Object Oriented Spatial Statistics for the analysis of georeferenced complex data

Alessandra Menafoglio

Fontainebleau, 7th September 2023

MOX, Department of Mathematics, Politecnico di Milano
Object Oriented Spatial Statistics

Why?

- Complex data increasingly available in industrial and environmental studies

**Example: Mortality data**

*Key problem: study of spatial structures & anomaly detection*

Incidence of mortality (2017-2020)

Daily death counts over 2017-2020, 70+ years

Scimone, Menafoglio, Sangalli, Secchi (SPASTA2022)
Object Oriented Spatial Statistics

Why?

- Complex data increasingly available in industrial and environmental studies

Example: functional data in the oil&gas industry

Data on sediment compaction curves and gas rate production curves

Dalla Rosa et al. (IAMG 2013); Menafoglio et al. (SPASTA 2016)
Object Oriented Spatial Statistics

Why?

- Complex data increasingly available in industrial and environmental studies

**Example: functional data as outputs of numerical models**

Response of a numerical model of fluid flow in a reservoir

*Simulation of fluid flow in subsurface*  
*Field Water Production Rate*

Grujic et al. (SERRA 2018)
Object Oriented Spatial Statistics

Why?

- Complex data increasingly available in industrial and environmental studies

Example: distributional data in environmental monitoring

Distribution of dissolved oxygen in the Chesapeake Bay

Menafoglio, Gaetani, Secchi (SERRA, 2018); Menafoglio, Pigoli, Secchi (JCGS, 2021)
Object Oriented Spatial Statistics

Why?

- Complex data increasingly available in industrial and environmental studies
- Classical approaches typically yield information loss

**Example: distributional data in environmental monitoring**

Distribution of dissolved oxygen in the Chesapeake Bay

(Menafoglio, Gaetani, Secchi, SERRA, 2018; Menafoglio, Pigoli, Secchi, JCGS)
**Object Oriented Spatial Statistics (O2S2)**

- Complex data increasingly available in industrial and environmental studies
- Classical approaches typically yield **information loss**

**What?**

- **The “atom” of the analysis is the entire object, rather than a limited number of selected features of the data**
- **A system of ideas & methods to**
  - Estimate the **spatial dependence**
  - **Prediction** (kriging) at unsampled locations
  - Uncertainty quantification
Object Oriented Spatial Statistics

What?

- Complex data increasingly available in industrial and environmental studies
- Classical approaches typically yield **information loss**

Object Oriented Spatial Statistics (O2S2)

In this talk:
- **Object Oriented Spatial Modeling**
- of functional data
- over complex domains
Outline

➢ Introduction and Motivations

➢ An overview on O2S2

➢ Modeling object data over complex domains
  • A divide-and-conquer approach
  • The case of stream network domains

➢ Concluding remarks
Outline

➢ Introduction and Motivations

➢ An overview on O2S2

➢ Modeling object data over complex domains
  • A divide-and-conquer approach
  • The case of stream network domains

➢ Concluding remarks
Data as Objects

The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space.
Data as Objects
The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space

*If the data were vectors of two dimensions…*

**Euclidean space** $\mathbb{R}^2$

- **Sum:** $v_1 + v_2 = (x_1 + x_2, y_1 + y_2)$
- **Product by a constant:** $c \cdot v = (c \cdot x, c \cdot y)$
- **Norm (length 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:** $\theta = \arccos \frac{\langle v_1, v_2 \rangle}{\|v_1\| \|v_2\|}$

*Operations* (+, ∙)

*Inner product*

$\langle v_1, v_2 \rangle = (x_1 \cdot x_2) + (y_1 \cdot y_2)$
Data as Objects

The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space

*If the data were vectors of two dimensions…*

*… for complex data of any (even infinite) dimension…*
Object Oriented Spatial Statistics

The foundational idea

**Data as Objects**

The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space.

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

---

**Hilbert space** $H$

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

- Sum: $f_1 + f_2$
- Product by a constant: $c \cdot f$
- Norm: $\|f\|$
- Distance: $\|f_1 - f_2\|$
- Angle: $\theta = \arccos \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$

Operations ($+$, $\cdot$)

Inner product $\langle f_1, f_2 \rangle$
Data as Objects
The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space

*If the data were vectors of two dimensions…*

*… for complex data of any (even infinite) dimension…*

**Example: Hilbert space of square integrable functions \( L^2 \)**

- Sum: \((f_1 + f_2)(t) = f_1(t) + f_2(t)\)
- Product by a constant: \((c \cdot f)(t) = c \cdot f(t)\)
Object Oriented Spatial Statistics
The foundational idea

Data as Objects
The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space

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

Hilbert space $H$

Example: Hilbert space of square integrable functions $L^2$

- Sum: $(f_1 + f_2)(t) = f_1(t) + f_2(t)$
- Product by a constant: $(c \cdot f)(t) = c \cdot f(t)$
- Norm: $\|f\| = \int (f(t))^2 dt$
- Distance: $\|f_1 - f_2\| = \int (f_1(t) - f_2(t))^2 dt$
- Angle: $\vartheta = \arccos \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}$
Object Oriented Spatial Statistics
The foundational idea

Data as Objects
The «atom» of the analysis is the entire object, considered as a point within an appropriate mathematical space

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

Hilbert space $H$

- Not just $L^2$ (space of square integrable functions)!
- Should capture the «appropriate» geometry for the data

*The choice of $H$ and its geometry should NOT be driven by mathematical convenience only, but by the data characteristics which are deemed important for the application*
A paradigmatic example: PDF data

Example: Monitoring dissolved oxygen in the Chesapeake Bay

Data object: Probability Density Functions (PDFs) of dissolved oxygen at 110 monitoring stations

(a) PDFs of DO
(b) Mean of DO
A paradigmatic example: PDF data

Example: Monitoring dissolved oxygen in the Chesapeake Bay

Data object: Probability Density Functions (PDFs) of dissolved oxygen at 110 monitoring stations

- Density data are constrained functional data
- The $L^2$ geometry becomes meaningless in the presence of density data
- FDA methods in $L^2$ may provide uninterpretable results

Example: $L^2$ sum of two Gaussian PDFs
A paradigmatic example: PDF data

Example: Monitoring dissolved oxygen in the Chesapeake Bay

Data object: Probability Density Functions (PDFs) of dissolved oxygen at 110 monitoring stations

- Density data are constrained functional data
- The $L^2$ geometry becomes meaningless in the presence of density data
- FDA methods in $L^2$ may provide uninterpretable results

The Problem: The space of PDFs is not a linear (Hilbert) space when the $L^2$ geometry is used

A Solution: Bayes Hilbert Space (from Compositional Data Analysis)
Discrete PDFs as multivariate compositions

\[(x_1, \ldots, x_D) \in \mathbb{R}^D\]
\[x_i > 0, \sum_i x_i = const\]

- \(x_i\) represents a parts of a whole according to a partition of the domain
- Convey only relative information: (log) ratios between parts provide the meaningful info.
PDFs as functional compositional data

Discrete PDFs as multivariate compositions

\[(x_1, \ldots, x_D) \in \mathbb{R}^D\]
\[x_i > 0, \sum_i x_i = \text{const}\]

- \(x_i\) represents a parts of a whole according to a partition of the domain
- Convey only relative information: (log) ratios between parts provide the meaningful info.

Continuous PDFs as functional compositions:

\[x : \mathcal{T} \to \mathbb{R}\]
\[x(t) > 0, \int_{\mathcal{T}} x(t)dt = \text{const}\]

- Infinite-dimensional object (a function)
- Point-wise values represent infinitesimal parts of a whole (e.g., unity).
PDFs as functional compositional data

Discrete PDFs as multivariate compositions

\[(x_1, \ldots, x_D) \in \mathbb{R}^D\]
\[x_i > 0, \sum_i x_i = \text{const}\]

- \(x_i\) represents a parts of a whole according to a partition of the domain
- Convey only relative information: (log) ratios between parts provide the meaningful info.

Continuous PDFs as functional compositions:

\[x : \mathcal{T} \to \mathbb{R}\]
\[x(t) > 0, \int_{\mathcal{T}} x(t) dt = \text{const}\]

- Infinite-dimensional object (a function)
- Point-wise values represent infinitesimal parts of a whole (e.g., unity).

Aitchison Geometry

\(\mathbb{R}^3\)

Data obj.

Refine the domain partition

\(L^2\)

Hilbert Geometry?
Bayes Hilbert geometry (Egozcue et al., 2006)

Perturbation + and powering ⋅

\[(f + g)(t) = \frac{f(t)g(t)}{\int_{\mathcal{F}} f(s)g(s) \, ds},\]

\[(\alpha \cdot f)(t) = \frac{f(t)^\alpha}{\int_{\mathcal{F}} f(s)^\alpha \, ds} ,\]

Inner product \(\langle \cdot, \cdot \rangle\)

\[\langle f, g \rangle = \frac{1}{2|\mathcal{V}|} \int_{\mathcal{V}} \int_{\mathcal{V}} \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} \, dt \, ds,\]

\[f, g \in A^2(\mathcal{F})\]

Only (log-)ratios between probabilities are meaningful (~ odds-ratio) as data represent the distribution of a total mass over a domain.

Example: \(B^2\) sum of two Gaussian PDFs
Bayes spaces for PDF data

Bayes Hilbert geometry (Egozcue et al., 2006)

**Perturbation + and powering •**

\[(f + g)(t) = \frac{f(t)g(t)}{\int_{\mathcal{F}} f(s)g(s) \, ds};\]

\[(\alpha \cdot f)(t) = \frac{f(t)^\alpha}{\int_{\mathcal{F}} f(s)^\alpha \, ds};\]

**Inner product \( \langle \cdot , \cdot \rangle \)**

\[\langle f, g \rangle = \frac{1}{2|\mathcal{F}|} \int_{\mathcal{F}} \int_{\mathcal{F}} \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} \, dt \, ds, \quad f, g \in A^2(\mathcal{F})\]

Only (log-)ratios between probabilities are meaningful (∼ odds-ratio) as data represent the distribution of a *total* mass over a domain

Meaningful interpretations in mathematical statistics, e.g.,

- Exponential families as affine finite-dimensional subspaces
- Perturbation + as a Bayes update of information

Remarks:

- Problems with zeros
- Other approaches are possible: alpha-transformations (Tsagris et al 2011, Clarotto et al 2022)
Examples of PDF data

PDFs of dissolved oxygen in the Chesapeake Bay

(a) PDFs of DO
(b) Mean of DO

Particle-size densities

PDF data might also be convenient exhaustive summaries of large databases
Object Oriented Spatial Statistics (O2S2)

- The “atom” of the analysis is the entire object, embedded in an appropriate mathematical (Hilbert) space.
- **Object-oriented approach** to:
  - Estimate the **spatial dependence**
  - **Prediction** (kriging) at unsampled locations
  - Uncertainty quantification
O2S2 (Hilbert data) in a nutshell

Review: Menafoglio & Secchi (2017)

Spatially dependent object data
O2S2 (Hilbert data) in a nutshell

Object Oriented approach

- $\{\chi_s : s \in D\}$ random field, $D \subset \mathbb{R}^d$
- $\chi_s$ : random element of a separable Hilbert space $H(+,\cdot,\langle \cdot,\cdot \rangle,\|\cdot\|)$ [feature space]
- $\chi_{s1}, \ldots, \chi_{sn}$: observations of the field at $n$ sampled locations

Spatially dependent object data

Points of a Hilbert space
Object Oriented approach

- \( \{ \chi_s : s \in D \} \) random field, \( D \subset \mathbb{R}^d \)
- \( \chi_s \): random element of a separable Hilbert space \( H(+,\cdot,\langle \cdot,\cdot \rangle,\|\cdot\|) \) [feature space]
- \( \chi_{s_1}, \ldots, \chi_{s_n} \): observations of the field at \( n \) sampled locations

Mean and Spatial Dependence

- Mean in \( H \) (Frechét) \( m_s = \mathbb{E}[\chi_s] \),
- Trace-variogram:
  \[
  2\gamma(s_i, s_j) = \text{Var}_H(\chi_{s_i} - \chi_{s_j}) = \\
  = \mathbb{E}[\|\chi_{s_i} - \chi_{s_i}\|^2] - \mathbb{E}[\chi_{s_i} - \chi_{s_j}]^2 
  \]

Stationarity:

- Constant mean
- Trace-variogram (and trace-covariogram) depending only on the increment between locations
Object Oriented approach

- $\left\{ \chi_s : s \in D \right\}$ random field, $D \subset \mathbb{R}^d$
- $\chi_s$: random element of a of a separable Hilbert space $H(\cdot,\cdot,\langle \cdot,\cdot \rangle,\|\cdot\|)$ [feature space]
- $\chi_{s_1}, \ldots, \chi_{s_n}$: observations of the field at $n$ sampled locations

Mean and Spatial Dependence

- Mean in $H$ (Frechét) $m_s = \mathbb{E}[\chi_s]$
- Trace-variogram:
  
  $$2\gamma(s_i, s_j) = \text{Var}_H(\chi_{s_i} - \chi_{s_j}) = \mathbb{E}[\|\chi_{s_i} - \chi_{s_j}\|^2] - \|\mathbb{E}[\chi_{s_i} - \chi_{s_j}]\|^2$$

Variogram estimation

1. Empirical estimate (MoM) (stationarity)
   
   $$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} \|\chi_{s_i} - \chi_{s_j}\|^2$$

2. Fit of a parametric model (LS) e.g., spherical, matérn
O2S2 (Hilbert data) in a nutshell

Object Oriented Kriging

Goal: Prediction of a generic $\chi_{s_0}$ at site $s_0$ through the Kriging Predictor (BLUP):

$$\chi_{s_0}^* = \sum_{i=1}^{n} \lambda_i^* \cdot \chi_{s_i}$$

Formulation:

- Find $(\lambda_1^*, \ldots, \lambda_n^*) \in \mathbb{R}^n$ that solve

$$\min_{(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n} \mathbb{E} \left[ \left\| \chi_{s_0} - \sum_{i=1}^{n} \lambda_i \cdot \chi_{s_i} \right\|^2 \right]$$

subject to $\mathbb{E}[\chi_{s_0}] = \mathbb{E} \left[ \sum_{i=1}^{n} \lambda_i \cdot \chi_{s_i} \right]$
Theorem (Ordinary Kriging)
The optimal weights $(\lambda_1^*, \ldots, \lambda_n^*) \in \mathbb{R}^n$ are found by solving

$\begin{bmatrix} \gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \zeta \end{bmatrix} = \begin{bmatrix} \gamma(s_0, s_1) \\ \vdots \\ \gamma(s_0, s_n) \end{bmatrix}$

Remarks:
- Under assumptions to the 1D case, non-stationary solutions are available (UK, KED)
- The result generalizes that of Giraldo et al. (2011), Nerini et al (2008) [stationarity, $H=\mathbb{L}^2$]
Outline

➢ Introduction and Motivations

➢ An overview on O2S2

➢ Modeling object data over complex domains
  • A divide-and-conquer approach
  • The case of stream network domains

➢ Concluding remarks
Two environmental examples: complex spatial domains

- Monitoring dissolved oxygen in the Chesapeake Bay
- Data object: PDFs of dissolved oxygen at 110 monitoring stations
- **Spatial domain:** Estuarine system
Two environmental examples: complex spatial domains

- Monitoring dissolved oxygen in the Chesapeake Bay
  - Data object: PDFs of dissolved oxygen at 110 monitoring stations
  - Spatial domain: Estuarine system
- Monitoring dissolved water temperature along the Middle Fork river
  - Data object: temperature profiles at 157 monitoring stations
  - Spatial domain: Stream network
Geostatistics over complex domains

- Non-Euclidean spatial domains: widely-used parametric covariance families may no longer be valid (Curriero et al., 2006)
- No valid model available for general domains
- Valid covariance models developed in special cases:
  - spherical domains (Jun & Stein, AOAS 2008; Huang et al., MATG, 2011; Porcu et al., SERRA 2013)
  - stream networks (Ver Hoef & Peterson, JASA 2010; Asadi et al., AOAS 2015)

(Jeong et al., STAT SCI 2017)

(O'Donnell et al., JRSS-C 2014)
O2S2 over complex domains

Two approaches to O2S2 over complex domains

Divide & conquer:
general validity, but no
global model

Derivation of global valid models (stream networks)
Two approaches to O2S2 over complex domains

Divide & conquer: general validity, but no global model

Derivation of global valid models (stream networks)
Globally
- The process is **non-stationary**
- The domain is **complex**: non-Euclidean metric for the spatial domain

Locally
- The process is **stationary** (proxy)
- The domain is **simple**: Euclidean metric for the spatial domain (proxy)
Main idea: a perturbation approach

- **DIVIDE ET IMPERA**

- Iterative *random decomposition* of the spatial *domain* (RDDs)
- **Local analyses**, conditioned on the data and the partition, under the *stationarity assumption* (and possible linearization)
- Final *aggregation* of the local analyses
Object-Oriented Kriging via RDD

I. **Initialization.**
Set model parameters.

II. **Bootstrap step.**
for $b = 1$ to $B$ do
1. Random domain decomposition
2. For each neighborhood: perform a local analysis
   • Local estimate of the spatial dependence;
   • Local Kriging predictions.
end for.

Result: $B$ weak Kriging maps, each of them is defined by $K$ local maps.

III. **Aggregation step.**
The weak Kriging maps are aggregated into a final strong Kriging map (e.g. by averaging).
A case study on Dissolved Oxygen in the Chesapeake Bay

- Monitoring dissolved oxygen in the Chesapeake Bay
- Data object: PDFs of dissolved oxygen at 110 monitoring stations
- Spatial domain: Estuarine system
A case study on Dissolved Oxygen in the Chesapeake Bay

- Monitoring dissolved oxygen in the Chesapeake Bay
- Data object: PDFs of dissolved oxygen at 110 monitoring stations
- Spatial domain: Estuarine system

Representation of the data: embedding into a Bayes Hilbert space

Bayes Hilbert geometry (Egozcue et al., 2006)

Perturbation $+\,$ and powering $\cdot$

$$(f + g)(t) = \frac{f(t)g(t)}{\int_\mathcal{F} f(s)g(s)\,ds},$$

$$(\alpha \cdot f)(t) = \frac{f(t)^\alpha}{\int_\mathcal{F} f(s)^\alpha\,ds},$$

Inner product $\langle \cdot, \cdot \rangle$

$$\langle f, g \rangle = \frac{1}{2|\mathcal{F}|} \int_\mathcal{F} \int_\mathcal{F} \ln f(t) \ln g(s) \, dt\, ds,$$

$$f, g \in A^2(\mathcal{F})$$
A case study on Dissolved Oxygen in the Chesapeake Bay

- Monitoring dissolved oxygen in the Chesapeake Bay
- Data object: PDFs of dissolved oxygen at 110 monitoring stations
- Spatial domain: Estuarine system

*Irregularly shaped* domain. The graph-based distance between two points is defined as the length of the shortest path on a graph defined by the *Delaunay triangulation* of the domain.
A case study on Dissolved Oxygen in the Chesapeake Bay

Menafoglio, Gaetani, Secchi (SERRA, 2018)

Results of Object Oriented Kriging via RDD

Dead zone: DO < 2 mg/l

Probability $p$ of observing {DO < 2 mg/l} estimated from the predicted density

Areas of attention: $p > 0.5$
A case study on Dissolved Oxygen in the Chesapeake Bay

Menafoglio, Gaetani, Secchi (SERRA, 2018)

(a) Probability of DO < 2 mg/l (K=16)  
Divide&Impera via RDD

(b) Probability of DO < 2 mg/l (K=1)  
Global Euclidean metric

Using a local metric has an impact on predictions!
Two approaches to O2S2 over complex domains

Divide & conquer:

**PROS:** general validity,

**CONS:** no global model
Two approaches to O2S2 over complex domains

**Divide & conquer:**
- **PROS:** general validity,
- **CONS:** no global model

**PROS:** Derivation of global valid models
- **CONS:** restrictions on topology of the domain (stream networks)

Go to conclusion
Two approaches to O2S2 over complex domains

**Divide & conquer:**

**PROS:** general validity,
**CONS:** no global model

**PROS:** Derivation of global valid models
**CONS:** restrictions on topology of the domain (stream networks)
An environmental example: stream network domains

- **Problem.** Monitoring water temperatures
- **Data object.** Temperature profiles at 157 sites
- **Spatial domain.** Stream network

**Goal:** spatial statistical analysis, i.e., covariance modelling, prediction, from the functional data
Scalar geostatistics over stream networks

- Ver Hoef, Peterson and Theobald (2006) build valid models for scalar data over stream networks by using a **moving average construction, integrated piece-wise along the stream**

\[
Z(s) = \int_{-\infty}^{+\infty} g(x - s|\theta)dW(x)
\]
Scalar geostatistics over stream networks

- Ver Hoef, Peterson and Theobald (2006) build valid models for scalar data over stream networks by using a moving average construction, integrated piece-wise along the stream

$$Z(s) = \int_{-\infty}^{+\infty} g(x - s|\theta) dW(x)$$

- Depending on the shape of the moving average function $g$, different valid covariogram models are derived

| Name         | Moving Average Function $g(x|\theta)$ | Autocovariance Function $C(h|\theta)$ |
|--------------|--------------------------------------|----------------------------------------|
| Linear with Sill | $g(x|\theta) = \theta_1 \mathbb{I}_{(0 \leq x \leq \theta_r)}$ | $C(h|\theta) = \begin{cases} \theta_v \left(1 - \frac{h}{\theta_r}\right) & \text{if } 0 \leq h \leq \theta_r, \\ 0 & \text{if } h > \theta_r. \end{cases}$ |
| Spherical    | $g(x|\theta) = \theta_1 (1 - \frac{x}{\theta_r}) \mathbb{I}_{(0 \leq x \leq \theta_r)}$ | $C(h|\theta) = \begin{cases} \theta_v \left(1 - \frac{3h}{2\theta_r} + \frac{h^3}{2\theta_r^3}\right) & \text{if } 0 \leq h \leq \theta_r, \\ 0 & \text{if } h > \theta_r. \end{cases}$ |
| Exponential  | $g(x|\theta) = \theta_1 e^{-x} \mathbb{I}_{(0 \leq x)}$ | $C(h|\theta) = \theta_v e^{-\frac{h}{\theta_r}}$ if $0 \leq h$ |
| Mariah       | $g(x|\theta) = \theta_1 \frac{1}{h_{r} + 1} \mathbb{I}_{(0 \leq x)}$ | $C(h|\theta) = \begin{cases} \theta_v & \text{if } h = 0, \\ \theta_v \frac{\ln \frac{h_{r} + 1}{h} \theta_r}{h_{r} + 1} & \text{if } 0 < h. \end{cases}$ |

$$C_t(h|\theta) = \begin{cases} \int_{-\infty}^{+\infty} (g(x|\theta))^2 dx + \eta & h = 0 \\ \int_{-\infty}^{+\infty} g(x|\theta)g(x - h|\theta) dx & h > 0, \end{cases}$$
Scalar geostatistics over stream networks

- Ver Hoef, Peterson and Theobald (2006) build valid models for scalar data over stream networks by using a moving average construction, integrated piece-wise along the stream

\[ Z(s) = \int_{-\infty}^{+\infty} g(x - s|\theta) dW(x) \]

- Depending on the shape of the moving average function \( g \), different valid covariogram models are derived
- In particular, tail-up and tail-down models can be built
  
  Note: In tail-up models, flow-unconnected locations are uncorrelated

How to build these types of models for complex data?
Object-oriented moving average models in Hilbert spaces

- The projections of the process on an orthonormal basis are built as in the scalar case:

\[ X(s) = m + \sum_{k \geq 1} \xi_k(s)e_k. \]

\[ \xi_k(s) = \int_{-\infty}^{+\infty} g^{(k)}(x - s|\theta)dW_{k}(x), \]

*The process is well defined under regularity conditions on the \( g^{(k)} \)
Object-oriented moving average models in Hilbert spaces

- The projections of the process on an orthonormal basis are built as in the scalar case:

\[ X_s = m + \sum_{k \geq 1} \xi_k(s)e_k. \]

\[ \xi_k(s) = \int_{-\infty}^{+\infty} g^{(k)}(x - s | \theta) dW_k(x), \]

- The trace-covariogram/trace-variogram of the field are obtained as a nested structure, using as building blocks the scalar valid models. The trace-covariogram reads:

\[ C'(s_i, s_j) = \sum_{k \geq 1} C^{(k)}_d(s_i, s_j | \theta) \]

- Estimation of the trace-variogram follows similar lines as in the scalar case.

*The process is well defined under regularity conditions on the \( g^{(k)} \)
Object-oriented geostatistics over stream networks
Barbi, Menafoglio, Secchi (MOX rep. 32/2022)

- **Flow-unconnected stream distance empirical trace-semivariogram (FUSD)**
  
  Computed on flow-unconnected pairs (uncorrelated under tail-up model)

  \[
  \hat{\gamma}_{FUSD}(h_k) = \frac{1}{2|N(U_k)|} \sum_{(s_i, s_j) \in N(U_k)} \| X_{s_i} - X_{s_j} \|^2, \quad k = 1, \ldots, K_U,
  \]

- **Flow-connected stream distance empirical trace-semivariogram (FCSD)**
  
  Computed on flow-connected pairs, and used for variogram fitting

  \[
  \hat{\gamma}_{FCSD}(h_k) = \frac{1}{2|N(C_k)|} \sum_{(s_i, s_j) \in N(C_k)} \| X_{s_i} - X_{s_j} \|^2, \quad k = 1, \ldots, K_C,
  \]
Object-oriented geostatistics over stream networks

Barbi, Menafoglio, Secchi (MOX rep. 32/2022)

- **Flow-unconnected stream distance empirical trace-semivariogram (FUSD)**
  Computed on flow-unconnected pairs (uncorrelated under tail-up model)

\[
\hat{\gamma}_{FUSD}(h_k) = \frac{1}{2|N(U_k)|} \sum_{(s_i, s_j) \in N(U_k)} \left\| X_{s_i} - X_{s_j} \right\|^2, \quad k = 1, \ldots, K_U,
\]

- **Flow-connected stream distance empirical trace-semivariogram (FCSD)**
  Computed on flow-connected pairs, and used for variogram fitting

\[
\hat{\gamma}_{FCSD}(h_k) = \frac{1}{2|N(C_k)|} \sum_{(s_i, s_j) \in N(C_k)} \left\| X_{s_i} - X_{s_j} \right\|^2, \quad k = 1, \ldots, K_C,
\]

The FUSD can be used as a diagnostic tool
Example: Estimation of the covariance structure for the Middle Fork river data

Data: smoothed daily temperature profiles 15 July to 31 Aug. 2005 at 157 locations along the Middle Fork river
Object-oriented geostatistics over stream networks

Barbi, Menafoglio, Secchi (MOX rep. 32/2022)

Example: Estimation of the covariance structure for the Middle Fork river data

Data: smoothed daily temperature profiles 15 July to 31 Aug. 2005 at 157 locations along the Middle Fork river

a) Empirical estimator after detrending (tail-up model)
b) Fitted model (exponential family)
Object-oriented geostatistics over stream networks

Barbi, Menafoglio, Secchi (MOX rep. 32/2022)

Example: Estimation of the covariance structure for the Middle Fork river data

The estimated valid model can be used for **spatial prediction**, e.g., by kriging (but based on the stream topology to compute distances)

**Example:** kriging at \( s_0 \)

\[
\chi^*_0 = \sum_{i=1}^{n} \lambda_i^* \cdot \chi_{s_i}
\]

where the weights solve

\[
\begin{bmatrix}
\gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) & 1 \\
1 & \cdots & 1 & 0
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\
\vdots \\
\lambda_n \\
\zeta
\end{bmatrix} =
\begin{bmatrix}
\gamma(s_0, s_1) \\
\vdots \\
\gamma(s_0, s_n) \\
1
\end{bmatrix}
\]

Data: smoothed daily temperature profiles 15 July to 31 Aug. 2005 at 157 locations along the Middle Fork river

a) Empirical estimator after detrending (tail-up model)
b) Fitted model (exponential family)
Results of Kriging

**Prediction error**

\[ SSE_i = \| x_{s_i} - x_{s_i}^* \|^2, \quad i = 1, \ldots, n, \quad SSE_i^{(rel.)} = \frac{SSE_i}{\| x_{s_i} \|^2} \]

<table>
<thead>
<tr>
<th></th>
<th>$SSE$</th>
<th>$SSE^{(rel)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.574</td>
<td>$3.44 \cdot 10^{-5}$</td>
</tr>
<tr>
<td>Median</td>
<td>43.346</td>
<td>$4.19 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>Mean</td>
<td>132.897</td>
<td>$1.6 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>Sum</td>
<td>20864.89</td>
<td>2.52</td>
</tr>
</tbody>
</table>

Fig 8: SSE leave-one-out error for each location on the Middle Fork River

Table 2: Summary indices of the distribution of $SSE$ and $SSE^{(rel)}$

Kriging results, assessed via LOO-CV, show a good ability of the predictor based on stream distances to reproduce the data.
Outline

➢ Introduction and Motivations

➢ An overview on O2S2

➢ Modeling object data over complex domains
  • A divide-and-conquer approach
  • The case of stream network domains

➢ Concluding remarks
Concluding remarks

Object Oriented Spatial Statistics (O2S2)

• a mathematical framework for the spatial statistical analysis of complex data
• aiming to modeling, prediction and UQ, possibly accounting prior knowledge on the field, secondary data
• distributed over potentially complex domains
Concluding remarks

A key role for the embedding space

• Not just $L^2$ (space of square integrable functions)!
• Can account for data constraints: Bayes Hilbert space for PDF data
A key role for the embedding space

- Not just $L^2$ (space of square integrable functions)
- Can account for data constraints: Bayes Hilbert space for PDF data
- Can be non-linear: Riemannian manifolds for tensor data

O2S2 in Riemannian manifolds

Pigoli, Menafoglio, Secchi (JMVA 2016, JCGS 2021)
Concluding remarks

A key role for the embedding space

• Not just $L^2$ (space of square integrable functions)
• Can account for data constraints: Bayes Hilbert space for PDF data
• Can be non-linear: Riemannian manifolds for tensor data

**O2S2 in Riemannian manifolds**

Temp.-prec. covariance

PD(p): Riemannian manifold Modeling in the linear (Hilbert) tangent space $\text{Sym}(p)$

Pigoli, Menafoglio, Secchi (JMVA 2016, JCGS 2021)
Concluding remarks

O2S2 for meta-modeling and UQ in numerical models

- Dependence may arise from proximity in a parameters space
- Kriging and cokriging for meta-modeling and multi-fidelity modeling

O2S2 for statistical metamodeling

Grugjic, Menafoglio, Yang, Caers (SERRA 2018)
Selected references


R packages

- fdagstat. O. Grujic & A. Menafoglio, [https://github.com/ogru/fdagstat](https://github.com/ogru/fdagstat)
- manifoldgstat, L. Torriani & I. Sartori, [https://github.com/LucaTorriani/KrigingManifoldData](https://github.com/LucaTorriani/KrigingManifoldData)