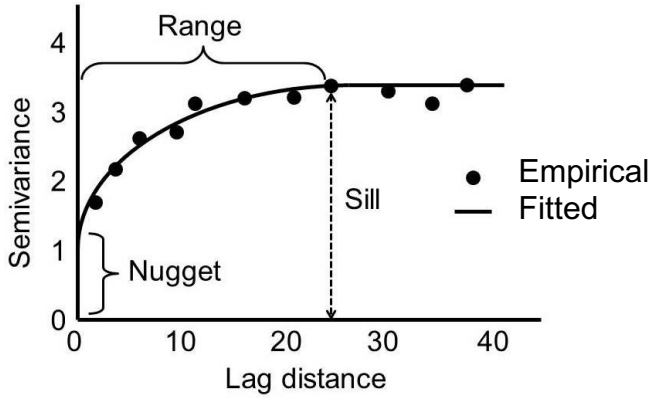
Estimating the spatial dependence

- To estimate the spatial dependence, the variogram is preferred to the covariogram
- Under stationarity, the estimation of the variogram is usually done in two steps
  - Empirical estimate or raw estimate

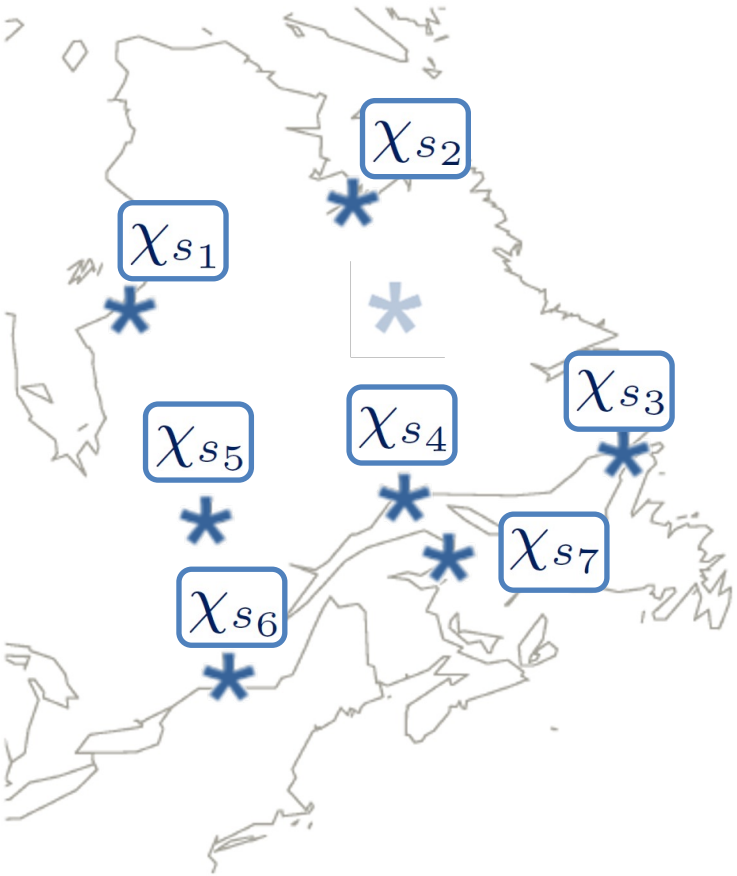
$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} [\chi_{s_i} - \chi_{s_j}]^2$$

with $N(h) = \{(i,j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| = h\}$

  - Fit of a valid model, e.g., via least squares



Spatial prediction



- **Goal:** predict the value of the field at an unsampled location
- **Methods:** Kriging
  - Best linear unbiased predictor from the data

$$\chi^*_{\boldsymbol{s}_0} = \sum_{i=1}^{n} \lambda^*_i \chi_{\boldsymbol{s}_i}$$

  - The optimal weights minimize the variance of prediction error under unbiasedness constraint
- According to the assumptions required, we distinguish: simple, ordinary or universal kriging

Universal Kriging

Goal: Prediction of a generic X_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$X^*_{s_0} = \sum_{i=1}^{n} \lambda^*_i X_{s_i}$$

Assumptions

- $\{X_s, s \in D\}$ non-stationary random field of the form:

$$X_s = \underbrace{m_s}_{\text{deterministic drift}} + \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$

- **Linear model** for the drift

$$m_s = \sum_{l=0}^{L} a_l f_l(s), \quad s \in D.$$

- **Second-order stationary** residual

Theorem

The optimal weights $\lambda^*_1, ..., \lambda^*_n \in \mathbb{R}$ are found by solving

$$\left(\begin{array}{c|c} \gamma(\boldsymbol{h}_{i,j}) & f_l(\boldsymbol{s}_i) \\ \hline f_l(\boldsymbol{s}_j) & 0 \end{array}\right) \left(\begin{array}{c} \lambda_i \\ \hline \zeta_l \end{array}\right) = \left(\begin{array}{c} \gamma(\boldsymbol{h}_{0,i}) \\ \hline f_l(\boldsymbol{s}_0) \end{array}\right)$$

Remarks

- The theory assumes the variogram to be known
- If it isn't, one estimates the variogram and plug it in the linear system
- In case of stationarity, the variogram can be estimated directly from the data
- For non-stationary fields, the variogram needs to be estimated from (estimated) residuals

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Recall: The Hilbert space model in FDA

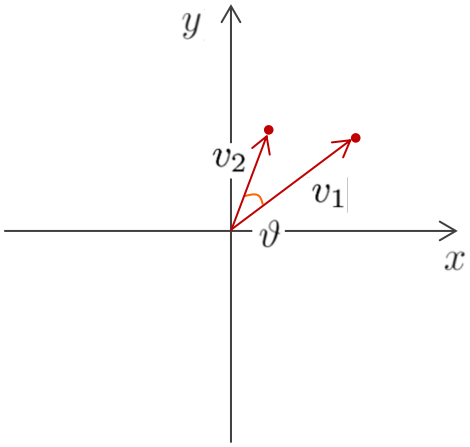
Main idea:
The data objects are considered as points within a Hilbert space

Recall: The Hilbert space model in FDA

Main idea:
The data objects are considered as points within a Hilbert space

If the data were vectors of two dimensions…

Euclidean space $\mathbb{R}^2$



Space whose points are two-dimensional objects (vectors)

- Sum: $v_1 + v_2 = (x_1 + x_2, y_1 + y_2)$
- Product by a constant: $c \cdot v = (c \cdot x, c \cdot y)$

} Operations (+, ·)

- Norm (lenght of a vector): $\|v\| = (x^2 + y^2)^{1/2}$
- Distance: $\|v_1 - v_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$
- Angle: $\vartheta = \arccos \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$

} Inner product $\langle v_1, v_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$

Recall: The Hilbert space model in FDA

Main idea:
The data objects are considered as points within a Hilbert space

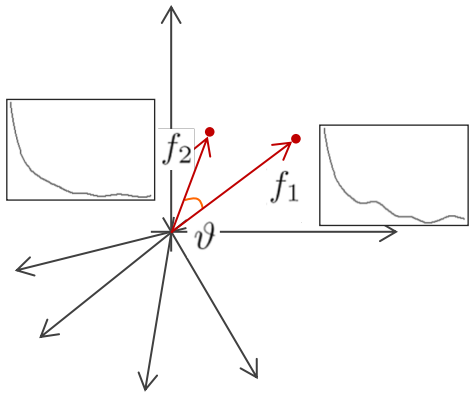
If the data were vectors of two dimensions…
… for complex data of any (even infinite) dimension…

Recall: The Hilbert space model in FDA

Main idea:
The data objects are considered as points within a Hilbert space

If the data were vectors of two dimensions…
… for complex data of any (even infinite) dimension…

Euclidean space $\mathbb{R}^2$



Space whose points are objects of any dimension (e.g., functions)

- Sum: $f_1 \oplus f_2$
- Product by a constant: $c \odot f$

} Operations ($\oplus$,$\odot$)

- Norm: $\|f\|$
- Distance: $\|f_1 \ominus f_2\|$
- Angle: $\vartheta = \arccos \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$

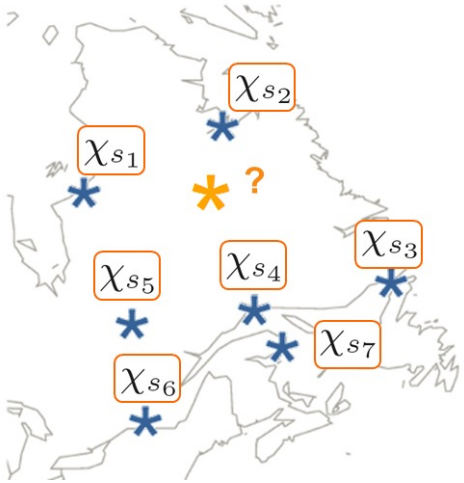
} Inner product $\langle f_1, f_2 \rangle$

A Hilbert space approach to the analysis of spatial data

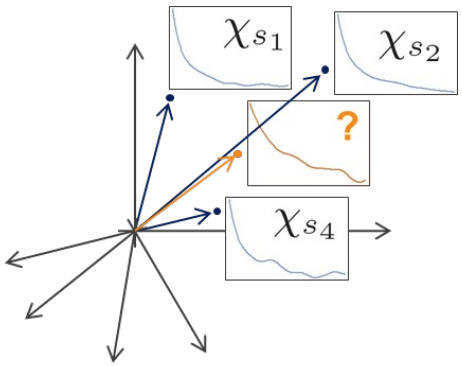
Main idea:
The data objects are considered as points within a Hilbert space

Strategy

Spatially dependent object data



Points of a Hilbert space



Geometry of the space

- Measure the **spatial dependence**: *similarities* between objects (distances)
- **Prediction** at unsampled locations through *linear combinations* of the data

A Hilbert space approach to the analysis of spatial data

Classical Geostatistics framework

- $\{\chi_s : s \in D\}$: stochastic process indexed by a **continuous spatial variable** $s \in D \subset \mathbb{R}^d$
- $\chi_{s_1}, \ldots, \chi_{s_n}$: **observations** of the process in *n* **sampled locations**

Object-Oriented approach

- $\overline{\chi_s}$: random element of an **infinite-dimensional** separable Hilbert space *H* $(\oplus, \odot, \langle \cdot, \cdot \rangle, \|\cdot\|)$
- $\chi_{s_1}, \ldots, \chi_{s_n}$: set of **object data**

Menafoglio, Secchi, Dalla Rosa (2013)

A Hilbert space approach to the analysis of spatial data

Classical Geostatistics framework	Object-Oriented approach
• $\{\chi_s : s \in D\}$: stochastic process indexed by a **continuous spatial variable** $s \in D \subset \mathbb{R}^d$	• $\overline{\chi_s}$: random element of an **infinite-dimensional** separable Hilbert space H $(\oplus, \odot, \langle \cdot, \cdot \rangle, \|\cdot\|)$
• $\chi_{s_1}, \ldots, \chi_{s_n}$: **observations** of the process in n **sampled locations**	• $\chi_{s_1}, \ldots, \chi_{s_n}$: set of **object data**

Strategy:

Extend classical geostatistics methods (e.g., variogram modeling, kriging) to object data, by analogy with the classical case, using the geometry of H

Menafoglio, Secchi, Dalla Rosa (2013)

A Hilbert space approach to the analysis of spatial data

Menafoglio, Secchi, Dalla Rosa (2013)

Classical Geostatistics framework	Object-Oriented approach
• $\{\chi_s : s \in D\}$: stochastic process indexed by a **continuous spatial variable** $s \in D \subset \mathbb{R}^d$	• χ_s: random element of an **infinite-dimensional** separable Hilbert space H $(\oplus, \odot, \langle\cdot,\cdot\rangle, \|\cdot\|)$
• $\chi_{s_1}, \ldots, \chi_{s_n}$: **observations** of the process in n **sampled locations**	• $\chi_{s_1}, \ldots, \chi_{s_n}$: set of **object data**

Strategy:

Extend classical geostatistics methods (e.g., variogram modeling, kriging) to object data, by analogy with the classical case, using the geometry of H

$\{\chi_s : s \in D\}$ ***real***-valued process:

- Spatial mean: $m_s = \mathbb{E}[\chi_s],$
- Covariogram: $$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Cov}_{\mathbb{R}}(\chi_{\boldsymbol{s}_i}, \chi_{\boldsymbol{s}_j}) = \\ = \mathbb{E}[(\chi_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_i})(\chi_{\boldsymbol{s}_j} - m_{\boldsymbol{s}_j})]$$
- Variogram $$2\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Var}_{\mathbb{R}}(\chi_{\boldsymbol{s}_i} - \chi_{\boldsymbol{s}_j}) = \\ = \mathbb{E}[(\chi_{\boldsymbol{s}_i} - \chi_{\boldsymbol{s}_j})^2] - (m_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_j})^2$$

A Hilbert space approach to the analysis of spatial data

Menafoglio, Secchi, Dalla Rosa (2013)

Classical Geostatistics framework	Object-Oriented approach
• $\{\boldsymbol{\chi}_s : s \in D\}$: stochastic process indexed by a **continuous spatial variable** $s \in D \subset \mathbb{R}^d$	• $\boldsymbol{\chi}_s$: random element of an **infinite-dimensional** separable Hilbert space H $(\oplus, \odot, \langle\cdot,\cdot\rangle, \|\cdot\|)$
• $\boldsymbol{\chi}_{s_1}, \ldots, \boldsymbol{\chi}_{s_n}$: **observations** of the process in n **sampled locations**	• $\boldsymbol{\chi}_{s_1}, \ldots, \boldsymbol{\chi}_{s_n}$: set of **object data**

Strategy:

Extend classical geostatistics methods (e.g., variogram modeling, kriging) to object data, by analogy with the classical case, using the geometry of H

$\{\boldsymbol{\chi}_s : s \in D\}$ ***Hilbert space***-valued process:

- Spatial mean (Bochner integral): $m_s = \mathbb{E}[\boldsymbol{\chi}_s]$,
- **Trace**-covariogram:

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Cov}_H(\boldsymbol{\chi}_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{\boldsymbol{s}_j}) = \\ = \mathbb{E}[\langle \boldsymbol{\chi}_{\boldsymbol{s}_i} \ominus m_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{\boldsymbol{s}_j} \ominus m_{\boldsymbol{s}_j} \rangle]$$

- **Trace**-variogram

$$2\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Var}_H(\boldsymbol{\chi}_{\boldsymbol{s}_i} \ominus \boldsymbol{\chi}_{\boldsymbol{s}_j}) = \\ = \mathbb{E}[\|\boldsymbol{\chi}_{\boldsymbol{s}_i} \ominus \boldsymbol{\chi}_{\boldsymbol{s}_j}\|^2] - \|m_{\boldsymbol{s}_i} \ominus m_{\boldsymbol{s}_j}\|^2$$

About the parameter estimation

Menafoglio, Secchi, Dalla Rosa (2013)

Second order stationarity

Process $\{\chi_s, s \in D\}$ is second-order stationary if:

i. $\mathbb{E}[\chi_{\boldsymbol{s}}] = m$ for all $s \in D$

ii. $Cov(\chi_{s_1}, \chi_{s_2}) = C(s_1 - s_2)$ for all $s_1, s_2 \in D$

Variogram estimate (under stationarity)

Classical Geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[(\chi_{\boldsymbol{s}} - \chi_{\boldsymbol{s}+\boldsymbol{h}})^2]$

1. Empirical estimate (MoM)

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} (\chi_{\boldsymbol{s}_i} - \chi_{\boldsymbol{s}_j})^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn

Object-Oriented geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[\|\chi_{\boldsymbol{s}} \ominus \chi_{\boldsymbol{s}+\boldsymbol{h}}\|^2]$

1. Empirical estimate (MoM)*

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} \|\chi_{\boldsymbol{s}_i} \ominus \chi_{\boldsymbol{s}_j}\|^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn

(Almost) the same modeling effort as in 1D geostatistics!

About the parameter estimation

Menafoglio, Secchi, Dalla Rosa (2013)

Second order stationarity

Process $\{\chi_s, s \in D\}$ is second-order stationary if:

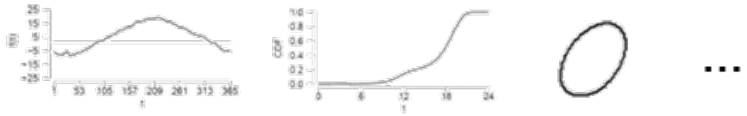
i. $\mathbb{E}[\chi_s] =$ [illegible]

ii. $Cov(\chi_s$ [illegible]

Key remark:

Data need to be embedded in a **meaningful** Hilbert space

Data Object χ_s



Variogram es [illegible]

Cla [illegible] tatistics

Variogram:

1. Empiri [illegible]

$\widehat{\gamma}(h) =$ [illegible] $\|\delta_{s_i} \ominus \delta_{s_j}\|^2$

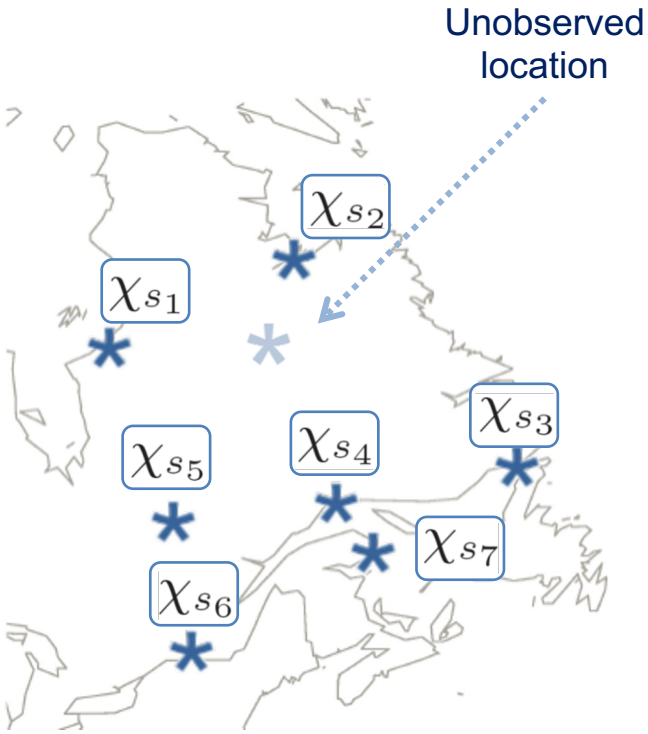
$(i,j) \in N(h)$ $(i,j) \in N(h)$

2. Fit of a parametric model (LS) e.g., spherical, matérn

2. Fit of a parametric model (LS) e.g., spherical, matérn

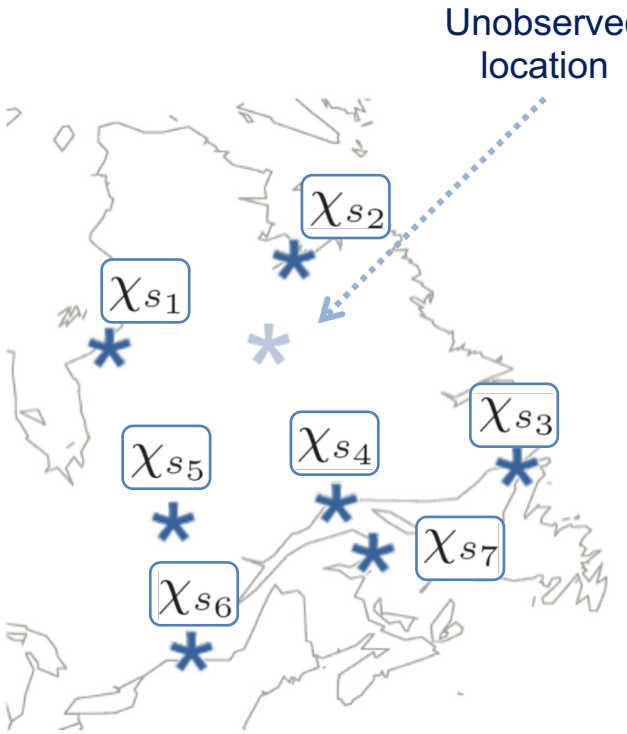
(Almost) the same modeling effort as in 1D geostatistics!

Functional Kriging in Hilbert spaces



Menafoglio, Secchi, Dalla Rosa (2013)

Functional Kriging in Hilbert spaces

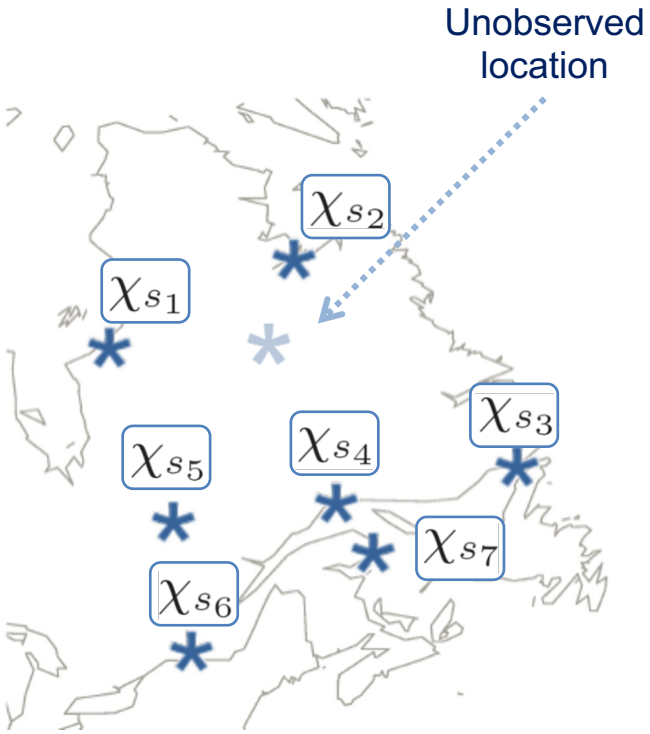


Goal: Prediction of a generic χ_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$\chi^*_{s_0} = \bigoplus_{i=1}^{n} \lambda^*_i \odot \chi_{s_i}$$

Menafoglio, Secchi, Dalla Rosa (2013)

Functional Kriging in Hilbert spaces



Goal: Prediction of a generic χ_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$\chi^*_{s_0} = \bigoplus_{i=1}^{n} \lambda_i^* \odot \chi_{s_i}$$

Assumptions

- $\{\chi_s : s \in D\}$ non-stationary random field of the form:

$$\chi_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$

- **Linear model** for the drift

$$m_s = \bigoplus_{l=0}^{L} (f_l(s) \odot a_l), \quad s \in D$$

- **Globally second-order stationary** residual (trace-variogram depending only on the increment vector between locations)

Menafoglio, Secchi, Dalla Rosa (2013)

Functional Kriging in Hilbert spaces

Goal: Prediction of a generic X_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$X^*_{s_0} = \sum_{i=1}^{n} \lambda^*_i X_{s_i}$$

Assumptions

- $\{X_s, s \in D\}$ non-stationary random field of the form:

$$X_s = \underbrace{m_s}_{\text{deterministic drift}} + \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$

- **Linear model** for the drift

$$m_s = \sum_{l=0}^{L} a_l \, f_l(s), \quad s \in D.$$

- **Second-order stationary** residual

Goal: Prediction of a generic χ_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$\chi^*_{s_0} = \bigoplus_{i=1}^{n} \lambda^*_i \odot \chi_{s_i}$$

Assumptions

- $\{\chi_s : s \in D\}$ non-stationary random field of the form:

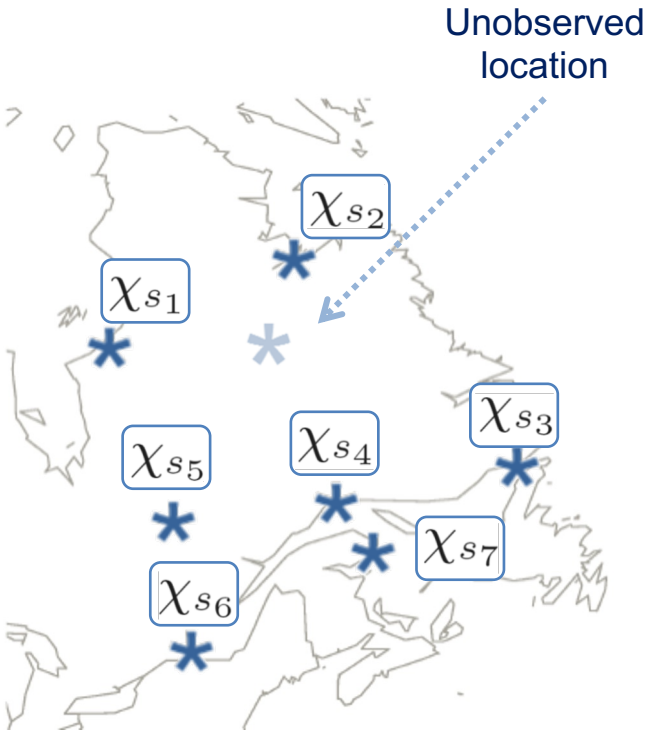
$$\chi_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$

- **Linear model** for the drift

$$m_s = \bigoplus_{l=0}^{L} (f_l(s) \odot a_l), \quad s \in D$$

- **Globally second-order stationary** residual (trace-variogram depending only on the increment vector between locations)

Functional Kriging in Hilbert spaces



Goal: Prediction of a generic χ_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$\chi^*_{s_0} = \bigoplus_{i=1}^{n} \lambda^*_i \odot \chi_{s_i}$$

Formulation

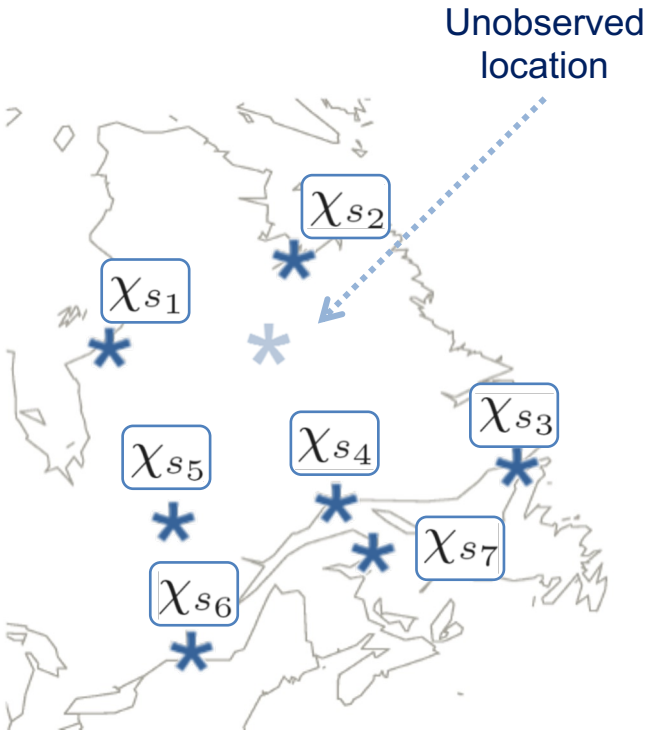
- Find $\lambda^*_1, \ldots, \lambda^*_n \in \mathbb{R}$ that solve

$$\min_{\lambda_1,\ldots,\lambda_n} \mathrm{Var}_H(\chi^*_{s_0} \ominus \chi_{s_0}),$$

$$\text{s. t.} \quad \mathbb{E}[\chi^*_{s_0} \ominus \chi_{s_0}] = 0.$$

Menafoglio, Secchi, Dalla Rosa (2013)

Functional Kriging in Hilbert spaces



Goal: Prediction of a generic χ_{s_0} at site s_0 through the **Universal Kriging Predictor** (BLUP):

$$\chi^*_{s_0} = \bigoplus_{i=1}^{n} \lambda^*_i \odot \chi_{s_i}$$

Theorem

The optimal weights $\lambda^*_1, \ldots, \lambda^*_n \in \mathbb{R}$ are found by solving

$$\left(\begin{array}{c|c} \gamma(\boldsymbol{h}_{i,j}) & f_l(\boldsymbol{s}_i) \\ \hline f_l(\boldsymbol{s}_j) & 0 \end{array}\right)\left(\begin{array}{c} \lambda_i \\ \hline \zeta_l \end{array}\right) = \left(\begin{array}{c} \gamma(\boldsymbol{h}_{0,i}) \\ \hline f_l(\boldsymbol{s}_0) \end{array}\right)$$

Same as in 1D geostatistics!

Remark: the results generalizes that of Giraldo et al. (2011), Nerini et al (2008) [stationarity, $H=L^2$]

Menafoglio, Secchi, Dalla Rosa (2013)

About the parameter estimation

Classical Geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[(\delta_{\boldsymbol{s}} - \delta_{\boldsymbol{s}+\boldsymbol{h}})^2]$

1. Empirical estimate (MoM)

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} (\delta_{\boldsymbol{s}_i} - \delta_{\boldsymbol{s}_j})^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn

Object-Oriented geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[\| \delta_{\boldsymbol{s}} \ominus \delta_{\boldsymbol{s}+\boldsymbol{h}} \|^2]$

1. Empirical estimate (MoM)

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} \| \delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j} \|^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn

Residual variogram and drift jointly estimated via an iterative algorithm (GLS)

(Almost) the same modeling effort as in 1D geostatistics!

About the parameter estimation

Classical Geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[(\delta_{\boldsymbol{s}} - \delta_{\boldsymbol{s}+\boldsymbol{h}})^2]$

1. Empirical estimate (MoM)

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} (\delta_{\boldsymbol{s}_i} - \delta_{\boldsymbol{s}_j})^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn

Object-Oriented geostatistics

Variogram: $2\gamma(\boldsymbol{h}) = \mathbb{E}[\| \delta_{\boldsymbol{s}} \ominus \delta_{\boldsymbol{s}+\boldsymbol{h}} \|^2]$

1. Empirical estimate (MoM)

$$\widehat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(i,j)\in N(h)} \| \delta_{\boldsymbol{s}_i} \ominus \delta_{\boldsymbol{s}_j} \|^2$$

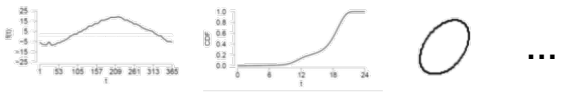
2. Fit of a parametric model (LS) e.g., spherical, matérn

Residual variogram and drift jointly estimated via an iterative algorithm (GLS)

(Almost) the same modeling effort as in 1D geostatistics!

Key remark: data needs to be embedded in a **meaningful** Hilbert space

Data Object χ_s



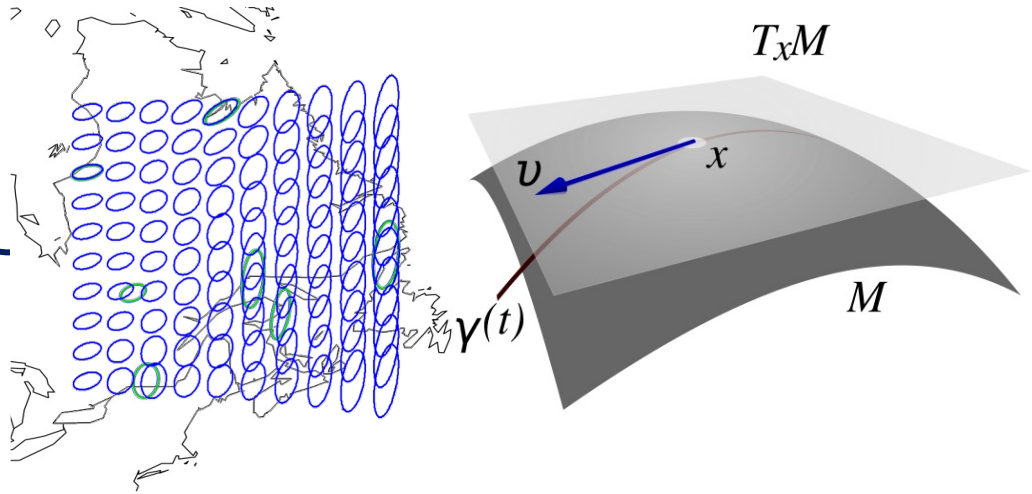
The advantages of a general approach: the application point of view

A new **general** and **coherent** framework for the treatment of **non-stationary** functional random fields valued in **any** separable **Hilbert Space**

General model

- Dichotomy model for $\{\chi_s : s \in D\}$, χ_s in H

$$\chi_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$



Pigoli, Menafoglio, Secchi (2013)

H separable Hilbert space:

- can account for **differential** information (Sobolev spaces)
- can account for data **constraints** (Aitchison geometry)
- can locally approximate a **Riemannian manifold**

Menafoglio, Secchi, Dalla Rosa (2013)
Menafoglio, Guadagnini, Secchi (2014,2016)

The advantages of a general approach: the application point of view

A new **general** and **coherent** framework for the treatment of **non-stationary** functional random fields valued in **any** separable **Hilbert Space**

General model

- Dichotomy model for $\{\boldsymbol{\chi}_s : s \in D\}$, $\boldsymbol{\chi}_s$ in H

$$\boldsymbol{\chi}_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\boldsymbol{\delta}_s}_{\text{stochastic residual}}, \quad s \in D$$

- Linear model for the drift:

$$m_s = \bigoplus_{l=0}^{L} (f_l(s) \odot a_l),$$

H separable Hilbert space:

- can account for **differential** information (Sobolev spaces)
- can account for data **constraints** (Aitchison geometry)

The scalar regressors $f_l(s)$:

- can be a priori known → UK
- can describe an **external** drift → KED
- can incorporate **cluster** information → Class-Kriging

Menafoglio, Secchi, Dalla Rosa (2013)
Menafoglio, Guadagnini, Secchi (2016)

The advantages of a general approach: the application point of view

A new **general** and **coherent** framework for the treatment of **non-stationary** functional random fields valued in **any** separable **Hilbert Space**

General model

- Dichotomy model for $\{\boldsymbol{\chi}_s : s \in D\}$, $\boldsymbol{\chi}_s$ in H

$$\boldsymbol{\chi}_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\boldsymbol{\delta}_s}_{\text{stochastic residual}}, \quad s \in D$$

- Linear model for the drift:

$$m_s = \bigoplus_{l=0}^{L} (f_l(s) \odot a_l),$$

H separable Hilbert space:

- can account for **differential** information (Sobolev spaces)
- can account for data **constraints** (Aitchison geometry)

The scalar regressors $f_l(s)$:

- can be a priori known → UK

Case Study 1

- can describe an **external** drift → KED
- can incorporate **cluster** information → Class-Kriging

Menafoglio, Secchi, Dalla Rosa (2013)
Menafoglio, Guadagnini, Secchi (2016)

The advantages of a general approach: the application point of view

A new **general** and **coherent** framework for the treatment of **non-stationary** functional random fields valued in **any** separable **Hilbert Space**

General model

- Dichotomy model for $\{\boldsymbol{\chi}_s : s \in D\}$, $\boldsymbol{\chi}_s$ in H

$$\boldsymbol{\chi}_s = \underbrace{m_s}_{\text{deterministic drift}} \oplus \underbrace{\boldsymbol{\delta}_s}_{\text{stochastic residual}}, \quad s \in D$$

- Linear model for the drift:

$$m_s = \bigoplus_{l=0}^{L} (f_l(s) \odot a_l),$$

H separable Hilbert space:

- can account for **differential** information (Sobolev spaces)
- can account for data **constraints** (Aitchison geometry)

Case Study 2

The scalar regressors $f_l(s)$:

- can be a priori known → UK
- can describe an **external** drift → KED
- can incorporate **cluster** information → Class-Kriging

Menafoglio, Guadagnini, Secchi (2014)

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

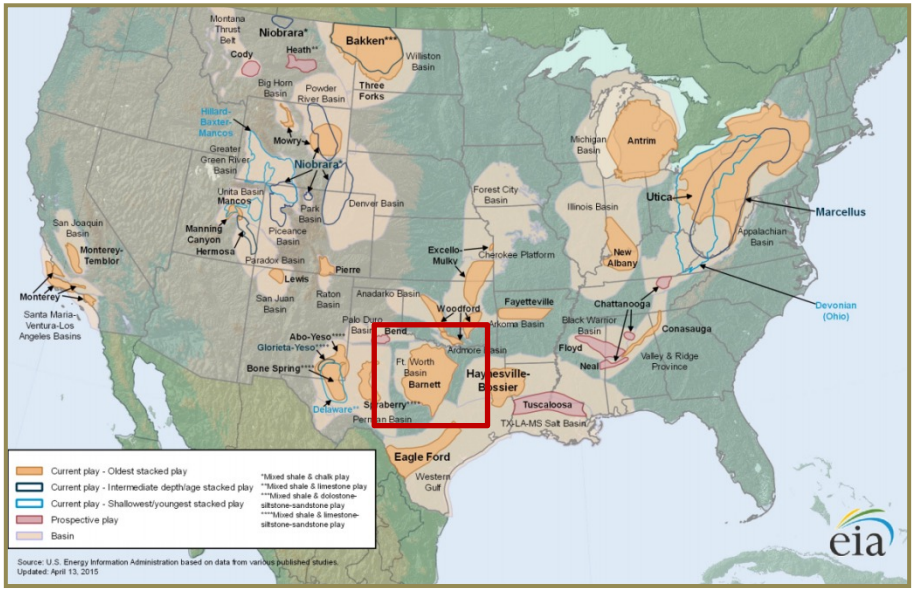
6.3. Spatial statistics for functional data

6.4. Two Case Studies

- Analysis of production profiles
- Analysis of particle-size distributions

Forecasting gas production rate curves at the Barnett shale system

Field site: Part of the Barnett shale system located in Texas



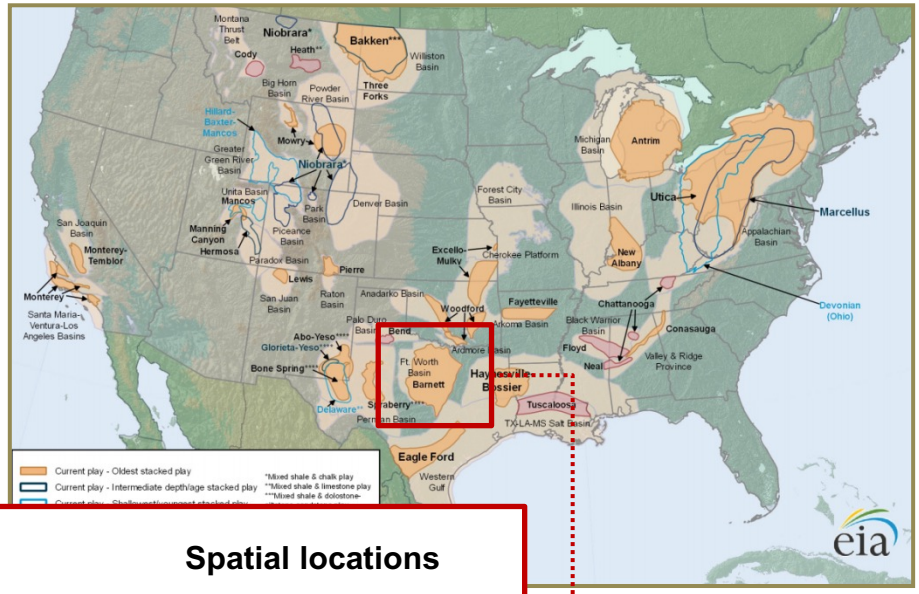
Data Source:
DrillingInfo.com

Forecasting gas production rate curves at the Barnett shale system

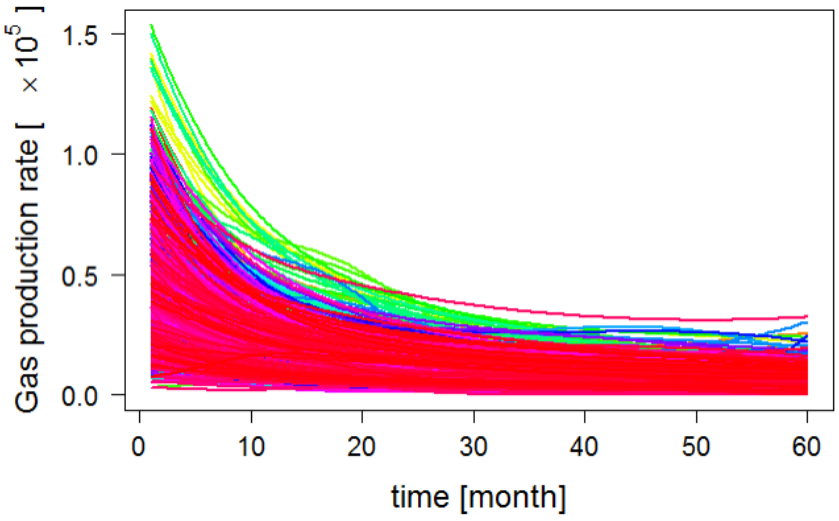
Field site: Part of the Barnett shale system located in Texas

The Data:

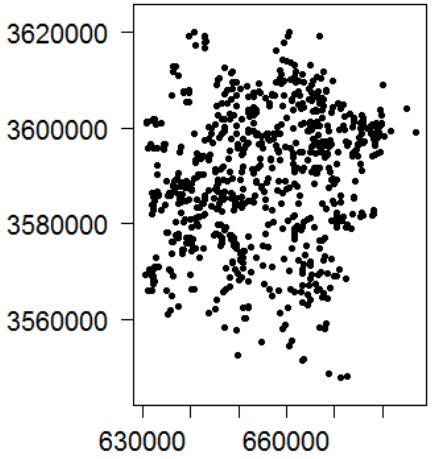
- Observations of the GPRCs at 922 wells in the area, along 60 months
- Data Preprocessing via smoothing (~denoising) with a B-spline basis



Smoothed Gas Production Rate Curves (GPRCs)



Spatial locations



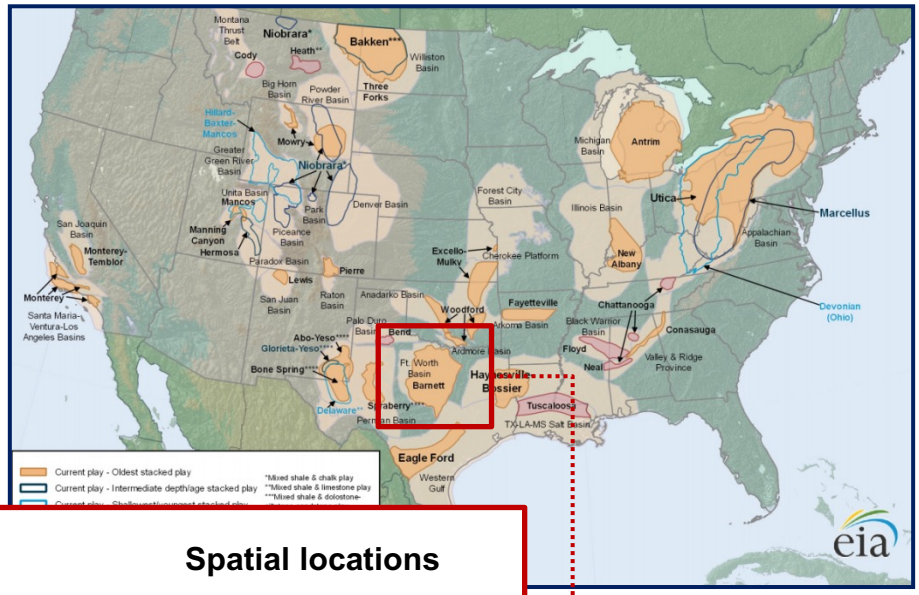
Data Source: DrillingInfo.com

Forecasting gas production rate curves at the Barnett shale system

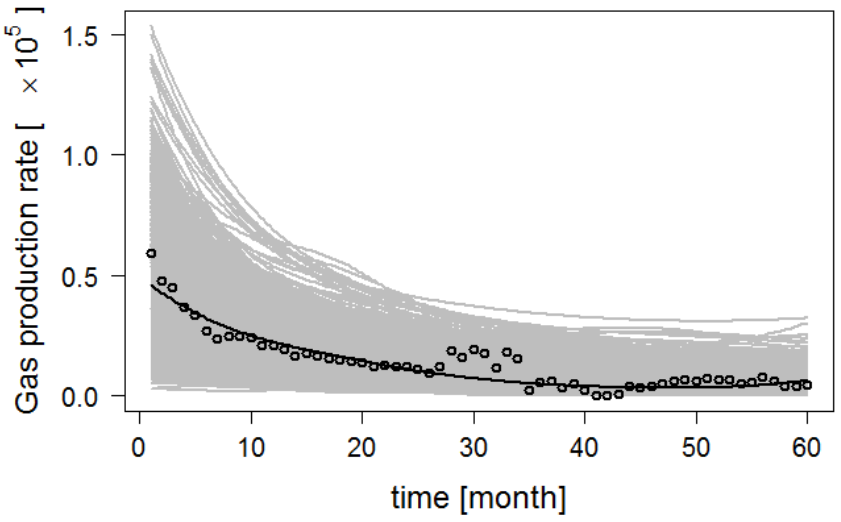
Field site: Part of the Barnett shale system located in Texas

The Data:

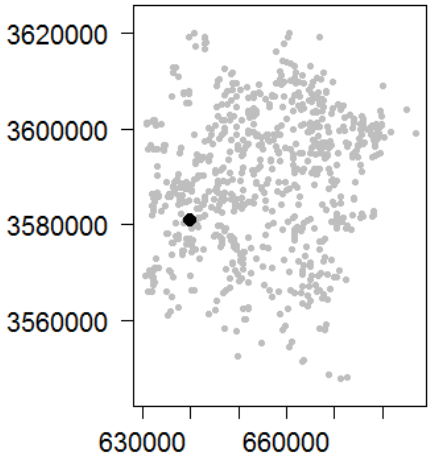
- Observations of the GPRCs at 922 wells in the area, along 60 months
- Data Preprocessing via smoothing (~denoising) with a B-spline basis



Smoothed Gas Production Rate Curves (GPRCs)



Spatial locations



Data Source: DrillingInfo.com

Forecasting gas production rate curves at the Barnett shale system

Field site: Part of the Barnett shale system located in Texas

The Data:

- Observations of the GPRCs at 922 wells in the area, along 60 months
- Data Preprocessing via smoothing (~denoising) with a B-spline basis

We randomly split the dataset in:

- Training set (461 wells)
- Test set (461 wells)

Forecasting gas production rate curves at the Barnett shale system

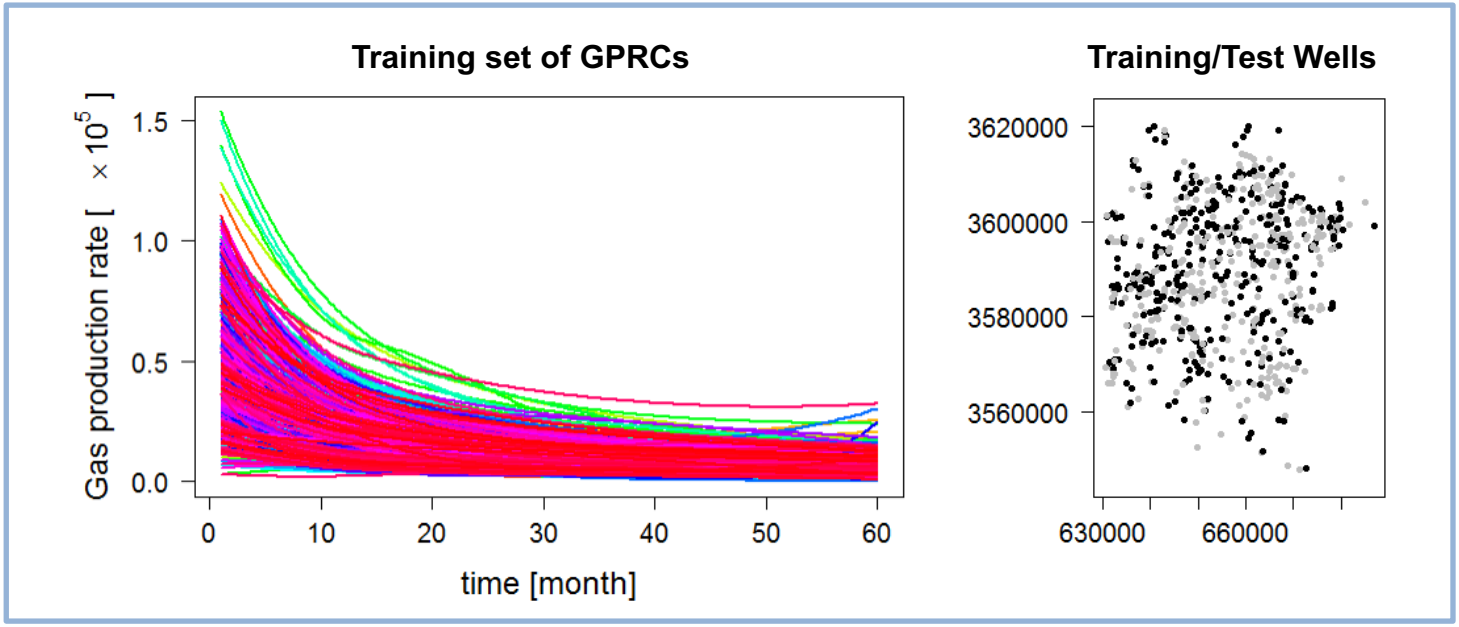
Field site: Part of the Barnett shale system located in Texas

The Data:

- Observations of the GPRCs at 922 wells in the area, along 60 months
- Data Preprocessing via smoothing (~denoising) with a B-spline basis

We randomly split the dataset in:

- Training set (461 wells)
- Test set (461 wells)



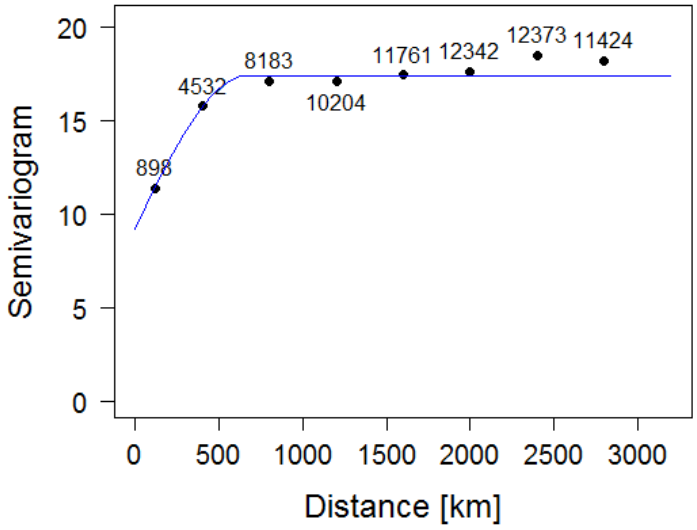
Model

$$\chi_s = \underbrace{m_s}_{\text{deterministic drift}} + \underbrace{\delta_s}_{\text{stochastic residual}}, \quad s \in D$$

Linear model for the drift:

$$m_s = a_0 + a_1 \cdot X + a_2 \cdot Y, \quad s = (X, Y) \in D$$

Trace-variogram: spherical model with nugget



Universal Kriging prediction of the test set

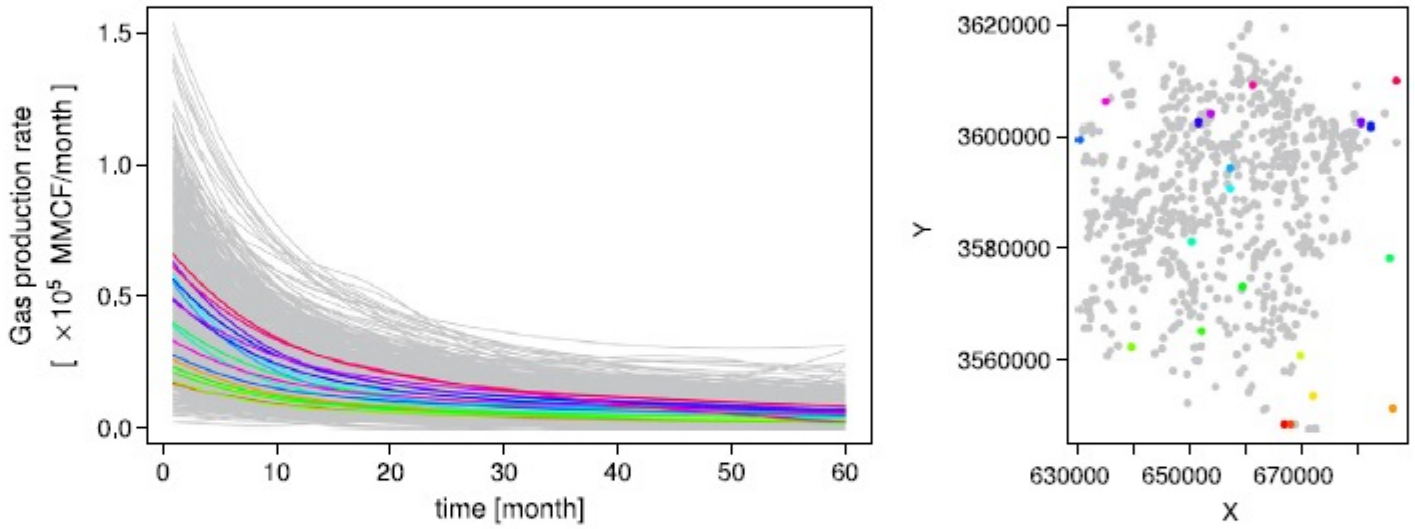
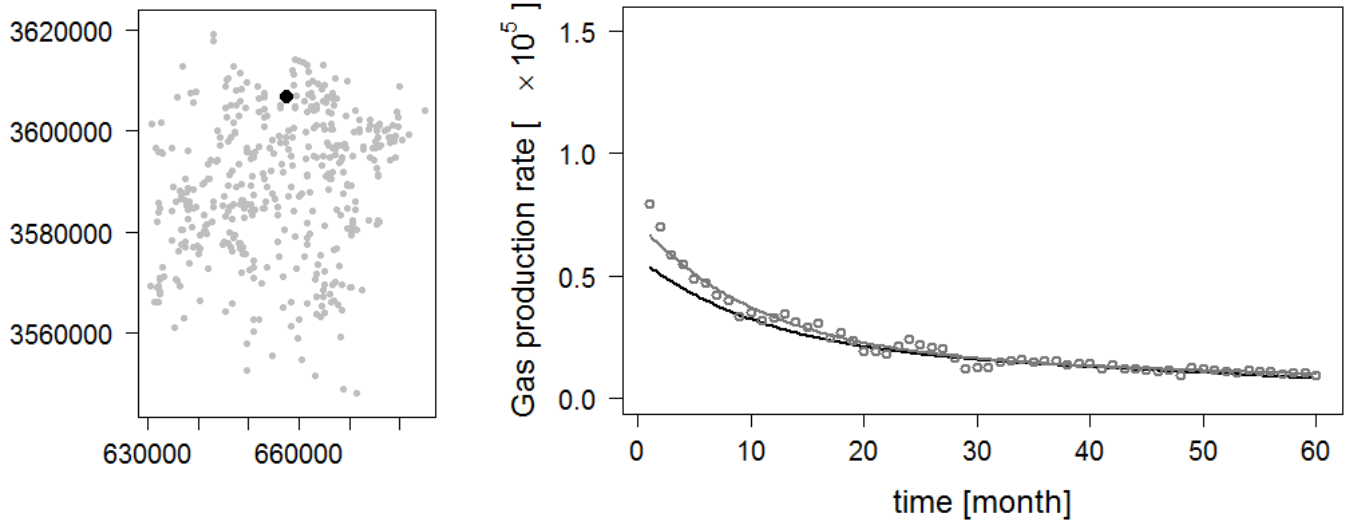


Fig. 3. Prediction by UTrK of GPRCs for 20 random locations at the Barnett shale. Left: smoothed data (gray lines) and predictions (colored lines). Right: sampled locations (gray symbols) and target locations (colored symbols). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from Menafoglio et al. (2016, SPATSTAT)

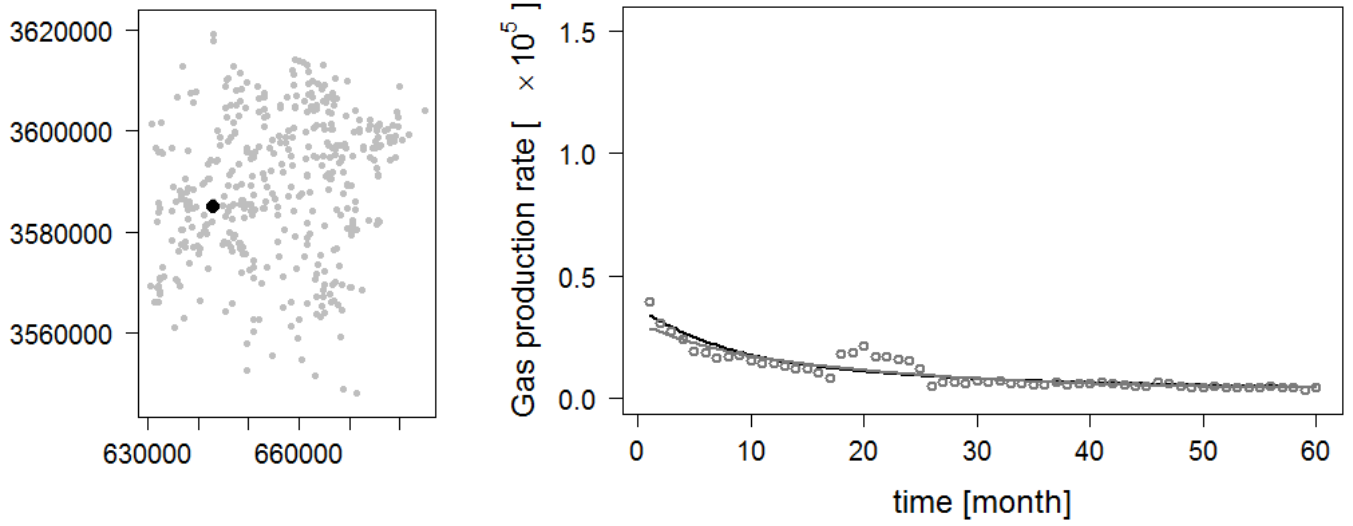
Universal Kriging prediction of the test set

Examples of predictions



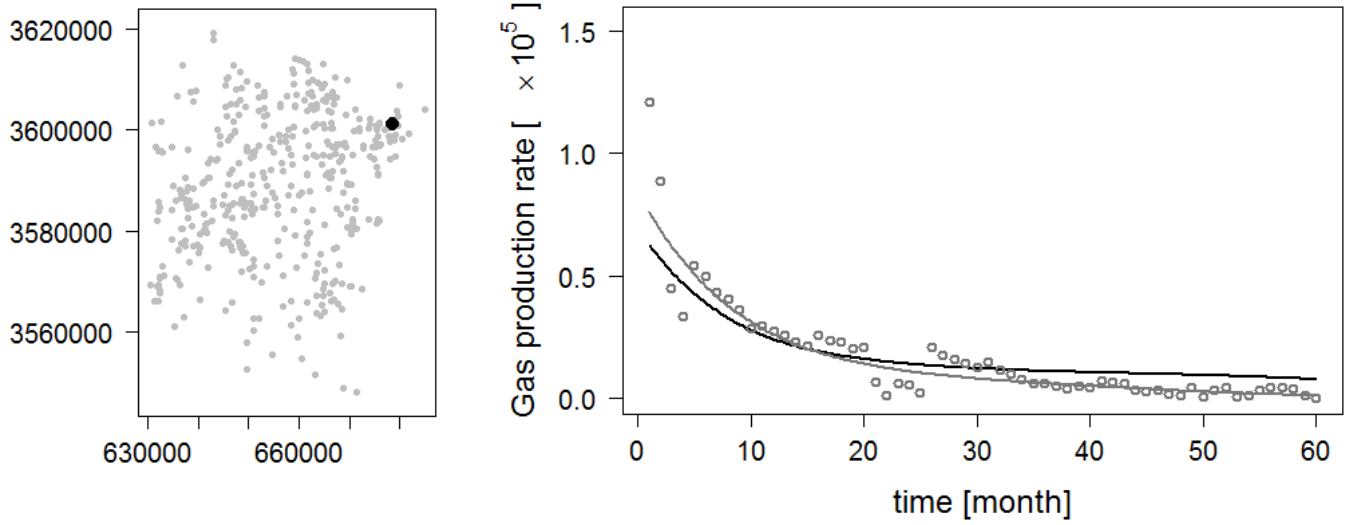
Universal Kriging prediction of the test set

Examples of predictions



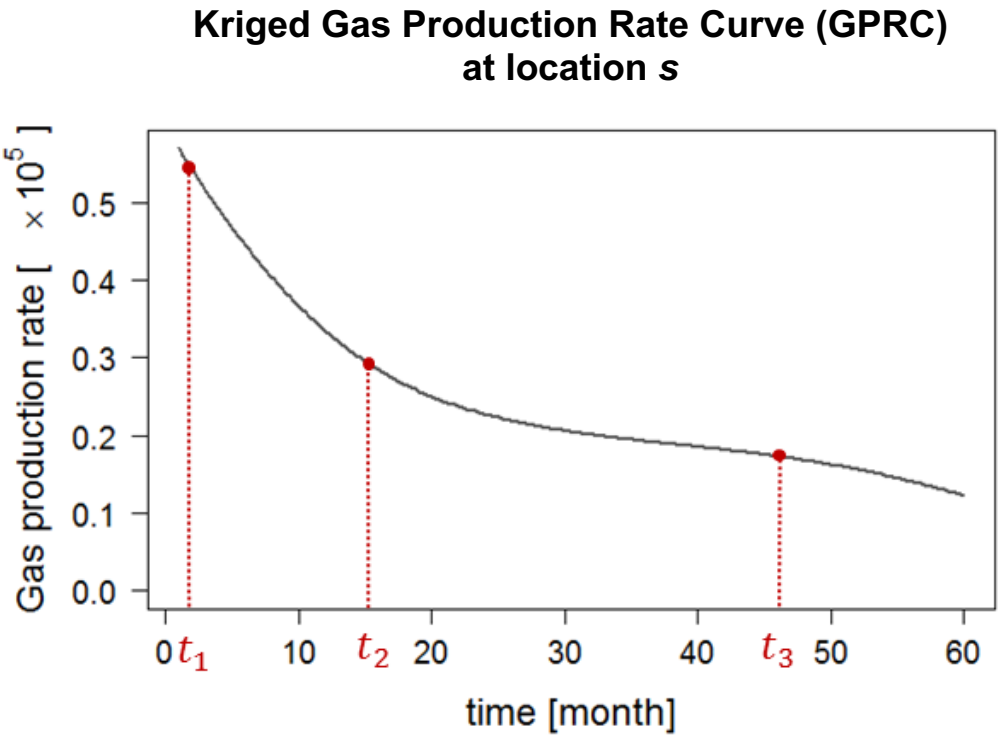
Universal Kriging prediction of the test set

Examples of predictions



Spatial maps via Functional Kriging

We can obtain maps of the gas production rate for significant time points



Spatial maps via Functional Kriging

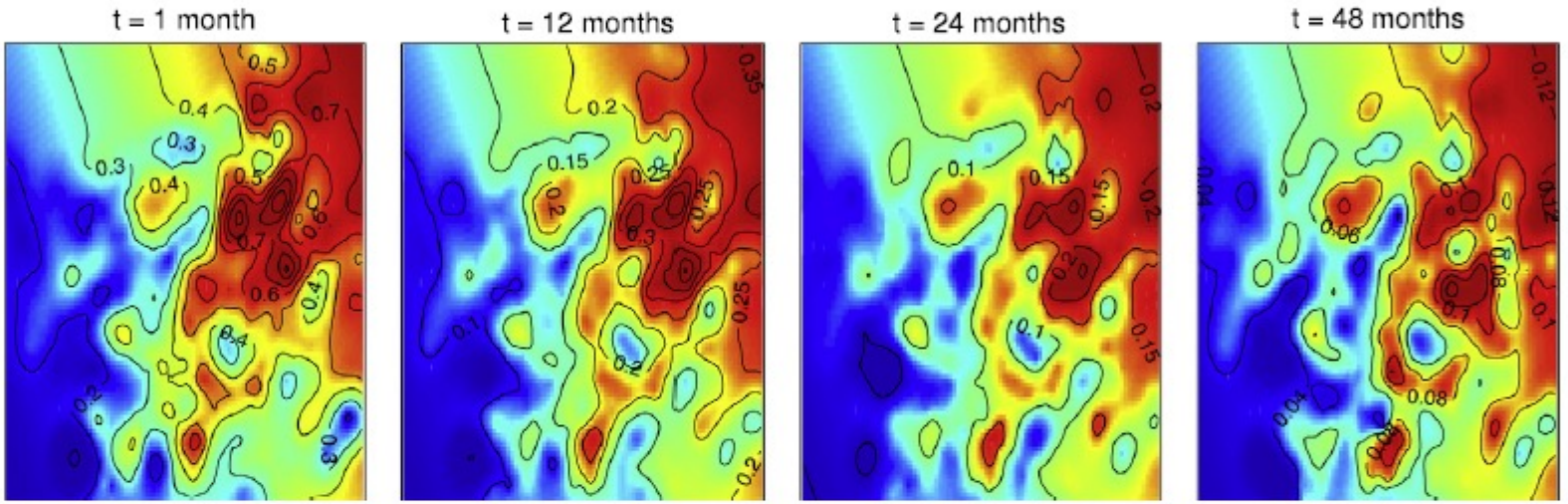


Fig. 4. Prediction maps obtained with UTrK at the Barnett shale, for $t = 1, 12, 24, 48$ months. Colors are given on a non-uniform scale. Gas rate reported on contour lines is meant up to a factor 10^5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from Menafoglio et al. (2016, SPATSTAT)

Agenda

6. Spatial statistics for functional data

6.1. Premiss: spatial statistics for object data

6.2. Basics of scalar geostatistics

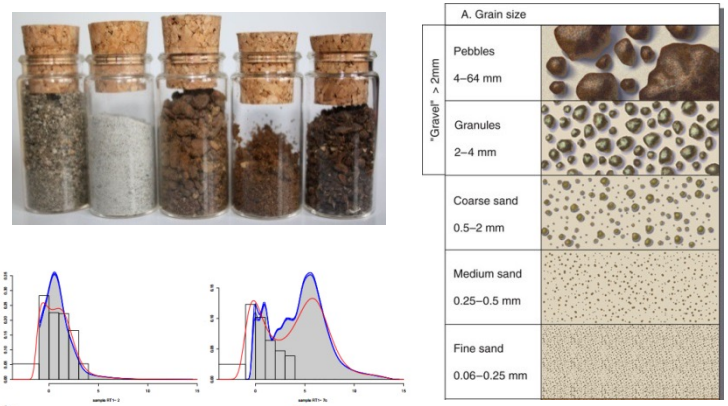
6.3. Spatial statistics for functional data

6.4. Two Case Studies

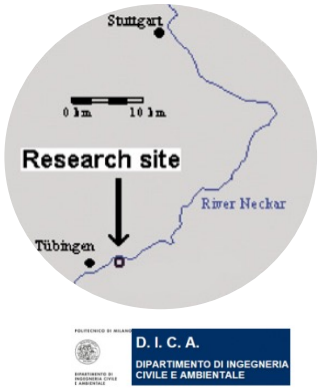
- Analysis of production profiles
- Analysis of particle-size distributions

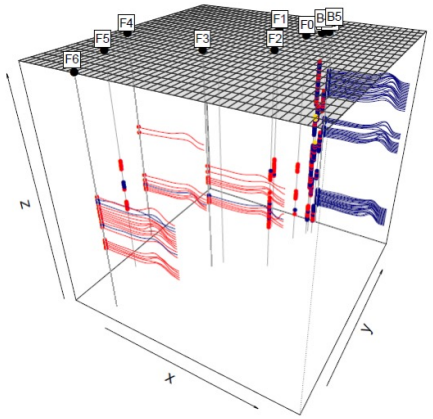
Analysis of particle-size curves in Heterogeneous Aquifers

Example: Particle-size distributions in heterogenous aquifer systems



Particle-size data





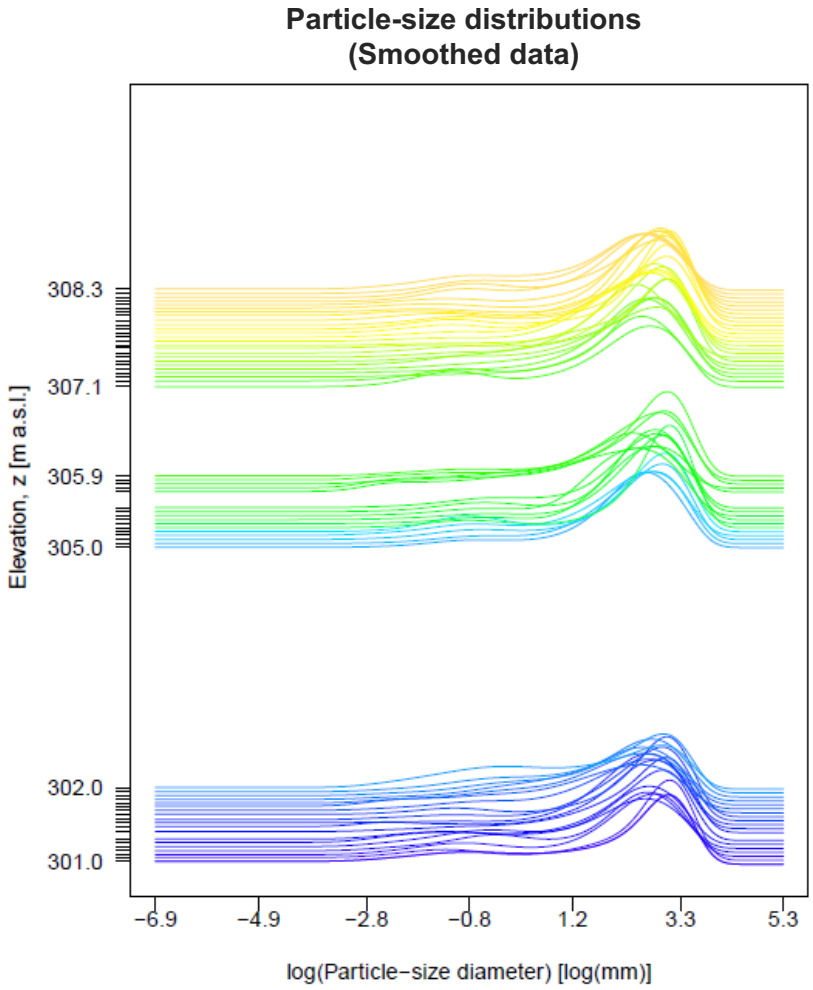
Data at the field site

Menafoglio, Guadagnini, Secchi (SERRA, 2014, 2016; MATHGEO 2016, WRR 2016)

Analysis of particle-size curves in Heterogeneous Aquifers

A closer look to the data

- **Test site**:
  - part of an alluvial heterogeneous aquifer in the Neckar river valley (Germany)
  - we first focus on data at the borehole B5
- **The data at B5**: 60 particle-size distributions reconstructed through **grain sieve analysis** based on 12 sieve diameters
- **Idea:**
  - Model particle-size distributions as cumulative distribution functions
  - Analyze their densities (after smoothing) as **functional compositional data** through an **appropriate geometry**



Menafoglio, Guadagnini, Secchi (2014a,b)

Hilbert space geometry for particle-size densities

Bayes Hilbert space $B^2(I)$

(van den Boogaart et al., 2014)

Space (of equivalence classes of) real valued functions on I with square-integrable log

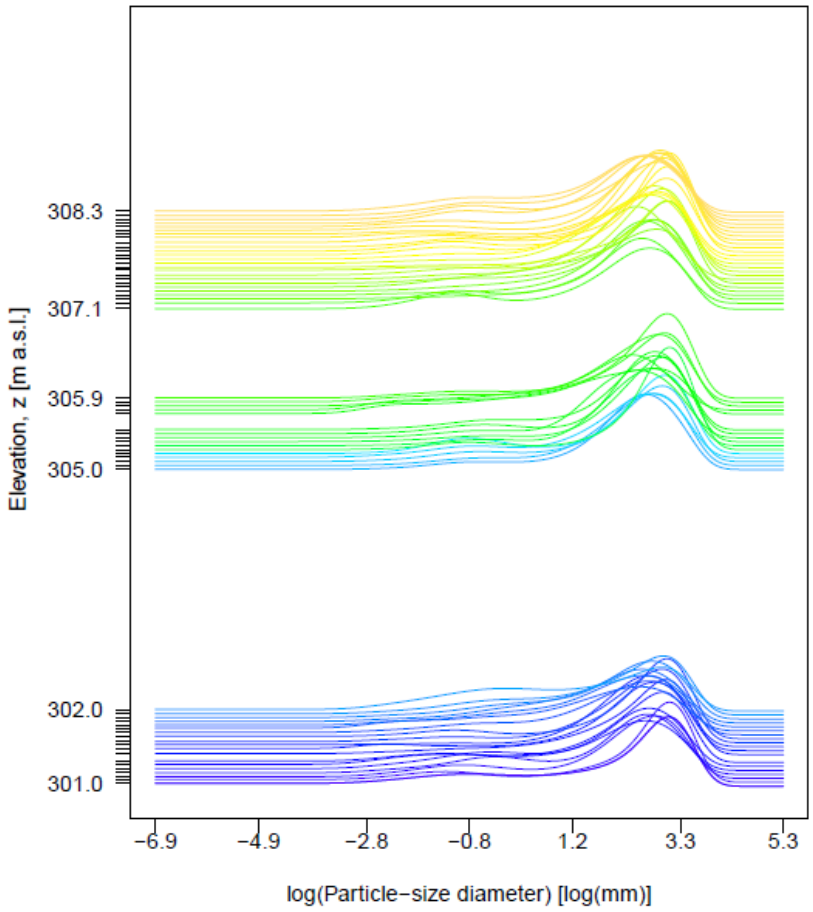
$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,ds}$$

$$(\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,ds}, \quad t \in I.$$

$$\langle f, g \rangle_{\mathcal{B}} = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)}\,dt\,ds$$

Data are embedded into the Hilbert space of functional compositions endowed with the generalized Aitchison geometry

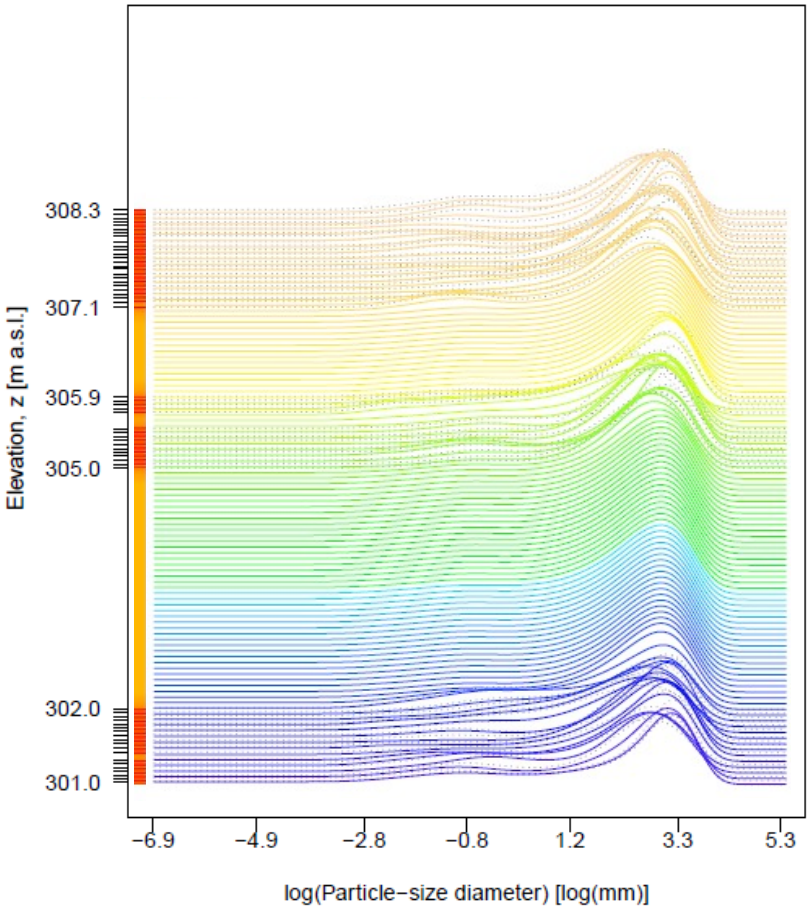
Particle-size densities (after smoothing)



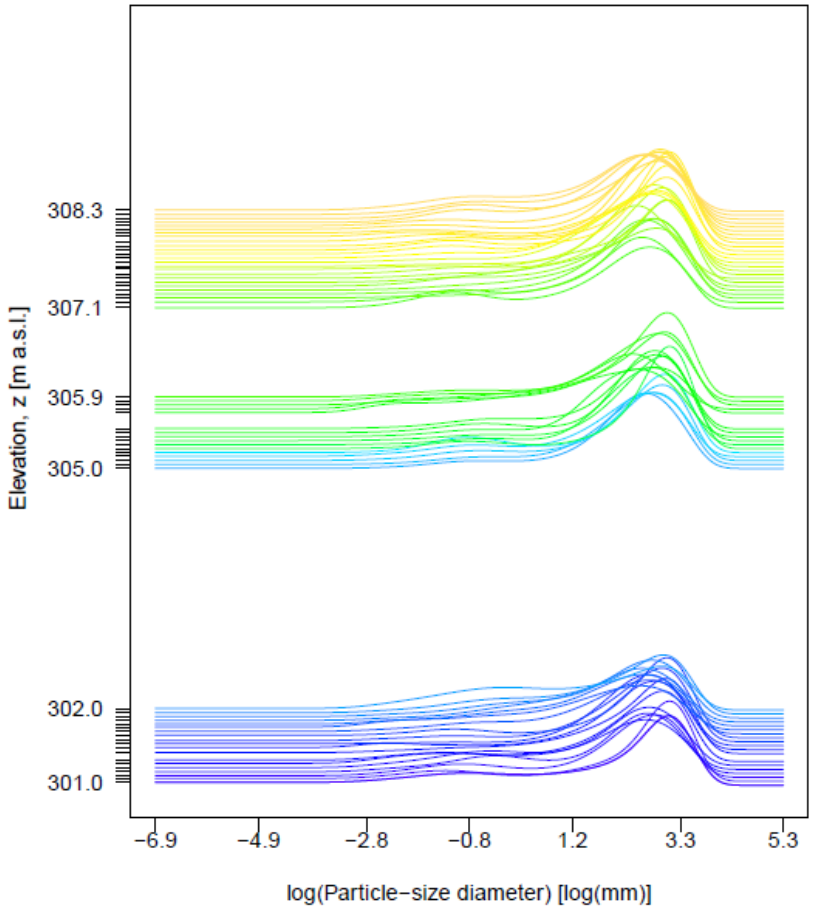
Menafoglio, Guadagnini, Secchi (2014)

Hilbert space geometry for particle-size densities

Kriging predictions and Kriging Variance



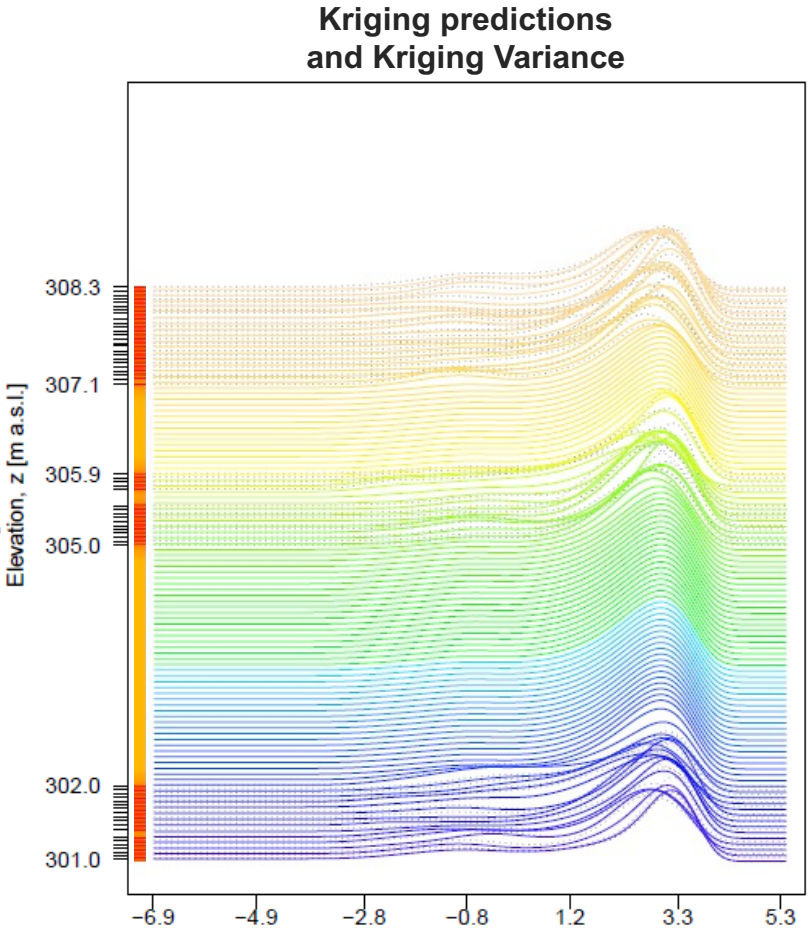
Particle-size densities (after smoothing)

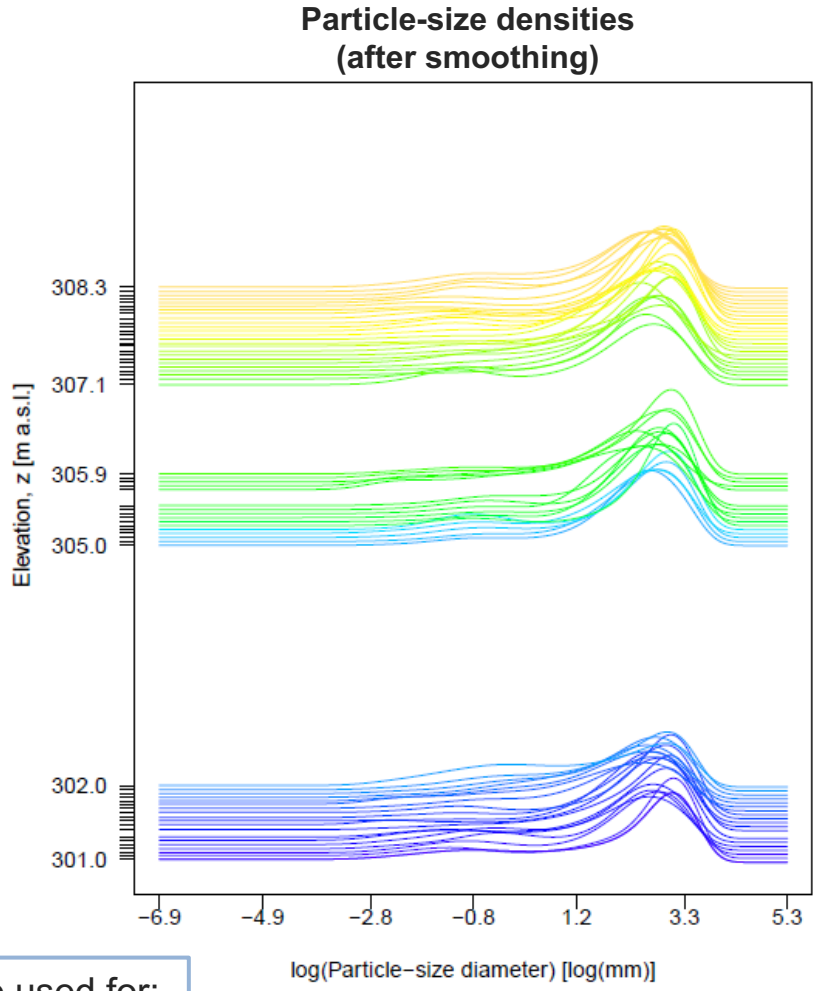


Functional Compositional Kriging

Menafoglio, Guadagnini, Secchi (2014)

Hilbert space geometry for particle-size densities





Kriged fields and the associated variance can be used for:

- Global quantile assessment
- Hydrofacies characterization in MC simulations

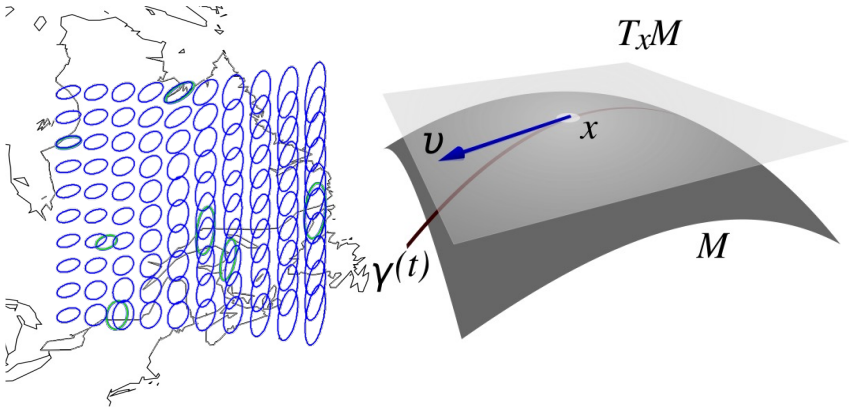
Menafoglio, Guadagnini, Secchi (2014)

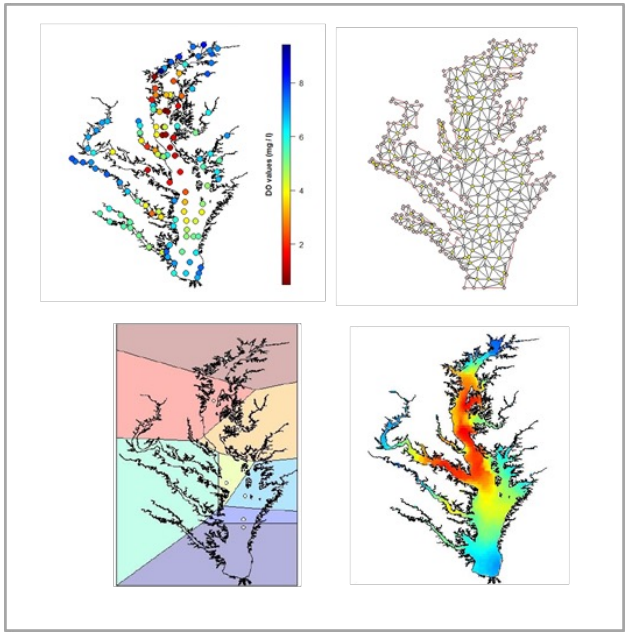
Take home messages

- Data with dependence arise in several environmental applications
- Neglecting the dependence may generate sub-optimal results
- In the presence of spatial dependence key typical issues are:
  - Estimation of the dependence
  - Prediction (kriging)
  - Uncertainty quantification
- Classical methods to face these issues can be generalized to the functional setting without substantial difference in the modeling effort with respect to the scalar case
- The embedding for the data (choice of the Hilbert space) is still crucial and strongly influences final results

Take home messages

- **Broad field of research**.
  Extensions include:
  - Beyond Hilbert spaces: Riemannian data, Banach data, more complex objects
  - Beyond simple domains: domains that are themselves objects





References

- J. P. Chilles and P. Definer (1999). Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, New York.
- N. Cressie. Statistics for Spatial data (1993). John Wiley & Sons, New York.
- N. Cressie and K. Wikle (2011). Statistics for Spatio-Temporal Data. Wiley.
- P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data. Environmetrics, 21(3-4):224-239.
- O. Gromenko, P. Kokoszka, L. Zhu, and J. Sojka. (2012) Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. Annals of Applied Statistics, 6(2):669-696.
- O. Grujic, A. Menafoglio, J. Caers (2017): “Functional Co-Kriging for multi-fidelity flow modeling”. Stochastic Environmental Research and Risk Assessment, forthcoming.
- K. Hron, A. Menafoglio, M. Templ, K. Hruzova, P. Filzmoser (2016): “Simplicial principal component analysis for density functions in Bayes spaces”, Computational Statistics & Data Analysis, 94, 330-350.
- J. S. Marron and A. M. Alonso. Overview of object oriented data analysis. Biometrical Journal, 56(5):732-753, 2014.
- A. Menafoglio, O. Grujic, J. Caers (2016): “Universal kriging of functional data: trace-variography vs cross-variography? Application to forecasting in unconventional shales”, Spatial Statistics, 15, 39–55.
- A. Menafoglio, A. Guadagnini, P. Secchi (2014): “A Kriging Approach based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers”, Stochastic Environmental Research and Risk Assessment, 28(7), 1835-1851.
- A. Menafoglio, A. Guadagnini, P. Secchi (2016): “Stochastic Simulation of Soil Particle-Size Curves in Heterogeneous Aquifer Systems through a Bayes space approach”, Water Resources Research, 52, 5708–5726.
- A. Menafoglio, G. Petris. (2017) Kriging for Hilbert-space valued random fields: the Operatorial point of view. Journal of Multivariate Analysis, 146, 84–94
- A. Menafoglio, P. Secchi (2017): “Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics”, European Journal of Operational Research, 258(2), pages 401–410.
- A. Menafoglio, P. Secchi, M. Dalla Rosa (2013): “A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space”, Electronic Journal of Statistics 7, 2209–2240.
- A. Menafoglio, P. Secchi, A. Guadagnini (2016): “A Class-Kriging predictor for Functional Compositions with Application to Particle-Size Curves in Heterogeneous Aquifers”, Mathematical Geosciences.
- D. Pigoli, A. Menafoglio and P. Secchi (2016): “Kriging prediction for manifold-valued random fields”, Journal of Multivariate Analysis, 145, 117–131.
- R. Tolosana-Delgado, K. G. van den Boogaart, and V. Pawlowsky-Glahn. (2011) Geostatistics for Compositions, pages 73{86. John Wiley & Sons, Ltd, Pawlowsky- Glahn & Buccianti edition, 2011
- K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. Australian & New Zealand Journal of Statistics, 56:171-194, 2014.