MOX–Report No. 34/2012

A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space

Menafoglio, A.; Dalla Rosa, M.; Secchi, P.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space

Alessandra Menafoglio[a], Matilde Dalla Rosa[b], Piercesare Secchi[a]

[a]MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy
[b]Eni S.p.A. Divisione E & P, San Donato M.se, Italy
alessandra.menafoglio@mail.polimi.it
matilde.dalla.rosa@eni.com
piercesare.secchi@polimi.it

Keywords: Functional data analysis, Spatial prediction, Variogram, Sobolev metrics.

Abstract

We address the problem of predicting spatially dependent functional data belonging to a Hilbert space, with a Functional Data Analysis approach. Having defined new global measures of spatial variability for functional random processes, we derive a Universal Kriging predictor for functional data. Consistently with the new established theoretical results, we develop a two-step procedure for predicting georeferenced functional data: first model selection and estimation of the spatial mean (drift), then Universal Kriging prediction on the basis of the identified dichotomy model, sum of deterministic drift and stochastic residuals. The proposed methodology is tested by means of a simulation study and finally applied to daily mean temperatures curves aiming at reconstructing the space-time field of temperatures of Canada's Maritimes Provinces.

1 Introduction

Functional Data Analysis (FDA, Ramsay and Silverman [2005]) has recently received a great deal of attention in the literature because of the increasing need to analyze infinite-dimensional data, such as curves, surfaces and images. Whenever functional data are spatially dependent, FDA methods relying on the assumption of independence among observations could fail because consistency problems may arise [Hörmann and Kokoszka, 2011].

In the presence of spatial dependence, not only *ad hoc* estimation and regression techniques need to be developed (e.g., Gromenko et al. [2012] and Yamanishi and Tanaka [2003]), but also other topics need to be faced. Among them, spatial prediction assumes a key role: the extension of kriging techniques [Cressie, 1993] to the functional setting meets the need of interpolating complex data collected in a limited number of spatial locations and thus could find application in different areas of industrial and environmental research. Nevertheless, little literature has been produced on this topic: indeed, theoretical results in this direction have been recently derived (Giraldo et al. [2008b], Giraldo et al. [2010a], Delicado et al. [2010], Giraldo et al. [2010b] and Monestiez and Nerini [2008]) but this theory is still limited to stationary functional stochastic processes valued in L^2.

However, in geophysical and environmental applications, natural phenomena are typically very complex and they rarely show a uniform behavior over the spatial domain: in these cases, non-stationary methods are needed, but, to the best of our knowledge, a non-stationary kriging methodology for functional data has yet to be developed. In this work, we tackle this problem both from a theoretical point of view and from a computational one.

The methodological effort is here devoted to establish a general and coherent theoretical framework for Universal Kriging prediction in any separable Hilbert space, not just L^2. For instance, in our setting both pointwise and differential properties characterizing the functional data can be explicitly incorporated in the measures of spatial dependence –namely trace-variogram and trace-covariogram– if data are assumed to belong to a proper Sobolev space (see Remark 5 and Subsection 4.2).

Together with the theoretical results –presented in Section 2–, new algorithms to perform spatial prediction are developed in Section 3, while their performance is tested through a simulation study in Section 4. Two main goals move this part of the work: first to select an optimal linear model for the spatial mean –i.e. the drift– in the absence of a priori information, second to estimate the structure of spatial dependence of the associated residuals, which is that involved in the kriging prediction.

Finally, the case study that first motivated this work is presented in Section 5. It originates from a meteorological application and concerns the analysis of daily mean temperature curves recorded in the Maritimes Provinces of Canada. The aim of the study is to predict the whole space-time field of temperatures on the basis of the available data, deriving furthermore estimates for the temperature spatial trend. The problem of spatial prediction of temperatures is of interest in microclimate prediction as well as in hydrological and forest ecosystem modeling. It has been already faced in the literature about kriging for functional data by means of stationary techniques (e.g., [Giraldo et al., 2010b]); here a non-Euclidean distance is adopted for the spatial domain and a drift term is modeled. We will show that the introduction of a drift term has a strong influence on the analysis in terms of cross-validation performance and prediction accuracy, besides allowing to deduce a climatical interpretation of the results.

2 Universal Kriging for Functional Random Fields

2.1 Preliminaries and definitions

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ a probability space and H a separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$, whose points are functions $\mathcal{X} : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a compact subset of $\mathbb{R}$. Call *functional random variable* a measurable function $\boldsymbol{\mathcal{X}} : \Omega \to H$, whose realization $\mathcal{X}$, called *functional data*, is an element of H [Ferraty and Vieu, 2006].

Consider on H a (functional) random field:

$$\{\boldsymbol{\chi_s}, \boldsymbol{s} \in D \in \mathbb{R}^d\}, \tag{1}$$

that is a set of functional random variables $\boldsymbol{\chi_s}$ of H, indexed by a continuous spatial vector $\boldsymbol{s}$ varying in $D \in \mathbb{R}^d$ (usually $d = 2$).

In this framework, a functional dataset $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$ is the collection of n observations of the random field (1) relative to n locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n \in D$; in non-trivial situations a vector of observations $\chi_{\boldsymbol{s}} = (\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n})^T$ is characterized by a structure of spatial dependence reflecting the covariance structure of the generating random process (1).

The aim of this work is the prediction of the realization $\chi_{\boldsymbol{s}_0}$ in an unsampled site $\boldsymbol{s}_0 \in D$, through a geostatistical approach, based on global definitions of covariogram and variogram.

For $1 \leq p < \infty$ denote with $L^p(\Omega; H)$ the vector space of (equivalence classes of) measurable functions $\mathcal{X} : \Omega \to H$ with $\|\mathcal{X}\| \in L^p(\Omega)$ — i.e. $\int_\Omega \|\mathcal{X}(\omega)\|^p \mathbb{P}(d\omega) = \mathbb{E}[\|\mathcal{X}\|^p] < \infty$ where $\mathbb{E}$ indicates the expected value —, that is a Banach space with respect to the norm:

$$\|\mathcal{X}\|_{L^p(\Omega;H)} := \left(\int_\Omega \|\mathcal{X}(\omega)\|^p \mathbb{P}(d\omega) \right)^{1/p} = (\mathbb{E}[\|\mathcal{X}\|^p])^{1/p}.$$

In this work, we assume that the following condition holds.

Assumption 1 (Square-integrability)**.** *Each element* $\boldsymbol{\chi_s}$*,* $\boldsymbol{s} \in D$*, of the random field* (1) *belongs to* $L^2(\Omega; H)$*.*

When Assumption 1 is true, the expected value $m_{\boldsymbol{s}}$ of the random field (1) can be defined as:

$$m_{\boldsymbol{s}} = \int_\Omega \boldsymbol{\chi_s}(\omega) \mathbb{P}(d\omega), \quad \boldsymbol{s} \in D.$$

A global measure of spatial dependence can be provided defining the *(global) covariance function* $C : D \times D \to \mathbb{R}$ as the function mapping each pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in D$ into:

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{s_j}) := \mathbb{E}[\langle \boldsymbol{\chi}_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{\boldsymbol{s}_j} - m_{\boldsymbol{s}_j} \rangle], \tag{2}$$

which is well-defined when Assumption 1 is true.

In particular, the covariance function (2) is a positive definite function:

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

and defines a scalar product on $L^2(\Omega; H)$.

The function C will be called *trace-covariogram* because of its relation –for every fixed $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$– with the cross-covariance operator $C_{\boldsymbol{s}_i,\boldsymbol{s}_j} : H \to H$ defined, for $x \in H$, by:

$$C_{\boldsymbol{s}_i,\boldsymbol{s}_j} x = \mathbb{E}[\langle \boldsymbol{\chi}_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_i}, x\rangle (\boldsymbol{\chi}_{\boldsymbol{s}_j} - m_{\boldsymbol{s}_j})]. \tag{3}$$

As proved in [Bosq, 2000], the operator $C_{\boldsymbol{s}_i,\boldsymbol{s}_j}$ is a trace-class Hilbert-Schmidt operator. Moreover, by applying Parsival Identity and following the arguments presented in (Hörmann and Kokoszka [2011], Section 3), one can easily prove the following:

Proposition 2. *For every couple of locations $\boldsymbol{s}_i, \boldsymbol{s}_j$ in D, $C(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the trace of the corresponding cross-covariance operator C_{s_i,s_j}:*

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sum_{k=1}^{\infty} \langle C_{\boldsymbol{s}_i,\boldsymbol{s}_j} e_k, e_k\rangle, \tag{4}$$

where $\{e_k, k \in \mathbb{N}\}$ is an orthonormal basis of H. In particular:

$$|C(\boldsymbol{s}_i, \boldsymbol{s}_j)| \leq \sum_{k=1}^{\infty} |\lambda_k^{(\boldsymbol{s}_i,\boldsymbol{s}_j)}|,$$

being $\lambda_k^{(\boldsymbol{s}_i,\boldsymbol{s}_j)}$, $k = 1, 2, ...$, the singular values of the cross-covariance operator C_{s_i,s_j}.

Notice that, the trace of C_{s_i,s_j} is well defined by $\sum_{k=1}^{\infty} \langle C_{\boldsymbol{s}_i,\boldsymbol{s}_j} e_k, e_k\rangle$, since this series converges absolutely for any orthonormal basis $\{e_k, k \geq 1\}$ of H and the sum does not depend on the choice of the orthonormal basis (Zhu [2007], Theorem 1.24).

Expression (2) induces a notion of global variance and of variogram, as well as new global definitions of second-order and intrinsic stationarity.

Definition 3. The *(global) variance* of the process (1) is the function $\sigma^2 : D \to [0, +\infty]$:

$$\sigma^2(\boldsymbol{s}) = \mathrm{Var}(\boldsymbol{\chi}_{\boldsymbol{s}}) = \mathbb{E}[\|\boldsymbol{\chi}_{\boldsymbol{s}} - m_{\boldsymbol{s}}\|^2], \quad \boldsymbol{s} \in D. \tag{5}$$

The *trace-semivariogram* of the process (1) is the function $\gamma : D \times D \to [0, +\infty]$:

$$\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{1}{2} \mathrm{Var}(\boldsymbol{\chi}_{\boldsymbol{s}_i} - \boldsymbol{\chi}_{\boldsymbol{s}_j}), \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{6}$$

Definition 4. A functional spatial random field $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D \subset \mathbb{R}^d\}$ is said to be *strongly stationary* if for every $\boldsymbol{h} \in D$, $k \geq 1$ and every collection $\boldsymbol{s}_1, ..., \boldsymbol{s}_k \in D$:

$$(\boldsymbol{\chi}_{\boldsymbol{s}_1}, \boldsymbol{\chi}_{\boldsymbol{s}_2}, ..., \boldsymbol{\chi}_{\boldsymbol{s}_k}) \sim (\boldsymbol{\chi}_{\boldsymbol{s}_1+\boldsymbol{h}}, \boldsymbol{\chi}_{\boldsymbol{s}_2+\boldsymbol{h}}, ..., \boldsymbol{\chi}_{\boldsymbol{s}_k+\boldsymbol{h}}). \tag{7}$$

A process $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D \in \mathbb{R}^d\}$ is said to be *(globally) second-order stationary* if the following conditions hold:

(i) $\mathbb{E}[\boldsymbol{\chi}_{\boldsymbol{s}}] = m, \quad \forall\, \boldsymbol{s} \in D \subseteq \mathbb{R}^d$;

(ii) $\mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \boldsymbol{\chi}_{\boldsymbol{s}_i} - m, \boldsymbol{\chi}_{\boldsymbol{s}_j} - m \rangle] = C(\boldsymbol{h}), \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

A process $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D \in \mathbb{R}^d\}$ is said to be *(globally) intrinsically stationary* if:

(i') $\mathbb{E}[\boldsymbol{\chi}_{\boldsymbol{s}}] = m, \quad \forall\, \boldsymbol{s} \in D \subseteq \mathbb{R}^d$;

(ii') $\mathrm{Var}(\boldsymbol{\chi}_{\boldsymbol{s}_i} - \boldsymbol{\chi}_{\boldsymbol{s}_j}) = \mathbb{E}[\|\boldsymbol{\chi}_{\boldsymbol{s}_i} - \boldsymbol{\chi}_{\boldsymbol{s}_j}\|^2] = \gamma(\boldsymbol{h}), \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

Function (6) has the same properties as its finite-dimensional analogue [Chilès and Delfiner, 1999]; in particular the trace-semivariogram is a conditionally negative definite function:

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

Remark 5. *When $H = L^2$ and global second-order stationarity and isotropy for the process* (1) *is in force, the trace-semivariogram* (6) *corresponds to the integrated version of pointwise variograms $\gamma(h_{i,j}; t) = \frac{1}{2}\mathrm{Var}(\boldsymbol{\chi}_{\boldsymbol{s}_i}(t) - \boldsymbol{\chi}_{\boldsymbol{s}_j}(t))$ (assumed to exist a.e.):*

$$\gamma(h_{i,j}) = \int_{\mathcal{T}} \gamma(h_{i,j}; t) dt. \tag{8}$$

This has been introduced in [Giraldo et al., 2008a] with the name of trace-semivariogram. *However our definition is more general and permits the analysis of functional data in more complex situations. For instance, we might want to take explicitly into account the regularity of the elements of H –which captures the dependence along the coordinate $t \in \mathcal{T}$– by assuming that H is an appropriate Sobolev space and working with the inner product consistent with this assumption.*

In particular, let $\mathcal{H}^k, k \geq 1$, be the subset of L^2 consisting of the equivalence classes of functions with weak derivatives $D^\alpha \mathcal{X}$, $\alpha \leq k$, in L^2:

$$\mathcal{H}^k(\mathcal{T}) = \{\mathcal{X} : \mathcal{T} \to \mathbb{R},\ s.t.\ D^\alpha \mathcal{X} \in L^2, \forall \alpha \leq k,\ \alpha \in \mathbb{N}\}.$$

By considering on $\mathcal{H}^k$ the usual inner product and norm, the resulting trace-variogram is ($D^0 \boldsymbol{\chi}_{\boldsymbol{s}} = \boldsymbol{\chi}_{\boldsymbol{s}}$):

$$2\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Var}(\boldsymbol{\chi}_{\boldsymbol{s}_i} - \boldsymbol{\chi}_{\boldsymbol{s}_j})_{\mathcal{H}^k} = \sum_{\alpha=0}^{k} \mathrm{Var}(D^\alpha \boldsymbol{\chi}_{\boldsymbol{s}_i} - D^\alpha \boldsymbol{\chi}_{\boldsymbol{s}_j})_{L^2},$$

where $\mathrm{Var}(D^\alpha \boldsymbol{\chi}_{\boldsymbol{s}_i} - D^\alpha \boldsymbol{\chi}_{\boldsymbol{s}_j})_{L^2}$ *are the trace-variograms in* L^2 *relative to the weak derivative random fields* $\{D^\alpha \boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, $0 \leq \alpha \leq k$ *(assumed to exist, for every* $\alpha = 0, 1, ..., k$*), which might significantly influence the overall trace-variogram.*

The choice of the most proper Sobolev space may allow to distinguish among functional random fields which might appear similar from a spatial dependence point of view, but indeed very different in the structure of dependence along the coordinate $t \in \mathcal{T}$ *(see Subsection 4.2).*

Moreover, suppose the random field to be the random path of a stochastic dynamical system, $\{\boldsymbol{\chi}_\tau, \tau \in \mathcal{D} \subset \mathbb{R}\}$*, whose state* $\boldsymbol{\chi}_\tau$ *is a functional random variable belonging to a Sobolev space* H *–determined by the equations which govern the dynamics of the system– [Arnold, 2003]. In dynamical system theory, the Sobolev norm of the state coincides with (twice) the energy of the system. Therefore, the choice of the most proper Sobolev space for geostatistical analysis implies a precise physical meaning for the measure of stochastic variability: indeed, the global variance represents (twice) the mean energy of the system, while the trace-variogram (twice) the mean energy of the increments between two states.*

In the light of the Proposition 2, existence of strong, second-order and intrinsic stationary functional processes can be established by direct construction as in [Hörmann and Kokoszka, 2011]. Considering a basis $\{e_j, j \geq 1\}$ of H, every functional random process (1) with constant mean m admits the following expansion:

$$\boldsymbol{\chi}_{\boldsymbol{s}} = m + \sum_{j \geq 1} \xi_j(\boldsymbol{s}) e_j. \tag{9}$$

The scalar fields $\xi_j(\boldsymbol{s}) = \langle \boldsymbol{\chi}_{\boldsymbol{s}} - m, e_j \rangle$, $j = 1, 2, ...$, determine the stationarity of the functional process. In fact, process (1) is strong stationary if and only if the scalar fields $\xi_j(\boldsymbol{s})$ are strictly stationary for all $j \geq 1$; moreover the random element $\boldsymbol{\chi}_{\boldsymbol{s}}$, $s \in D$, belongs to $L^2(\Omega; H)$ if and only if the sequence $\{\xi_j(\boldsymbol{s})\}_{j \geq 1}$ belongs to $\ell^2(\Omega; \mathbb{R})$ (i.e. $\sum_{j \geq 1} \mathbb{E}[\xi_j(\boldsymbol{s})] < \infty$), [Hörmann and Kokoszka, 2011]. Furthermore, second-order stationarity of each scalar field $\xi_j(\boldsymbol{s})$, $j = 1, 2, ...$, ensures that the cross-covariance operator $C_{\boldsymbol{s}, \boldsymbol{s}+\boldsymbol{h}}$ depends only on the increment vector $\boldsymbol{h} \in D$, for every $\boldsymbol{s} \in D$, which is in fact a sufficient condition for the functional process to be global second-order stationary. This condition can be weakened in order to obtain the following necessary and sufficient condition for global second-order stationarity:

$$\sum_{j \geq 1} \mathbb{E}[\xi_j(\boldsymbol{s}), \xi_j(\boldsymbol{s} + \boldsymbol{h})] = C(\boldsymbol{h}), \tag{10}$$

for each $\boldsymbol{s}, \boldsymbol{h} \in D$ for some real-valued function C. As a consequence, a necessary condition for global second-order stationarity is the independence of the ℓ^2-norm of the sequence $\{\xi_j(\boldsymbol{s})\}_{j \geq 1}$ from the location $\boldsymbol{s} \in D$.

As in finite-dimensional theory, intrinsic stationarity is a weaker condition than second-order stationarity. Indeed, a d-dimensional isotropic Brownian motion $\{W_{\boldsymbol{s}}, \boldsymbol{s} \in D \subseteq \mathbb{R}^d\}$ can be seen as a functional random field on $H =$

$L^2([0,1])$, such that each element $\mathcal{W}_{\boldsymbol{s}} : [0,1] \to \mathbb{R}$, $\boldsymbol{s} \in D$, is a functional random variable whose realization $\mathcal{W}_{\boldsymbol{s}}(\omega, \cdot)$, $\omega \in \Omega$, is constant over the domain $[0,1]$: $\mathcal{W}_{\boldsymbol{s}}(\omega, t) = W_s(\omega)$, for all $t \in [0,1]$. Obviously, each $\mathcal{W}_s$ belongs to $L^2(\Omega, H)$ and:

$$\mathrm{Var}(\mathcal{W}_{\boldsymbol{s}_i} - \mathcal{W}_{\boldsymbol{s}_j}) = \mathbb{E}[(W_{\boldsymbol{s}_i} - W_{\boldsymbol{s}_j})] = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|,$$

while

$$\mathrm{Cov}(\mathcal{W}_{\boldsymbol{s}_i}, \mathcal{W}_{\boldsymbol{s}_j}) = \mathbb{E}[W_{\boldsymbol{s}_i} W_{\boldsymbol{s}_j}] = (\|\boldsymbol{s}_i\| + \|\boldsymbol{s}_j\| - \|\boldsymbol{s}_i - \boldsymbol{s}_j\|),$$

which is not a function of $(\boldsymbol{s}_i - \boldsymbol{s}_j)$.

Finally, the condition of isotropy can be established as follow.

Definition 6. A second-order stationary random process is said to be *isotropic* if:

$$\mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}_i}, \boldsymbol{\chi}_{\boldsymbol{s}_j}) = C(\|\boldsymbol{h}\|), \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j,$$

where $\|\cdot\|$ is a norm on D.

2.2 Universal Kriging predictor

Consider a non-stationary random process $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, whose elements are representable as:

$$\boldsymbol{\chi}_{\boldsymbol{s}} = m_{\boldsymbol{s}} + \boldsymbol{\delta}_{\boldsymbol{s}}, \tag{11}$$

where m_s, the drift, describes the non-constant spatial mean variation, while the residual term $\boldsymbol{\delta}_{\boldsymbol{s}}$ is supposed to be a zero-mean, second-order stationary and isotropic random field, i.e.:

$$\begin{cases} \mathbb{E}[\boldsymbol{\chi}_{\boldsymbol{s}}] = m_{\boldsymbol{s}}, & \boldsymbol{s} \in D \subseteq \mathbb{R}^d; \\ \mathbb{E}[\boldsymbol{\delta}_{\boldsymbol{s}}] = 0, & \boldsymbol{s} \in D \subseteq \mathbb{R}^d; \\ \mathrm{Cov}(\boldsymbol{\delta}_{\boldsymbol{s}_i}, \boldsymbol{\delta}_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \boldsymbol{\delta}_{\boldsymbol{s}_i}, \boldsymbol{\delta}_{\boldsymbol{s}_j} \rangle] = C(\|\boldsymbol{h}\|), & \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j. \end{cases}$$

As in classical geostatistics [Cressie, 1993], assume the following linear model for the drift m_s:

$$m_{\boldsymbol{s}}(t) = \sum_{l=0}^{L} a_l(t) f_l(\boldsymbol{s}), \quad \boldsymbol{s} \in D,\ t \in \mathcal{T}, \tag{12}$$

where $f_0(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$, $f_l(\cdot)$, $l = 1, ..., L$, are known functions of the spatial variable $\boldsymbol{s} \in D$ and $a_l(\cdot) \in H$, $l = 0, ..., L$, are functional coefficients independent from the spatial location. Hence it is supposed that the dependence of the mean $m_{\boldsymbol{s}}$ on the spatial variable $\boldsymbol{s} \in D$ is explained by the family $\{f_l(\cdot)\}_{l=1,...,L}$, that is scalar with respect to the variable $t \in \mathcal{T}$; in the meantime, the functional nature of the drift $m_{\boldsymbol{s}}$ is preserved thanks to the introduction of the functional coefficients $a_l(\cdot)$, $l = 0, ..., L$.

For most applications, these assumptions are not too restrictive: in fact this model is able to describe precisely the drift term whenever it is a separable function or in the presence of a scalar external drift.

Given n observations $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$ sampled from a realization of $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, our next goal is the formulation of the Universal Kriging predictor of the variable $\boldsymbol{\chi}_{\boldsymbol{s}_0}$ located in $\boldsymbol{s}_0 \in D$, which is the best linear unbiased predictor (BLUP):

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

whose weights $\lambda_1^*, ..., \lambda_n^* \in \mathbb{R}$ minimize the global variance of the prediction error under the unbiasedness constraint:

$$(\lambda_1^*, ..., \lambda_n^*) = \underset{\substack{\lambda_1,...,\lambda_n \in \mathbb{R}: \\ \boldsymbol{\chi}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} = \sum_{i=1}^n \lambda_i \boldsymbol{\chi}_{\boldsymbol{s}_i}}}{\operatorname{argmin}} \operatorname{Var}(\boldsymbol{\chi}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} - \boldsymbol{\chi}_{\boldsymbol{s}_0}) \quad \text{s.t.} \quad \mathbb{E}[\boldsymbol{\chi}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0}] = m_{\boldsymbol{s}_0}. \tag{13}$$

In (13) both the variance to be minimized and the unbiasedness constraint are well defined since the linear predictor $\boldsymbol{\chi}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0}$ (and thus $\boldsymbol{\chi}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} - \boldsymbol{\chi}_{\boldsymbol{s}_0}$) belongs to the same space H as the variables $\boldsymbol{\chi}_{\boldsymbol{s}_1}, ..., \boldsymbol{\chi}_{\boldsymbol{s}_n}$, because H is closed with respect to linear combinations of its elements.

From the unbiasedness constraint, the following set of restrictions on the weights can be easily derived:

$$\sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall\, l = 0, ..., L. \tag{14}$$

By including (14) in the minimization problem through $L+1$ Lagrange multipliers, $\mu_0, ..., \mu_L$, problem (13) can be solved by minimizing the functional Φ:

$$\Phi = \operatorname{Var}(\boldsymbol{\chi}^*_{\boldsymbol{s}_0} - \boldsymbol{\chi}_{\boldsymbol{s}_0}) + 2 \sum_{l=0}^{L} \mu_l \left(\sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0) \right),$$

easily reduced to:

$$\Phi = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(\boldsymbol{s}_i, \boldsymbol{s}_j) + C(0) - 2 \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_i, \boldsymbol{s}_0) + 2 \sum_{l=0}^{L} \mu_l \left(\sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0) \right). \tag{15}$$

Under suitable assumptions on the sampling design –namely $\Sigma = (C(h_{i,j})) \in \mathbb{R}^{n \times n}$ definite positive and $\mathbb{F}_{\boldsymbol{s}} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times (L+1)}$ of full rank–, the functional (15) admits a unique global minimum that can be found solving the following

linear system:

$$\begin{pmatrix} C(0) & \cdots & C(h_{1,n}) & 1 & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C(h_{n,1}) & \cdots & C(0) & 1 & f_1(\boldsymbol{s}_n) & \cdots & f_L(\boldsymbol{s}_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(\boldsymbol{s}_1) & \cdots & f_1(\boldsymbol{s}_n) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_L(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_n) & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} C(h_{0,1}) \\ \vdots \\ C(h_{0,n}) \\ 1 \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{pmatrix}, \tag{16}$$

where $C(h_{i,j})$ denotes the trace-covariogram function of the residual process $\{\boldsymbol{\delta}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, evaluated in $h_{i,j} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$.

Moreover, since:

$$\gamma(h_{i,j}) = C(0) - C(h_{i,j}),$$

–an extension to the functional case of a well-known result in geostatistics for real-valued processes which can be easily derived combining (2) and (6)– the linear system (16) can also be expressed in terms of the semivariogram function γ as:

$$\begin{pmatrix} \gamma(0) & \cdots & \gamma(h_{1,n}) & 1 & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(h_{n,1}) & \cdots & \gamma(0) & 1 & f_1(\boldsymbol{s}_n) & \cdots & f_L(\boldsymbol{s}_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(\boldsymbol{s}_1) & \cdots & f_1(\boldsymbol{s}_n) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_L(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_n) & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} \gamma(h_{0,1}) \\ \vdots \\ \gamma(h_{0,n}) \\ 1 \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{pmatrix}. \tag{17}$$

In addition, we can associate to the pointwise prediction $\boldsymbol{\chi}^*_{\boldsymbol{s}_0}$ in $\boldsymbol{s}_0$ a measure of its global variability through the *universal kriging variance*, defined as:

$$\begin{aligned} \sigma^2_{UK}(\boldsymbol{s}_0) &= C(0) - \sum_{i=1}^{n} \lambda_i C(h_{i,0}) - \sum_{l=0}^{L} \mu_l f_l(\boldsymbol{s}_0) = \\ &= \sum_{i=1}^{n} \lambda_i \gamma(h_{i,0}) + \sum_{l=0}^{L} \mu_l f_l(\boldsymbol{s}_0), \quad \boldsymbol{s}_0 \in D;\ f_0(\boldsymbol{s}) = 1,\ \forall \boldsymbol{s} \in D. \end{aligned} \tag{18}$$

Observe that both kriging systems (16) and (17), as well as the kriging variance (18), have exactly the same form of the finite-dimensional corresponding expressions, indicating the consistency of our extensions with the real-valued random field case (if $H = \mathbb{R}$, the trace-covariogram reduces to the usual covariogram).

Moreover, by considering the very specific case treated in Remark 5 the Universal Kriging system (17) reduces to the Ordinary Kriging system already presented in [Giraldo et al., 2008a].

Finally, global second-order stationarity of the residual process has been assumed for the construction of the optimal predictor: however, as in classical theory, the Ordinary Kriging predictor is also well defined under the hypothesis of intrinsic stationarity. In fact, second-order stationarity for the residuals has to be required whenever the mean $m_{\boldsymbol{s}}$ is not constant and the residual trace-variogram is unknown: in such a case, although the trace-covariogram is not directly involved in the kriging system (17), it is needed for the generalized least squares estimate of the drift (see Subsection 2.4).

2.3 Variogram Estimation

In order to determine the universal kriging predictor in $\boldsymbol{s}_0$ by solving (16) or (17), an estimation of the trace-covariogram or, as usually preferred, of the trace-semivariogram is needed.

As in classical geostatistics, the variogram estimation can be performed in two steps: determination of an empirical estimator and fitting of a variogram valid model. The latter step is necessary in order to fulfil the requirements on the trace-variogram function, e.g. conditional negative definiteness.

Suppose to know the realization $\delta_{\boldsymbol{s}_1}, ..., \delta_{\boldsymbol{s}_n}$ of the residual process $\{\boldsymbol{\delta}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, in the n sampling locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ of the domain D in which we observe the functional dataset $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$. Recall that the residual process is zero-mean second-order stationary and isotropic, so that:

$$\gamma(h) = \mathbb{E}[\|\boldsymbol{\delta}_{\boldsymbol{s}_i} - \boldsymbol{\delta}_{\boldsymbol{s}_j}\|^2], \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|.$$

Following the approach adopted in [Giraldo et al., 2008a] and proceeding by analogy with the finite-dimensional case, a method-of-moments estimator can be used:

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \|\boldsymbol{\delta}_{\boldsymbol{s}_i} - \boldsymbol{\delta}_{\boldsymbol{s}_j}\|^2, \tag{19}$$

where $N(h)$ indicates the set of all couples of sites separated by a distance h and $|N(h)|$ is its cardinality. In applications, since it is hardly possible to calculate an estimate $\widehat{\gamma}(h)$ for every value of h, a discretized version $\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_K))$ of $\widehat{\gamma}(h)$ can be used instead, defining K classes of distance and calculating for each $k = 1, ..., K$:

$$\widehat{\gamma}(h_k) = \frac{1}{2|N(h_k)|} \sum_{(i,j)\in N(h_k)} \|\boldsymbol{\delta}_{\boldsymbol{s}_i} - \boldsymbol{\delta}_{\boldsymbol{s}_j}\|^2. \tag{20}$$

For the fitting step, a least squares criterion can be used, minimizing the distance between the empirical estimate $\widehat{\gamma}(\boldsymbol{h})$ and a parametric valid model $\gamma(\boldsymbol{h}; \boldsymbol{\vartheta})$,

properly chosen among the classical families of valid variogram models [Cressie, 1993]. Indeed, in classical geostatistics there exists a number of parametric families of valid variogram models that can be used in the functional case as well, since the trace-variogram is a real valued function which has to fulfil the same set of requirements as its finite-dimensional analogue. As an alternative, ad hoc constructed valid models can be tested for conditional negative definiteness by means of spectral methods [Armstrong and Diamond, 1984].

2.4 Drift Estimation

Although the drift coefficients are not directly included in the Universal Kriging systems (16) and (17), their estimation is necessary in order to assess the trace-variogram of the residual process $\{\boldsymbol{\delta_s}, \boldsymbol{s} \in D\}$, since, in general, this is unobserved.

Assuming the dichotomy (11) and the linear model (12), the original process can be expressed as:

$$\boldsymbol{\chi_s} = \sum_{l=0}^{L} a_l f_l(s) + \boldsymbol{\delta_s}, \quad \boldsymbol{s} \in D. \tag{21}$$

Hence, the compact matrix form for model (21) for the random vector $\boldsymbol{\chi_{\vec{s}}} = (\boldsymbol{\chi_{s_1}}, ..., \boldsymbol{\chi_{s_n}})$ –whose realization $\chi_{\vec{s}}$ belongs to the product space $H^n = H \times H \times \cdots \times H$– is:

$$\boldsymbol{\chi_{\vec{s}}} = \mathbb{F}_{\vec{s}} a_{\vec{l}} + \boldsymbol{\delta_{\vec{s}}}, \tag{22}$$

where $a_{\vec{l}} = (a_0, ..., a_L)$ is the vector of coefficients, $\boldsymbol{\delta_{\vec{s}}} = (\boldsymbol{\delta_{s_1}}, ..., \boldsymbol{\delta_{s_n}})$ is the random vector of spatially-correlated residuals and $\mathbb{F}_{\vec{s}}$ is the design matrix:

$$\mathbb{F}_{\vec{s}} = \begin{pmatrix} 1 & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ 1 & f_1(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_2) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & f_1(\boldsymbol{s}_n) & \cdots & f_L(\boldsymbol{s}_n) \end{pmatrix}.$$

The theory of linear models in functional data analysis (FDA) Ramsay and Silverman [2005] has been developed under the founding hypothesis of independent and identically distributed residuals, so that the ordinary least squares approach developed in that framework inevitably turns out to be somewhat inadequate in the presence of correlated residuals.

In order to properly take into account the structure of spatial dependence existing among observations, we propose a generalized least squares criterion (GLS) with weighting matrix Σ^{-1}, the inverse of the $n \times n$ covariance matrix Σ of $\boldsymbol{\chi_{\vec{s}}}$.

Indeed, a measure of the statistical distance among functional random variables $\mathcal{X}, \mathcal{Y}$ in H^n can be provided through the following extension of the notion of Mahalanobis distance [Mahalanobis, 1936]:

$$d_{\Sigma^{-1}}(\mathcal{X}, \mathcal{Y}) = \|\mathcal{X} - \mathcal{Y}\|_{\Sigma^{-1} - H^n} = \|\Sigma^{-1/2}(\mathcal{X} - \mathcal{Y})\|_{H^n},$$

where $\Sigma = \Sigma^{1/2}\Sigma^{T/2}$ and $\|\cdot\|_{H^n}$ denotes the norm in H^n, defined as $\|\mathcal{X}\|^2_{H^n} = \sum_{i=1}^n \|\mathcal{X}_i\|^2_H$, for $\mathcal{X} = (\mathcal{X}_1, ..., \mathcal{X}_n) \in H^n$.

The GLS estimator $\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS} = (\widehat{\boldsymbol{a}}_0^{GLS}, ..., \widehat{\boldsymbol{a}}_L^{GLS})^T$ can be determined solving the following optimal problem:

$$\min_{\widehat{\boldsymbol{a}}_{\vec{l}} \in H^{L+1}} \Phi^{GLS}(\widehat{\boldsymbol{a}}_{\vec{l}}) \tag{23}$$

where the functional Φ^{GLS} to be minimized corresponds to the functional Mahalanobis distance between fitted values $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} = \mathbb{F}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{a}}_{\vec{l}}$ and observed data:

$$\Phi^{GLS}(\widehat{\boldsymbol{a}}_{\vec{l}}) = \|\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - \mathbb{F}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{a}}_{\vec{l}}\|^2_{\Sigma^{-1}-H^n} = \|\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - \widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}\|^2_{\Sigma^{-1}-H^n}. \tag{24}$$

Proposition 7. *If* $\mathrm{rank}(\mathbb{F}_{\vec{\boldsymbol{s}}}) = L + 1$ *and* $\mathrm{rank}(\Sigma) = n$*, there exists a unique vector* $\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS}$ *solving the estimation problem* (23)*, which admits the following explicit representation:*

$$\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS} = (\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}\Sigma^{-1}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}}. \tag{25}$$

Moreover, the (unique) GLS drift estimator $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}$ *is:*

$$\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} = \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}\Sigma^{-1}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}}. \tag{26}$$

Since estimators (25) and (26) are linear, their mean and variance-covariance matrix can be easily derived obtaining:

$$\mathbb{E}[\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS}] = a_{\vec{l}}; \quad \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS}) = (\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}; \tag{27}$$

$$\mathbb{E}[\widehat{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}] = m_{\vec{\boldsymbol{s}}}; \qquad \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}) = \mathbb{F}_{\vec{\boldsymbol{s}}}^T(\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}. \tag{28}$$

Besides being unbiased, the following result holds.

Proposition 8. *The estimator* $\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS}$ *is the BLUE (*Best Linear Unbiased Estimator*) for the coefficients* $a_{\vec{l}}$*, i.e. for any other linear unbiased estimator* $\widehat{\boldsymbol{a}}_{\vec{l}} = \mathbb{A}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} + \boldsymbol{b}$ *of* $a_{\vec{l}}$*, the matrix:*

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}^{BLUE})$$

is positive semi-definite. As a consequence, $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}$ *is the BLUE for the mean vector* $m_{\vec{\boldsymbol{s}}}$*.*

Proposition 8 not only provides an optimality result in terms of drift estimation, but also in terms of bias involved in the estimation of Σ.

Indeed, let Σ^{GLS} be the $n \times n$ covariance matrix of the estimator $\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}} = \boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - \widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}$, $\Sigma^{GLS} = \mathbb{E}[\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}^T]$, then the identity:

$$\Sigma = \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}) + \Sigma^{GLS}, \tag{29}$$

can be verified through orthogonality arguments (for details see the Appendix). Expression (29) provides a decomposition of the covariance matrix Σ in a part depending on the variability of the drift estimator $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}$ and a component representing the dependence structure of the estimated residual process.

Although the covariance matrix Σ^{GLS} of $\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}$ represents the natural estimator of Σ, it provides a biased estimation of the spatial-dependence structure, underestimating it for a quantity:

$$\mathbb{B} = \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}) = \mathbb{F}_{\vec{\boldsymbol{s}}}^{T}(\mathbb{F}_{\vec{\boldsymbol{s}}}^{T}\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}. \tag{30}$$

However, the bias $\mathbb{B}$ of the estimator Σ^{GLS} coincides with the minimum bias obtainable by a least squares estimation procedure, since $\widehat{\boldsymbol{a}}_{\vec{l}}^{GLS}$ and $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}^{GLS}$ are BLUE. Indeed, every other choice of the weighting matrix for the generalized least squares criterion would lead to a linear unbiased estimator $\widehat{\boldsymbol{m}}'_{\vec{\boldsymbol{s}}}$, whose resulting bias matrix $\mathbb{B}'$ would be higher than $\mathbb{B}$ (i.e. $\mathbb{B}'-\mathbb{B}$ would be semi-definite positive).

For these reasons, precision in the least squares estimation procedure of the drift assumes a double role, determining both the accuracy in the estimate of the deterministic component and the bias in the assessment of the spatial-dependence structure relative to the stochastic component, that plays a key role in kriging prediction.

The magnitude of the bias reduction due to the adoption of a GLS criterion instead of an OLS one is clearly dependent on the structure of spatial dependence. Indeed, such a reduction can be explicitly computed as:

$$\Delta_{\mathbb{B}} = \mathbb{B}^{OLS} - \mathbb{B} = \mathbb{F}_{\vec{\boldsymbol{s}}}(\mathbb{F}_{\vec{\boldsymbol{s}}}^{T}\mathbb{F}_{\boldsymbol{s}})^{-1}(\mathbb{F}_{\vec{\boldsymbol{s}}}^{T}\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})(\mathbb{F}_{\vec{\boldsymbol{s}}}^{T}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}^{T} - \mathbb{F}_{\vec{\boldsymbol{s}}}^{T}(\mathbb{F}_{\vec{\boldsymbol{s}}}^{T}\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}.$$

Observe that $\Delta_{\mathbb{B}}$ annihilates when $\Sigma = \sigma^2\mathbb{I}$, that is precisely the case of (globally) uncorrelated residuals for which OLS and GLS criteria coincide.

From the point of view of the residual variogram, the (global) uncorrelated case corresponds to a pure nugget structure, because the mean squared norm of the discrepancy among uncorrelated observations is equal to the variance σ^2 of the process, being thus independent from their separating distance. Therefore, the estimation of the residuals variogram, besides allowing the analysis of the spatial dependence structure, will be the leading tool in the determination of the most proper procedures for the statistical treatment of the observations, as will be clarified in the next sections.

3 Algorithms

Drift Estimation In order to compute the Universal Kriging prediction coherently with the established theoretical results, an iterative algorithm is necessary. Indeed, both the GLS drift estimator $\widehat{m}_{\vec{\boldsymbol{s}}}^{GLS}$ and the system (17) depend substantially on the residual covariance structure, that can be assessed only once an estimation of the residual process –obtainable by difference from the estimate $\widehat{m}_{\vec{\boldsymbol{s}}}^{GLS}$– is available.

Therefore we propose to initialize the procedure to the ordinary least squares (OLS) estimate, computing at each step the residual estimate and the related trace-variogram structure, as well as the update of the drift estimate on the basis of the structure of spatial dependence currently available.

Having reached convergence, proved to be within five iterations by simulations, the final estimate of the variogram model can be used to solve the Universal Kriging system (17), deriving the desired prediction. The described algorithm is summarized in Algorithm 9.

Algorithm 9. *Given a realization* $\chi_{\vec{\boldsymbol{s}}} = (\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n})$ *of the nonstationary random field* $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, $D \subset \mathbb{R}^d$, *representable as in* (11)*:*

1. *Estimate the drift vector* $m_{\vec{\boldsymbol{s}}}$ *through the OLS method* ($\widehat{m}^{OLS}_{\vec{\boldsymbol{s}}} = \mathbb{F}_{\vec{\boldsymbol{s}}}(\mathbb{F}^T_{\vec{\boldsymbol{s}}} \mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}^T_{\vec{\boldsymbol{s}}}\chi_{\vec{\boldsymbol{s}}}$) *and set* $\widehat{m}_{\vec{\boldsymbol{s}}} := \widehat{m}^{OLS}_{\vec{\boldsymbol{s}}}$.

2. *Compute the residual estimate* $\widehat{\delta}_{\vec{\boldsymbol{s}}} = (\widehat{\delta}_{s_1}, ..., \widehat{\delta}_{s_n})$ *by difference* $\widehat{\delta}_{\vec{\boldsymbol{s}}} = \chi_{\vec{\boldsymbol{s}}} - \widehat{m}_{\vec{\boldsymbol{s}}}$.

3. *Estimate the trace-variogram* $2\gamma(\cdot)$ *of the residual process* $\{\boldsymbol{\delta}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ *from* $\widehat{\delta}_{\vec{\boldsymbol{s}}}$ *first with the empirical estimator* (20)*, then fitting a valid variogram model* $\gamma(\cdot; \widehat{\boldsymbol{\vartheta}})$*. Derive from* $\gamma(\cdot; \widehat{\boldsymbol{\vartheta}})$ *the estimate* $\widehat{\Sigma}$ *of* Σ.

4. *Estimate the drift vector* $m_{\vec{\boldsymbol{s}}}$ *with* $\widehat{m}^{GLS}_{\vec{\boldsymbol{s}}}$*, obtained from* $\chi_{\vec{\boldsymbol{s}}}$ *using* (26).

5. *Repeat 2.-4. until convergence has been reached.*

For computational efficiency reasons, the step 4. can be performed through the auxiliary uncorrelated vector $\widehat{\widetilde{\chi}}_{\vec{\boldsymbol{s}}} = L^{-1}\chi_{\vec{\boldsymbol{s}}}$, where L appears in the Cholesky decomposition $\widehat{\Sigma} = LL^T$. Indeed, $\widehat{\widetilde{\chi}}_{\vec{\boldsymbol{s}}}$ is an estimate of $\widetilde{\chi}_{\vec{\boldsymbol{s}}} = \Sigma^{-1/2}\chi_{\vec{\boldsymbol{s}}}$ since the inverse Cholesky factor L^{-1} provides an estimate of $\Sigma^{-1/2}$ (see the proof of Proposition 7 for details).

Drift Model Selection Although knowledge of the functions f_l, for $l = 1, ..., L$ ($f_0(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$), is one of the underlying assumptions for the procedure detailed in Algorithm 9, in most applications no 'a priori' information is available about the family $\{f_l\}_{l=0,...,L} = \{f_0, ..., f_L\}$ (e.g., no scalar external drift for the observed phenomenon is known). Therefore a model selection step before the application of Algorithm 9 is needed. In order to handle the model selection problem, we propose first to choose a number of candidate regressors families –e.g. the 2^5 polynomials of order lower than 2– then to select the optimal set of regressors with predictive criterion.

Formally, consider N_f collections of functions $f^k_l = \{f^k_0, ..., f^k_L\}$, corresponding to N_f possible drifts $m^k_{\boldsymbol{s}} = \sum_{l=0}^{L} a_l f^k_l(\boldsymbol{s})$, $\boldsymbol{s} \in D$, $k = 1, ..., N_f$. The aim of the proposed method is the determination of a permutation $\{(1), ..., (N_f)\}$ of the set of indexes $\{1, ..., N_f\}$ according to the mean squared error of prediction:

$$MSE_k = \mathbb{E}[\|\boldsymbol{\chi}_{\boldsymbol{s}} - \boldsymbol{\chi}^{*k}_{\boldsymbol{s}}\|^2], \quad k = 1, ..., N_f.$$

that can be assessed by a cross-validation (leave-on-out) technique combined with a Universal Kriging prediction, based on a proper drift estimate. For computational efficiency, an OLS drift estimation can be used in Universal Kriging prediction, which actually corresponds to the very first iteration of the Algorithm 9. The proposed procedure is summed up in the following Algorithm.

Algorithm 10. *Given a realization $\chi_{\boldsymbol{s}_1}, ..., \chi_{\boldsymbol{s}_n}$ of the nonstationary random field $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ and N_f collections of functions $f^k_{\vec{l}} = \{f^k_0, ..., f^k_L\}$ (candidate forms for the drift):*

1. *Fix a collection $f^k_{\vec{l}}$, $k = 1, ..., N_f$;*
2. *Compute the GLS drift estimate $\widehat{m}^{GLS,k}_{\vec{\boldsymbol{s}}}$, the residual estimate $\widehat{\delta}^k_{\vec{\boldsymbol{s}}}$ and the corresponding trace-variogram model $\gamma_k(\cdot)$ applying M iterations of Algorithm 9 ($M = 1$ for OLS estimate);*
3. *Per each fixed $i = 1, ..., n$, predict $\chi_{\boldsymbol{s}_i}$ from $\chi_{\vec{\boldsymbol{s}}^{-i}} = (\chi_{\boldsymbol{s}_j})_{j \neq i}$ through the Universal Kriging predictor $\chi^{*k}_{\boldsymbol{s}_i}$ solving (17) with $\gamma = \gamma_k$ and $f_{\vec{l}} = f^k_{\vec{l}}$;*
4. *Compute the sample mean squared error: $MSE_k = \frac{1}{n} \sum_{i=1}^n \|\chi_{\boldsymbol{s}_i} - \chi^{*k}_{\boldsymbol{s}_i}\|^2$;*
5. *Repeat 1.-4. for every collection $f^k_{\vec{l}}$, $k = 1, ..., N_f$;*
6. *Sort $\{MSE_1, ..., MSE_{N_f}\}$ in increasing order, determining the optimal permutation $\{(1), ..., (N_f)\}$ of $\{1, ..., N_f\}$; order the collections $\{f^k_{\vec{l}}\}_{k=1,...,N_f}$ according to $\{(1), ..., (N_f)\}$, $\{f^{(k)}_{\vec{l}}\}_{(k)=1,...,N_f}$;*
7. *For $k = 1, ..., N_f$:*
   a. *Check the second-order stationarity of the residual variogram model $\gamma_{(k)}(\cdot)$ relative to the (k)-th model;*
   b. *If $\gamma_{(k)}(\cdot)$ proves to be second-order stationary, select the optimal drift model as:*

$$m^{opt}_{\boldsymbol{s}} = \sum_{l=0}^{L} a_l f^{(k)}_l(\boldsymbol{s}), \quad \boldsymbol{s} \in D,$$

   and stop the procedure.

Note that step 7. of Algorithm 10 guarantees the stationarity of the residuals and thus ensures that the Universal Kriging hypotheses are fulfilled by the selected drift model. The residual second-order stationarity can be checked through the analysis of the residual empirical variogram with the same criteria used in finite-dimensional geostatistics (e.g. presence of a sill close to the estimated variance and sub-quadratic growth for increasing distances).

Moreover the adoption of a predictive criterion contributes to avoid overfitting: very complex models are unable to filter the noise in the observed data,

therefore the selection of a too complex drift structure would also catch part of their stochastic variability, reducing considerably the predictive power of the model.

In order to obtained the final Universal Kriging prediction, Algorithms 9 and 10 need to be combined. Two main choices can be made, according to computational efficiency or estimation accuracy criteria.

The first possibility is to consider for step 2. of Algorithm 10 the OLS estimation method. By making this choice a three-step procedure is finally obtained: first drift model selection by Algorithm 10, second GLS estimation by Algorithm 9, finally Universal Kriging prediction. This choice aims mainly in controlling the computational costs, ignoring the possible influences of the drift estimation method on the prediction (not always negligible).

The second possible choice is the integration of Algorithms 9 and 10, by considering GLS estimation method during step 2. of Algorithm 10 and then using the drift estimate of the selected model, available at the end of Algorithm 10, for Universal Kriging prediction. This choice does not preserve the computational costs from becoming high in the presence of many candidates families, but permits to perform a more precise drift model selection, which contributes to make the kriging prediction more accurate. Moreover, the fairly high speed of convergence of Algorithms 9 –by 5 iterations in all the simulations– and the consideration of a moderate number of drift candidates contribute to control the computational efficiency of the procedure. For these reasons, the choice made in this work is the latter.

4 Examples and Simulation Study

4.1 Simulation of Non-stationary Functional Processes

The simulation of functional stochastic processes $\{\boldsymbol{\chi}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ of the form (11) can be performed first by simulating a second-order stationary and isotropic residual field $\{\boldsymbol{\delta}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ by direct construction as in (9), then by generating the drift term and finally by summing residuals and drift term.

For this Section, the residual fields have been simulated considering the space $H = L^2([0,1])$, endowed with the Fourier orthonormal basis $\{e_j, j \geq 1\}$. Expansion (9) has been truncated to the 7^{th} order for the first dataset of Subsection 4.2 and for the 5 collections of datasets analyzed in Subsection 4.3; for the second dataset of Subsection 4.2 expansion (9) has been truncated to the 25^{th} order, assuming all coefficients to be zero except for the last seven. The generation of the 7 non-null scalar fields $\{\xi_j(\boldsymbol{s}), \boldsymbol{s} \in D\}$, $1 \leq j \leq 7$, involved in expansion (9) –which in fact determine the structure of spatial dependence of the functional random field– has been performed by means of the geostatistical software ISATIS®. In particular, each scalar field has been independently simulated on a fine grid over the domain $D = [0,2] \times [0,3] \subset \mathbb{R}^2$, according to a gaussian second-order sta-

	Structure	(Sill, Range, Nug.)
ξ_1	Exp.	(16, 0.75, 0)
ξ_2	Sph.	(16, 0.75, 0)
ξ_3	Exp.	(16, 1.50, 0)
ξ_4	Sph.	(16, 1.50, 0)
ξ_5	Sph.; Exp.	(8, 0.75, 0); (8, 0.75, 0)
ξ_6	Sph.; Exp.	(8, 0.75, 0); (8, 0.75, 0)
ξ_7	Sph.; Exp.	(12, 1.50, 0); (4, 0.75, 0)

Table 1: Generating variogram models for the scalar fields $\{\xi_j(\boldsymbol{s})\}$, $j = 1, ..., 7$. Fields ξ_5, ξ_6, ξ_7 are generated by the sum of the indicated variogram structures.

tionary and isotropic distribution; the generating variograms are listed in Table 1. Having obtained the functional residuals over the whole grid by combining the scalar grid realizations, the residual datasets have been finally obtained by sampling uniformly $n = 100$ grid locations.

For Subsection 4.2, only stationary datasets –obtained directly from the residuals realizations – have been considered. For Subsection 4.3, non-stationary datasets have been built instead. For generating the drift terms, polynomials of degree lower than two have been considered:

$$m_{\boldsymbol{s}}(t) = a_0(t)+a_1(t)x+a_2(t)y+a_3(t)x^2+a_4(t)y^2+a_5(t)xy, \quad t \in [0,1], \boldsymbol{s} = (x,y) \in D, \tag{31}$$

where a_l are deterministic functional (possibly null) coefficients belonging to L^2. For the construction of a_l, $l = 0, ..., 5$, the same basis with the same truncation as for the residuals has been fixed:

$$a_l(t) = \sum_{j=1}^{7} \beta_{j,l} e_j(t), \quad t \in \mathcal{T}, \tag{32}$$

where $\beta_{j,l} \in \mathbb{R}$, $1 \leq j \leq 7$, $0 \leq l \leq 5$, are the deterministic coefficients of the expansion on the Fourier basis. In Table 2 the coefficients $\beta_{j,l}$ relative to the complete model are listed; drift models used in the considered synthetic examples are obtained as sub-models of the complete model as will be specified later on.

4.2 Trace-variograms in Sobolev Spaces: an example

The purpose of this first example is to show how the choice of the space H to which data are assumed to belong might heavily influence the way in which the spatial dependence is modeled (see also Remark 5 in Section 2).

To see this, we now consider two functional random fields, $\{\boldsymbol{\chi}_{\boldsymbol{s}}^{(m)}, \boldsymbol{s} \in D\}$, $m = 1, 2$, built by combining the scalar random fields $\{\xi_j(\boldsymbol{s}), \boldsymbol{s} \in D\}$, $j = 1, ..., 7$,

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
$\beta_{1,l}$	1.247	-5.050	4.011	0.389	1.734	1.572
$\beta_{2,l}$	0.979	1.651	-1.531	1.535	0.086	0.710
$\beta_{3,l}$	0.558	4.008	3.096	0.289	1.246	0.502
$\beta_{4,l}$	-0.047	-0.020	0.045	-0.031	0.008	-0.001
$\beta_{5,l}$	0.032	0.022	-0.024	-0.005	-0.008	-0.047
$\beta_{6,l}$	0.029	0.028	0.033	0.046	0.047	-0.0002
$\beta_{7,l}$	0.063	0.042	0.016	0.109	0.057	0.004

Table 2: Coefficients $\beta_{j,l}$, $1 \leq j \leq 7$, $0 \leq l \leq 5$ of the drift expansion (32) used for the construction of the complete model and the relative sub-models.

–introduced in Subsection 4.1– as:

$$\boldsymbol{\chi_s}^{(1)} = \sum_{k=1}^{7} \xi_k^{(1)}(\boldsymbol{s}) e_k = \sum_{k=1}^{7} \xi_k(\boldsymbol{s}) e_k \tag{33}$$

$$\boldsymbol{\chi_s}^{(2)} = \sum_{k=1}^{25} \xi_k^{(2)}(\boldsymbol{s}) e_k = \sum_{k=19}^{25} \xi_{k-18}(\boldsymbol{s}) e_k. \tag{34}$$

The corresponding functional datasets $\chi_{\boldsymbol{s}_1}^{(m)}, ..., \chi_{\boldsymbol{s}_{100}}^{(m)}$, $m = 1, 2$, have been obtained by combining according to (33) and (34) the set of realizations of the scalar fields ξ_j, $j = 1, ..., 7$, simulated as specified in Subsection 4.1. The functional datasets are represented in the left panels of Figure 1a and 1b. The different behavior of the curves is evident: the first dataset has a less fluctuating pattern along the coordinate $t \in [0, 1]$, since only the first 3 frequencies are excited; conversely, the second dataset is characterized by a very fluctuating pattern, due to the higher order basis truncation involving only the 10^{th} to 12^{th} frequencies. First order derivatives are represented in the right panels of Figure 1a and 1b.

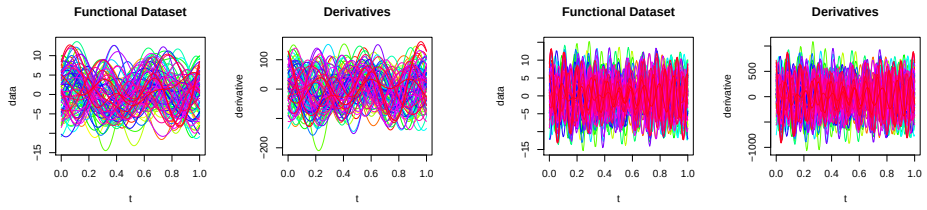
Notice that, by construction, each realization of both processes belongs not only to L^2, but also to $\mathcal{H}^1$; moreover, both processes are globally second-order stationary either in L^2 or in $\mathcal{H}^1$. Indeed, L^2 trace-variograms, as well as $\mathcal{H}^1$ trace-variograms, can be explicitly computed.

Assume first $H = L^2$. For $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, $m = 1, 2$, ($N_1 = 7$, $N_2 = 25$):

$$2\gamma^{(m)}(\boldsymbol{s}_i, \boldsymbol{s}_j)_{L^2} = \mathbb{E}[\|\boldsymbol{\chi}_{\boldsymbol{s}_i}^{(m)} - \boldsymbol{\chi}_{\boldsymbol{s}_j}^{(m)}\|_{L^2}^2] = \sum_{k=1}^{N_m} \mathbb{E}\left[|\xi_k^{(m)}(\boldsymbol{s}_i) - \xi_k^{(m)}(\boldsymbol{s}_j)|^2\right].$$

Therefore, for both functional random fields, the L^2 trace-variograms coincide:

$$2\gamma_{L^2}^{(1)} = 2\gamma_{L^2}^{(2)} = \sum_{k=1}^{7} 2\gamma_{\xi_k},$$



(a) 7 basis functions

(b) 25 basis functions

Figure 1: Functional datasets and corresponding derivatives. On the left: first dataset, built on a 7 Fourier functions basis. On the right: second dataset, built on a 25 Fourier functions basis, assuming non-zero only the last 7 coefficients.

where $2\gamma_{\xi_k}$ indicate the variogram of the field ξ_k, $k = 1, ..., 7$. Obviously, also their empirical estimates coincide (Figure 2a and 2b, left panels). Notice that the different behavior of the two datasets along the coordinate t is lost when inspecting L^2 trace-variograms: they are able to capture only the structure of spatial dependence determined by the fields ξ_j, $j = 1, ..., 7$, ignoring the possibly different associated frequencies.

Information regarding curve fluctuation can be modeled through first derivative: we thus assume $H = \mathcal{H}^1$. Trace-variograms in $\mathcal{H}^1$ can be computed by:

$$2\gamma^{(m)}(\boldsymbol{s}_i, \boldsymbol{s}_j)_{\mathcal{H}^1} = 2\gamma^{(m)}(\boldsymbol{s}_i, \boldsymbol{s}_j)_{L^2} + \mathrm{Var}(D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_i} - D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_j})_{L^2}.$$

However, since:

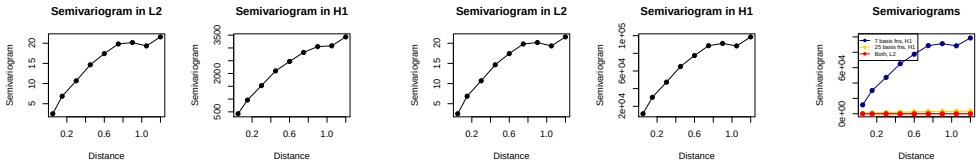
$$\mathrm{Var}(D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_i} - D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_j})_{L^2} = \mathbb{E}[\|D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_i} - D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_j}\|^2_{L^2}] - \|\mathbb{E}[D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_i} - D\boldsymbol{\chi}^{(m)}_{\boldsymbol{s}_j}]\|^2_{L^2} =$$
$$= \sum_{k=1}^{N_m} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 \mathbb{E}\left[|\xi_k(\boldsymbol{s}_i) - \xi_k(\boldsymbol{s}_j)|^2\right],$$

$2\gamma^{(1)}_{\mathcal{H}^1}$ does not coincide with $2\gamma^{(2)}_{\mathcal{H}^1}$. Indeed:

$$2\gamma^{(1)}_{\mathcal{H}^1} = 2\gamma^{(1)}_{L^2} + \sum_{k=2}^{7} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 2\gamma_{\xi_k} = \sum_{k=1}^{7} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_k};$$
$$2\gamma^{(2)}_{\mathcal{H}^1} = 2\gamma^{(2)}_{L^2} + \sum_{k=19}^{25} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 \gamma_{\xi_{k-18}} = \sum_{k=19}^{25} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_{k-18}}.$$

Notice that, for $k = 1, ..., 7$, the weights associated to the variogram γ_{ξ_k} depends on the frequency associated to ξ_k, a greater weight being assigned to a higher frequency.

The second panels of Figure 2a and 2b show the empirical $\mathcal{H}^1$ trace-variograms estimated from the two datasets. Although the shapes of the estimates are similar, without showing notable differences with respect to L^2 estimates, the orders



(a) 7 basis functions (b) 25 basis functions (c) All superimposed

Figure 2: Empirical trace-variograms in L^2 and $\mathcal{H}^1$.

of magnitude of the sills are significantly different. Observing Figure 2c, the different variances characterizing the two fields appear clearly: the curve corresponding to $\gamma^{(2)}_{\mathcal{H}^1}$ (in blue) is much higher than the other two, since the energy of the random field $\{\boldsymbol{\chi}^{(2)}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ is much higher than that of the others. In conclusion, the example clearly evidences as the choice of the space for the analysis has to be carefully taken, according to the dataset structure and, above all, to the purposes of the analysis. Indeed, if the aim of the analysis is purely spatial, then L^2 space is rich enough for exploratory analysis and for kriging prediction. In other situations, the choice of a Sobolev space might be needed instead. This is the case of the dynamical system example in Remark 5 of Section 2 or of the example presented here.

4.3 Simulation study

Model Selection Procedure The first goal of our simulation study is to evaluate the performance of Algorithm 10 in terms of error of model selection (referred to as *model misclassification*), which occurs when the selected model does not coincide with that generating the data, and type of error (in particular over-fitting or under-fitting[1]).

In order to analyze the behavior of the algorithm in different scenarios, pointing out possible tendency to over-fit or under-fit the data, 5 collections of 32 datasets each have been considered.

Data generation follows this scheme. First a set of 32 drift models has been built by considering the complete drift model (31) –constructed as previously specified and evaluated in the sampled locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_{100}$– and its 31 sub-models. In particular, for $t \in [0, 1]$, $\boldsymbol{s} = (x, y)$, drift model $k = 1, ..., 32$, has been represented through five binary variables $\{\zeta^{(k)}_1 \zeta^{(k)}_2, ..., \zeta^{(k)}_5\}$, by representing:

$$m^{(k)}_{\boldsymbol{s}}(t) = a_0(t) + \zeta^{(k)}_1 a_1(t)x + \zeta^{(k)}_2 a_2(t)y + \zeta^{(k)}_3 a_3(t)x^2 + \zeta^{(k)}_4 a_4(t)y^2 + \zeta^{(k)}_5 a_5(t)xy, \tag{35}$$

[1]We say that the algorithm finds an over-fitting solution when it selects a drift model including all the generating regressors plus at least one; analogously, we say that the algorithm selects an under-fitting solution when it selects a sub-model of the true model.

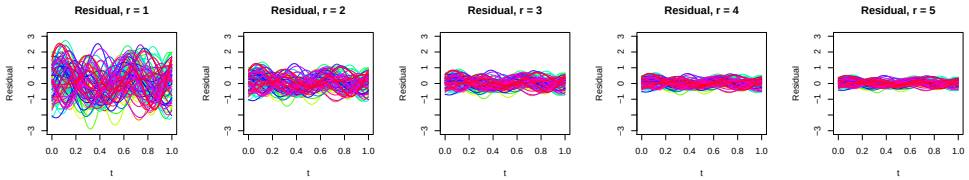


Figure 3: Residuals curves for $r = 1, 2, 3, 4, 5$ (from left to right). Curves in the panel r, $r = 2, 3, 4, 5$, are obtained dividing the first panel residuals by r, which is the same as considering the same realization of the first residual field but with a lower sill for the generating variogram, namely a sill divided by r.

and setting $\zeta_l^{(k)} = 1$ if the l-th regressor is included in the sub-model k, $\zeta_l^{(k)} = 0$ otherwise, for $l = 1, ..., 5$. Hence, the sub-models have been ordered according to the bijective relation among the binary numbers $\zeta_5^{(k)}\zeta_4^{(k)}\zeta_3^{(k)}\zeta_2^{(k)}\zeta_1^{(k)}$, $k = 1, ..., 32$, and their decimal representations plus 1 (the constant term is always included): the complete model is thus model $32 = 1 + 31$ ($31 \leftrightarrow 11111$), the spatially constant model is model 1 ($0 \leftrightarrow 00000$), while, for example, sub-model $18 = 1+17$ is the model with regressors $\{1, x, xy\}$ ($17 \leftrightarrow 10001$).

Given the set of drift terms, the first collection of 32 functional datasets, $\{\chi_{\vec{s}}^{(k,1)}, k = 1, ..., 32\}$, has been obtained by summing to each drift sub-model $m_{\vec{s}}^{(k)}$ the residuals generated as specified in Subsection 4.1 (Figure 3, left panel); the remaining 4 collections $\{\chi_{\vec{s}}^{(k,r)}, k = 1, ..., 32\}$, $r = 2, 3, 4, 5$, have been obtained with the same construction but dividing the residual realization by r (Figure 3, four panels on the right):

$$\chi_{\vec{s}}^{(k,r)} = m_{\vec{s}}^{(k)} + \delta_{\vec{s}}/r,$$

which in fact corresponds to a reduction by a factor $r = 2, ..., 5$ of the variogram sills reported in Table 1.

To illustrate how the datasets collections depend on the amplitude of the stochastic component, consider the drift model 18, which is of the form

$$m_{\boldsymbol{s}}^{(18)}(t) = a_0(t) + a_1(t)x + a_5(t)xy, \tag{36}$$

and consider the corresponding non-stationary data $\chi_{\vec{s}}^{(18,r)} = m_{\vec{s}}^{(k)} + \delta_{\vec{s}}/r$, $r = 1, ..., 5$ (Figure 4, upper panels). The increasing importance of the drift term is made already explicit by graphical inspection, but it is even more stressed by the empirical estimate of the variogram computed from the data (Figure 4, lower panels). Indeed, for higher levels of residuals amplitude ($r = 1$), the variogram is only slightly affected by the drift, presenting an almost stationary behavior (e.g., downwards concavity near the origin, presence of a horizontal asymptote for higher distances); on the contrary, for decreasing amplitude of the residuals

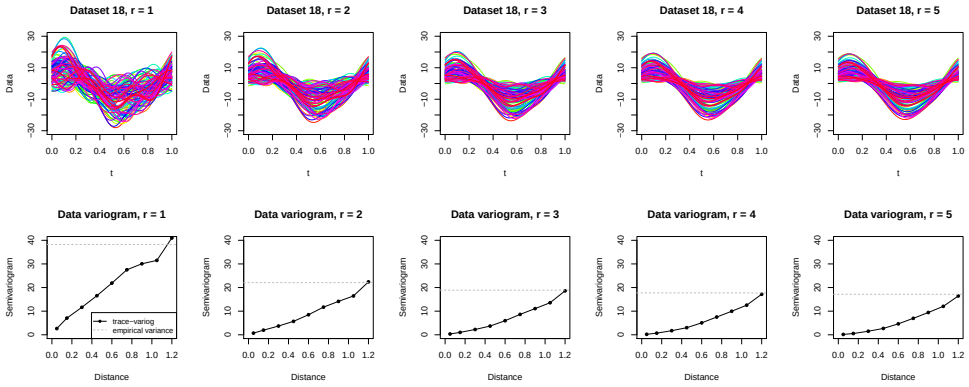


Figure 4: From left to right: Dataset $\chi_{\vec{s}}^{(18,r)}$ (upper panel) and the empirical variogram computed from the data (lower panel), for $r = 1, 2, 3, 4, 5$.

$(r = 2, 3, 4, 5)$, the experimental variograms assume a non-stationary aspect (e.g., upwards concavity near the origin, super-quadratic growth for higher distances). This behavior is mainly due to the increasing influence of the drift term on the data, since the empirical variogram estimator (19) computed from non-stationary data becomes severely biased when the drift component is predominant with respect to the residual one.

The generated collection of datasets has been used for testing the procedure as follows. For each collection r, $r = 1, ..., 5$, the model selection step has been separately applied to each of the corresponding 32 datasets, considering as candidate models all the 32 polynomials of degree lower than two, namely the complete model (31) and all its sub-models, and setting the number of the GLS iterations equal to $M = 5$, which seems sufficient for Algorithm 9 to converge. For each dataset $\chi_{\vec{s}}^{(k,r)}$, $k = 1, ..., 32$, the selected model and, in case of model misclassification, the type of error (over-fitting, under-fitting or none of them) has been recorded. Simulation results are shown in Figure 5. It is clear that the number of misclassified models sensibly decreases when the residual amplitude decreased. Indeed, as the r parameter increases, the drift term becomes more significant in the prediction: we thus expect a better performance of Algorithm 10 in cases high signal-to-noise ratio $(r = 3, 4, 5)$.

Consider now the behavior of the procedure in terms of over-fitting or under-fitting. Figure 5 shows that in response to a decrease in the residual amplitude $(r = 3, 4, 5)$ the behavior moves mainly from under-fitting to correct selection, except for a few cases in which over-fitting occurs (for $r = 5$, only datasets 5, 7, 13, 17, 25 are slightly over-fitted).

What is even more interesting to notice is that in very critical scenarios $(r = 1, 2)$, the most common misclassification error is under-fitting (Figure 6,

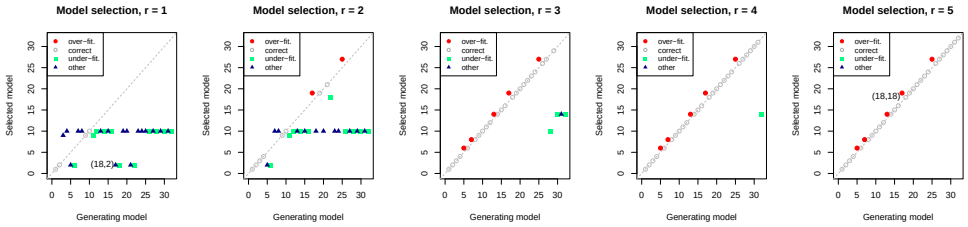


Figure 5: Simulation results for the model selection algorithm. From left to right: results applying Algorithm 10 to the collection of 32 datasets $\{\chi_{\vec{s}}^{(k,r)}, k = 1, ..., 32\}$ for $r = 1, 2, 3, 4, 5$. The horizontal axis identifies the number of the generating model, the vertical axis the number of the selected model; grey empty dots indicate correct selection, red full dots over-fitting, green square dots under-fitting, blue triangular dots the other cases. The points (18,2) –under-fitting– and (18,18) –correct selection– correspond to the dataset $\chi_{\vec{s}}^{(18,r)}$, for $r = 1, 5$.

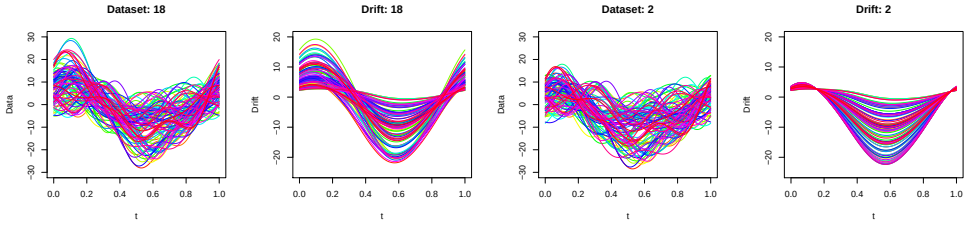
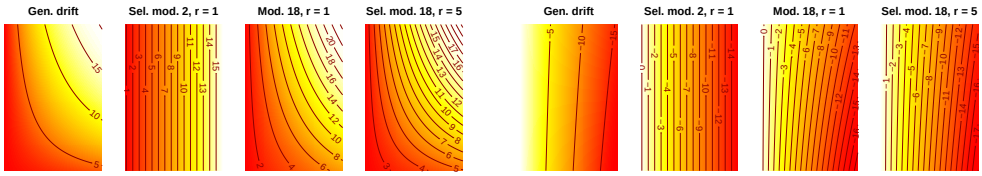


Figure 6: From left to right: dataset $\chi_{\vec{s}}^{(18,1)}$ and its drift (first and second panels), dataset $\chi_{\vec{s}}^{(2,1)}$ and its drift (third and fourth panels).

first and second panels). As an example, consider the dataset $\chi_{\vec{s}}^{(18,1)}$ (Figure 6, first panel). Comparing the non-stationary data with the drift curves (Figure 6, second panel), it is evident that the residuals heavily affect the shape of the curves, making them almost indistinguishable from the curves of dataset $\chi_{\vec{s}}^{(2,1)}$ (Figure 6, third panel). Here under-fitting occurs, but, as a matter of fact, drift models 2 and 18 are equally likely for the dataset $\chi_{\vec{s}}^{(18,1)}$: indeed, the larger fluctuations for t in $[0, 0.2]$ presented by dataset 18 –due to the xy component in drift model 18, which is missing in drift model 2– might be due to the residual fluctuation, and thus the simplest model is selected. Hence, in the presence of highly correlated residuals, Algorithm 10 proves to be very parsimonious, which is a very desirable property for a model selection procedure.

Drift Estimation and Universal Kriging Prediction Our next goals are the analysis of the performance of Algorithm 9 and the evaluation of Universal



(a) $t = 0.1$ (b) $t = 0.7$

Figure 7: Comparison of true and estimated drift through contour plots, for $t = 0.1, t = 0.7$. In each sub-figure, from left to right: generated drift grid, drift estimated with Algorithm 9 from $\chi_{\vec{s}}^{(18,1)}$ with model 2 –selected with Algorithm 10–, from $\chi_{\vec{s}}^{(18,1)}$ with model 18 and from $\chi_{\vec{s}}^{(18,5)}$ with model 18 –selected with Algorithm 10–.

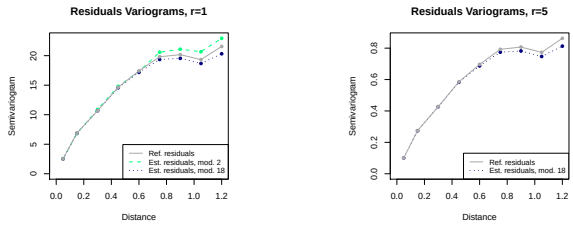
Kriging predictions.

Simulations have been performed on the data generated before, focusing on drift model 18 in the presence of residuals $\delta_{\vec{s}}/r$, with $r = 1, 5$ (Figure 3, first and fifth upper panels). Dataset $\chi_{\vec{s}}^{(18,1)}$ has been used in order to evaluate the effect of under-fitting both on drift estimates and on Universal Kriging predictions; in particular, Algorithm 9 and Universal Kriging prediction have been carried out first having selected the drift model by means of Algorithm 10 and then assuming that the true drift model is known. Dataset $\chi_{\vec{s}}^{(18,5)}$ has been used in order to study the influence of the residuals amplitude on the results, analyzing the performance of the procedures on less noisy data.

In all simulations the number of GLS iterations has been fixed to $M = 5$, which proved to be sufficient for Algorithm 9 to converge. For each of the three cases sketched before, the drift estimation, as well as the Universal Kriging prediction, has been performed over the whole generated grid for every t in $[0, 1]$. Since it is hardly possible to show at once a space-time grid of values –which is 4-dimensional– we consider two kind of visualizations: the functional visualization, obtained by plotting $t \in [0, 1]$ on the horizontal axis, and the value $\chi_{\boldsymbol{s}}(t)$ on the vertical axis for different $\boldsymbol{s} \in D$ –thus ignoring the spatial location– and the space contour representation, obtained by slicing the 4D space-time grid at some fixed t –thus loosing the functional variation–.

Figure 7 shows the contour plots of the GLS drift estimation, for all the three considered situations, namely for dataset $\chi_{\vec{s}}^{(18,1)}$ with drift model 2 –selected by Algorithm 10– (second panels of Figure 7a and 7b), with drift model 18 (third panels of Figure 7a and 7b) and for dataset $\chi_{\vec{s}}^{(18,5)}$ with drift model 18 (fourth panels of Figure 7a and 7b). Recall that both datasets are characterized by the same generating drift model.

It is clear from Figure 7a that the error in the drift estimate is not negligible in most critical situations, namely for $r = 1$, $t \in [0, 0.2]$ and when model 2 is



(a) $r = 1$ (b) $r = 5$

Figure 8: Residuals variograms: empirical variogram computed from generated residuals (solid grey lines), empirical variogram of the estimated residuals with model 18 (dotted blue lines), variogram of the residuals estimated with drift model 2 (only in left panel, dashed green line).

chosen by Algorithm 10 (Figure 7a second panel; recall Figure 6), and it becomes even more severe where the non-linear behavior of the drift is more apparent, as in the bottom-right part of the spatial domain. However, for higher values of t, the linear model 2 seems appropriate (Figure 7b second panel), although it is more parsimonious than the generating one.

The choice of the drift model has consequences not only on the drift maps, but also on the residuals variogram estimate. As a matter of fact, the deterministic variability not captured by under-fitted drifts, is picked up by the corresponding residual variogram, leading to its over-estimation (Figure 8a, green line). Such over-estimation is partially balanced by the downward bias that occurs estimating the variogram from estimated residuals –due to the variance decomposition (29)– which seems not to be very severe in the considered cases (compare gray lines and blue lines in Figure 8a and 8b).

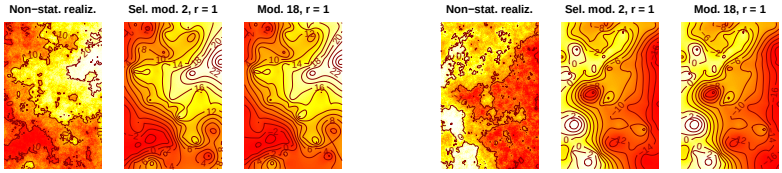
Even though the drift estimate as well as the variogram estimation obtained by selecting model 2 instead of model 18 might not seem very satisfactory, the Universal Kriging prediction appears not to be affected by under-fitting, both for $t = 0.1$ and for $t = 0.7$ (Figure 9a and 9b). Indeed, all the patterns presented by the original grid realization (first panels in Figure 9a and 9b) are well reproduced by both interpolations, with very similar results for models 2 and 18.

Cross-validation results, shown in Table 3, confirm these graphical observations. The statistics relative to the $n = 100$ cross-validation squared errors $\|\chi_{\boldsymbol{s}_i} - \chi^{*k}_{\boldsymbol{s}_i}\|^2$, $i = 1, ..., 100$ and $k = 2, 18$, are very similar, with slightly better results for the selected model –which is not surprisingly, since the optimality criterion in Algorithm 10 is precisely based on cross-validation error–.

Moreover, notice that dataset $\chi^{(18,1)}_{\vec{\boldsymbol{s}}}$ is characterized by a high residuals amplitude and thus the prediction turns out to be only slightly drift-driven: therefore, the Universal Kriging prediction proves to be very robust with respect to under-fitting of the drift model. On the contrary, in the presence of low resid-

	Selected Model: 2	Correct Model: 18
Median	5.19	5.33
Mean	7.47	7.54
Sum	747.3	753.7

Table 3: Cross-validation squared error for $\chi_{\vec{s}}^{(18,1)}$, considering model 2 –selected by Algorithm 10– and model 18.



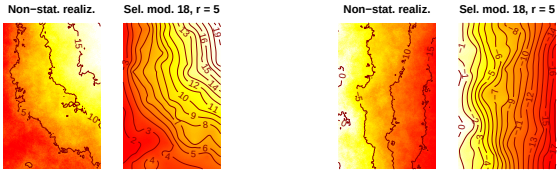
(a) $t = 0.1$ (b) $t = 0.7$

Figure 9: Comparison of simulated grid and UK prediction for $\chi_{\vec{s}}^{(18,1)}$ through contour plots, for $t = 0.1, 0.7$. In each sub-figure, from left to right: generated grid, UK prediction for with drift model 2 –selected with Algorithm 10– and with drift model 18.

uals amplitude, as for $r = 5$, the drift term becomes very influential on the data and on the prediction: in such a case, the performance of the model selection algorithm is much more satisfactory than before (Figure 5), as well as the drift estimation (fourth panels of Figure 7a and 7b), and the Universal Kriging prediction appears to be very accurate (Figure 10).

The increasing precision in estimating the drift for decreasing residuals variability is even more apparent by comparing the estimated drift coefficients $\widehat{a}_l$, $l = 0, 1, 5$, computed from dataset $\chi_{\vec{s}}^{(18,1)}$ and $\chi_{\vec{s}}^{(18,5)}$, adopting in both cases drift model 18, (Figure 11). Indeed, the coefficients relative to the case $r = 1$ (Figure 11, upper panels) capture also part of the stochastic variability and are thus much more fluctuating than the reference ones, while the coefficients computed from $\chi_{\vec{s}}^{(18,5)}$ (Figure 11, lower panels) are much more smooth reproducing more precisely the reference ones. Notice that the first situation is particularly critical because both residuals and drift curves are built on the same truncated basis and thus excite the same set of frequencies. The presence of more uncertainty in the estimates for noisier data is confirmed by the curves $\widehat{a}_l \pm 2\sqrt{\Lambda_{ll}}$, $\Lambda = \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}})$, $l = 0, 1, 5$, reported in Figure 11, which provide measures of the estimates variability.

Several other scenarios have been considered in the simulation study, obtaining further evidence of the results shown here: the performance of the proposed



(a) $t = 0.1$ (b) $t = 0.7$

Figure 10: Comparison of simulated grid and UK prediction for $\chi_{\vec{s}}^{(18,5)}$ through contour plots, for $t = 0.1, t = 0.7$. In each sub-figure: generated grid (left panel), UK prediction for with drift model 18 –selected with Algorithm 10– (right panel).

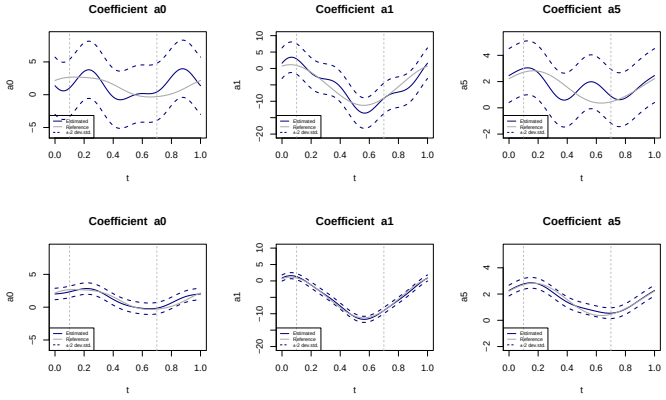


Figure 11: Comparison of the coefficients estimates $\widehat{a}_l$, $l = 0, 1, 5$, computed from dataset $\chi_{\vec{s}}^{(18,1)}$ (upper panels) and from dataset $\chi_{\vec{s}}^{(18,5)}$ (lower panels): estimated functional coefficients $\widehat{a}_l$, $l = 0, 1, 5$ (solid blue lines), generated coefficients a_l, $l = 0, 1, 5$ (solid grey lines). Dashed blue lines correspond to $\widehat{a}_l \pm 2\sqrt{\Lambda_{ll}}$, $\Lambda = \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}})$, $l = 0, 1, 5$. Vertical dashed grey lines indicate $t = 0.1$ and $t = 0.7$.

methodology on simulated data confirmed to be very satisfactory.

Indeed, the combination of Algorithms 9 and 10 leads to a very robust and flexible procedure. On one hand, in the presence of highly correlated data, the adoption of a predictive criterion for selecting the drift model proves to be appropriate to avoid over-fitting in favor of more parsimonious models; at the same time, in such cases Universal Kriging prediction proves to be very robust to under-fitting. On the other hand, in the presence of less noisy data, the results obtained by Algorithm 10 become more reliable, as well as the drift estimations by Algorithm 9, leading to a very precise prediction.

In any case, the obtained predictions appear very accurate in reproducing all the main patterns presented by the generated realizations, with only a moderate smoothing effect. This is a remarkable result especially given the simplicity of the kriging predictor which involves only global definitions of spatial dependence, besides being linear in the data through scalar coefficients. Indeed, simulations show that the proposed methodology is so flexible that not only global features, but also local structures can be well reproduced by this kind of predictor.

5 A Case Study: Analysis of Canada's Maritime Provinces Temperatures

Analysis of Averaged Temperatures Data The proposed methodology will be now applied to the Canada's Maritime Provinces Temperatures dataset (available in R package `geofd` [Giraldo et al., 2010c]), that collects daily mean temperatures data, observed in 35 meteorological stations located in Canada's Maritimes Provinces (Figure 12). This region consists of three provinces, Nova Scotia, New Brunswick and Prince Edward Island, located in the south-eastern part of Canada (Figure 12, first panel), whose very distinctive feature is the exposition toward the sea: indeed, especially because of the Gulf Stream coming from the Ocean, the Provinces climate is temperate, characterized by mild winters and cool summers [Stanley, 2002].

For each sampled site (Figure 12, second panel), identified by geographical coordinates (longitude, latitude), the original raw data consist of 365 measurements (one per day), obtained by averaging, over the years 1960 to 1994, the daily mean temperatures recorded by the Meteorological Service of Canada. This dataset, besides being very similar to the Canadian Weather dataset handled in [Ramsay and Silverman, 2005], has been often analyzed in the literature concerning geostatistical theory for stationary and isotropic functional processes (e.g., [Giraldo, 2009], [Delicado et al., 2010], [Giraldo et al., 2010b]).

Coherently with previous analyses, the Hilbert space H has been set to be L^2 and row data (Figure 12, third panel) have been projected on a basis of 65 Fourier function, selected in [Giraldo, 2009] through a non parametric functional cross-validation procedure (Figure 12, last right panel).

Denote with $\{\boldsymbol{\chi}_s, s \in D \subset \mathbb{R}^d\}$ the random field of temperature functions

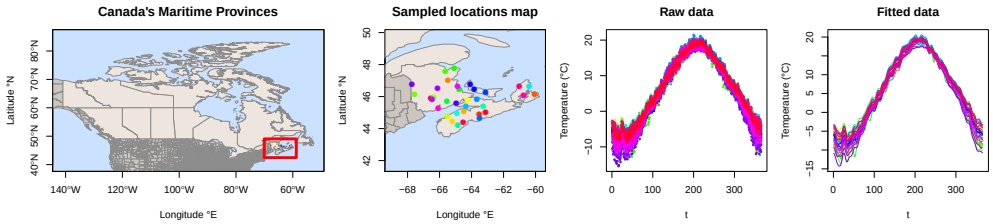


Figure 12: Canada's Maritime Provinces Temperatures dataset, averaged over 1960-1994. From left to right: map of Canada highlighting the Maritimes region; zoom of Maritime Provinces and sampled locations; raw data; fitted data.

and call D the spatial domain, endowed with the non-Euclidean metric induced by the geodesic distance that, assuming a spherical approximation for the Earth, can be explicitly computed as:

$$d_g(s_1, s_2) = 2R_m \arcsin\left(\sqrt{\sin^2\left(\frac{\zeta_1 - \zeta_2}{2}\right) + \cos(\zeta_1)\cos(\zeta_2)\sin^2\left(\frac{\varphi_1 - \varphi_2}{2}\right)}\right), \tag{37}$$

where $s_i = (\zeta_i, \varphi_i)$, (longitude and latitude) $i = 1, 2$ e $R_m \simeq 6371$ km indicates the Earth's mean radius. As noticed in [Banerjee, 2005], when working with geographical coordinates the above metrics is preferable to the Euclidean one –adopted in previous works (e.g. [Delicado et al., 2010])–; moreover, although the validity of usual parametric variogram models is not guaranteed in non-Euclidean spaces [Curriero, 2006], both the spherical and the exponential models are valid in the spherical geometry [Huang et al., 2011] and thus can be used in this case. By a first stationary analysis of the data through the trace-variogram empirical estimate, represented in the left panel of Figure 13, the non-stationarity of the field is apparent (super-quadratic growth for increasing distances, no evidence of a sill close to the sample variance of the data). Therefore, we analyze the data by means of Algorithms 9 and 10, searching the optimal drift model among polynomials of degree lower than 2.

The linear model singled out by the Algorithm 10 is model 23:

$$m(s,t) = a_0(t) + a_1(t)y + a_2(t)x^2 + a_3(t)xy, \quad s = (x,y),\ t \in \mathcal{T} = [0, 365], \tag{38}$$

where the coordinates are identified with latitude and longitude, $(x, y) = (\zeta, \varphi)$. Concerning the residuals structure of spatial dependence, the right panel of Figure 13 shows that the parametric model that best fits the empirical trace-variogram estimate is a pure-nugget model, meaning that the estimated residuals are uncorrelated. Therefore, the spatial variability characterizing the data is mostly explained by the deterministic drift term, while the residuals do not seem to contribute to the spatial correlation of the stochastic process.

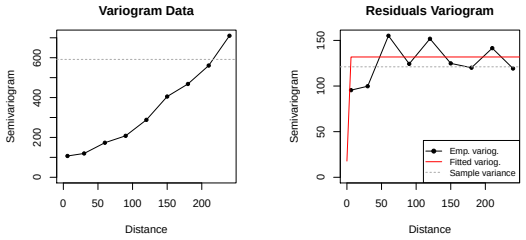


Figure 13: Estimated trace-variograms from data (on the left) and from residuals (on the right).

	UKFD (Pure nugget; Geod. dist.)	OKFD (Geod. dist.)	OKFD (Eucl. dist.)	FKTM (Eucl. dist.)
Median	99.1 (↓ 31%, 32%, 30%)	144.3	144.6	142.6
Mean (MSE)	155.4 (↓ 8%, 13%, 13%)	168.8	179.2	178.6

Table 4: Comparison of cross-validation squared error statistics computed by Universal Kriging (UKFD), Ordinary Kriging (OKFD, Delicado et al. [2010]) –using geodetic and Euclidean distance – and Functional Kriging Total Model (FKTM, Delicado et al. [2010]). The reduction of UKFD error with respect to OKFD and FKTM is reported between brackets.

In such a case, Universal Kriging predictor reduces to the drift estimate –i.e. the prediction which would have been obtained via FDA linear models–, which in fact provides the best predictive performance among the functional forms tested by the Algorithm 10. In particular, it is better performing –in terms of cross-validation errors– than the Ordinary Kriging predictor (i.e. drift model 1) computed by using geodesic distance (Table 4, second column), as well as by using Euclidean distance (Table 4, third column). Not even the Functional Kriging Total Model predictor (FKTM, Delicado et al. [2010]), which is the most complex stationary kriging predictor available in the literature, achieves results as satisfactory as those of the UKFD predictor (fourth column). Indeed, cross-validation statistics obtained with the proposed methodology are improved at least of 8% with respect to the stationary methods and at least of 13% with respect to the analyses already presented in the literature[2]. Notice in particular that the presented methodology is much simpler and computational efficient than FKTM.

The fact that the residuals do not show a non-trivial structure of spatial

[2]The codes for computing the stationary predictors are available in `geofd` R package. Cross-validation statistics are here computed with respect to fitted data and are thus different from statistics reported in previous works (e.g., [Delicado et al., 2010]), that refers to raw data instead.

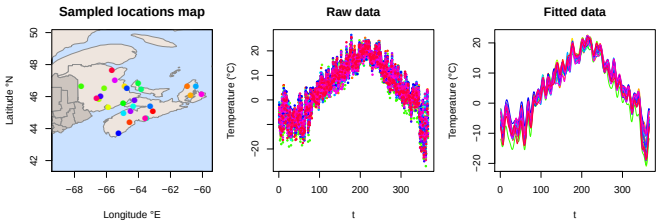


Figure 14: Canada's Maritime Provinces Temperatures dataset, year 1980. From left to right: map of Maritime Provinces and sampled locations; raw data; fitted temperature curves.

dependence might be due to the average over 34 years made on the original data, which may have masked the small scale variability. For this reason, we will now apply our methodology to a one-year dataset, collecting the measurements recorded in the same area, during the year 1980.

Analysis of 1980 Temperatures Data The dataset analyzed in this second part of the case study collects daily mean temperatures recorded, along the (leap) year 1980, in 27 meteorological stations located in the same region considered before (Figure 14, left panel). The raw data (Figure 14, central panel), available on the Natural Resources of Canada website [2012] website, have been projected as before on a basis of 65 Fourier functions, obtaining the functional dataset represented in the right panel of Figure 14. Choices for the functional and spatial metrics have been taken coherently with the previous analysis: the functional space is $H = L^2$, while distances between spatial locations have been computed by geodesic distance (37).

Although the empirical trace-variogram estimated from the data (Figure 15, left panel) seems not so far from stationarity, we proceed in applying our procedure, since among polynomials of degree lower than two also the stationary model is tested by Algorithm 10. The selected model is model 31:

$$m(s,t) = a_0(t)+a_1(t)y+a_2(t)x^2+a_3(t)y^2+a_4(t)xy, \quad s=(x,y),\ t \in \mathcal{T} = [0,366],$$

which provided the best cross-validation results: thus, from a predictive point of view, a non-stationary model seems the most appropriate for describing the data. Moreover, by observing the residuals trace-variogram, a strong correlation among residuals can be recognized and the exponential structure appears suitable for fitting the empirical variogram. Therefore, in this case, GLS method is the most appropriate for estimating the drift, while Universal Kriging is needed to perform optimal spatial prediction.

Figure 16 shows the contour plots of the GLS drift estimate (upper panels) and of the Universal Kriging prediction (lower panels), obtained by fixing the time coordinate t to the Spring Equinox (21^{st} March, first panels), the Summer

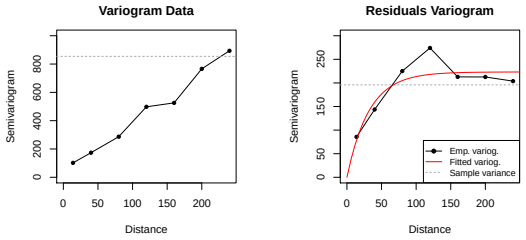


Figure 15: Estimated trace-variograms from data (on the left) and from residuals (on the right).

Solstice (21st June, second panels), the Autumn Equinox (23rd September, third panels) and the Winter Solstice (21st December, fourth panels). The first interesting result to be noticed is the climatical interpretation emerging from the obtained maps. The exposition of the Maritimes region towards the sea plays a key role indeed, due to the alternation of Atlantic warm-humid currents with freezing streams coming from the internal Canadian regions. These currents circulations significantly influences the temperatures and clearly reflects on drift contour lines (Figure 16, upper panels) with a clear rotation in drift contour lines, which begins during the springtime and continues until September under the influence of Gulf Stream from South (third panel): the early spring drift map (first panel) presents colder temperatures in the internal part of New Brunswick and warmer temperatures in the South; early summer panel (second panel) presents the opposite spatial behavior instead, featured by a warmer zone in the continental region and a cooler area along the sea. Moreover, notice the different rotation speed during the year –much faster in the transition from spring to summer and from summer to autumn than during the other months– that reflects the climatical trend in the region, featured by long lasting cold seasons and shorter warm periods.

Together with the drift rotation speed, the complexity in the spatial behavior (Figure 16, lower panels) seems to change along the temporal coordinate. Indeed, Universal Kriging maps relative to colder seasons (first, third and fourth panels) point out a much stronger influence of the drift component on the prediction with respect to the summer season (second panel); the latter is featured by very local structures instead, which seem to be strongly related to the geographical configuration of the area –notice in particular the low temperature zones marked off by the Bay of Fundy and by the Atlantic Ocean–. The interpretability of our results support the assertion made at the end of Section 4: our methodology applied to real data provides fairly accurate results also locally, although curves are handled as points of an infinite-dimensional space, under global assumptions.

Besides being climatically interpretable, the obtained results are consistent with the seasonal reference maps published by Natural Resources Canada, pro-

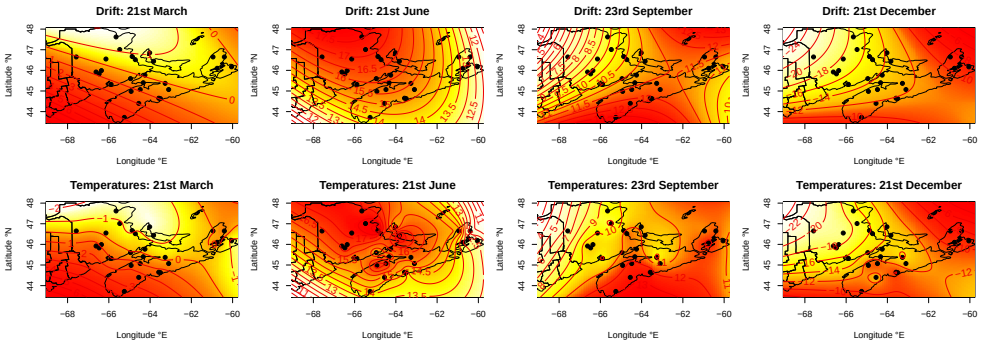


Figure 16: Drift estimation and Universal Kriging prediction contour plots for the Spring Equinox (21st March), to the Summer Solstice (21st June), to the Autumn Equinox (23rd September) and to the Winter Solstice (21st December).

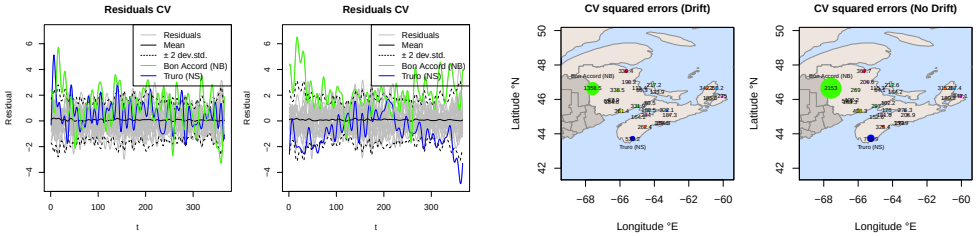
	UKFD (Geod. dist.)	OKFD (Geod. dist.)	OKFD (Eucl. dist.)	FKTM (Eucl. dist.)
Median	190.2 (↓ 8%, 13%, 13%)	205.9	218.2	218.1
Mean (MSE)	263.5 (↓ 14%, 18%, 19%)	306.8	323.2	323.8

Table 5: Comparison of cross-validation squared error statistics computed by Universal Kriging (UKFD), Ordinary Kriging (OKFD, Delicado et al. [2010]) –using geodetic and Euclidean distance – and Functional Kriging Total Model (FKTM, Delicado et al. [2010]). The reduction of UKFD error with respect to OKFD and FKTM is reported between brackets.

viding a further validation of the model.

Finally, cross-validation analysis has been performed, comparing our results with those obtained by applying a stationary model. Table 5 reports cross-validation statistics relative to Universal Kriging and Ordinary Kriging using the geodesic distance (first and second columns) and Ordinary Kriging and Functional Kriging Total Model based on the Euclidean metric (third and fourth columns). In the Euclidean setting, OKFD and FKTM show very similar predictive performances, with just slightly better results obtained in mean (MSE) with OKFD. By moving from Euclidean to geodesic distance, a first improvement in cross-validation results is achieved, but the most significant error reduction is due to the introduction of the drift term: indeed, moving from Ordinary to Universal Kriging –in the geodesic setting– the error decreases at least of 8% –if we consider the median value–, presenting a 14% reduction in mean.

Concerning the local errors along the temporal coordinate t, the modeling of a non-constant spatial mean makes the prediction unbiased and thus prevent the systematic overestimation or underestimation of the data, which occurs instead



(a) Cross-validation functional residuals

(b) Cross-validation squared errors maps

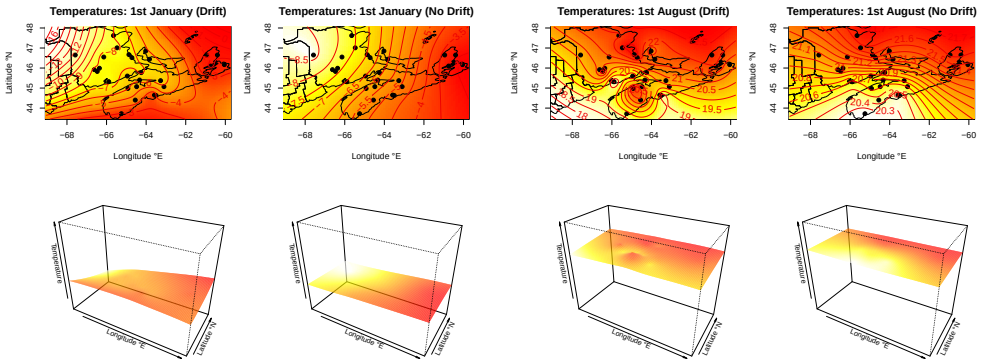
Figure 17: Comparison of cross-validation results for Universal Kriging and Ordinary Kriging, using in both cases the geodesic distance. In Subfigure (a): functional residuals for the 27 locations (grey lines) highlighting Bon Accord (NS) (green line) and Truro (NB) (blue line). In Subfigure (b): squared errors map; the dimension of the points is proportional to the cross-validation squared error; Bon Accord (NS) (green points) and Truro (NB) (blue points) are marked in the western part of the maps.

in OKFD prediction, respectively in Bon Accord and Truro (Figure 17a).

Moreover, by considering the spatial distribution of cross-validation errors (Figure 17b), it clearly appears that the most significant growth of the predictive power is obtained in peripheral zones, in particular for Bon Accord (NS) and Truro (NB) (western part of the maps). This kind of improvement is explained by the increased flexibility reached through the introduction of a drift term. This drives the prediction in peripheral areas and allows to reach more extreme predicted values, above all during the winter season where the drift is more influent on the prediction. For instance, observe the NW corner of Universal Kriging maps computed for the 1st January (Figure 18a): with OKFD the most extreme predicted temperatures are around -9°C, while UKFD prediction reaches values below -16°C.

On the other hand, the additional flexibility obtained by introducing the drift term contributes to mitigate the smoothing effect of kriging; this reflects on a very accurate local prediction that reproduces the local structures much better than the Ordinary Kriging interpolation. For example, look at the local structures that arise during the summer period between the Bay of Fundy and the Atlantic Ocean in Figure 18b: they are very well reproduced by UKFD prediction, while they are severely smoothed in the OKFD interpolation.

Therefore, the non-stationary prediction, obtained by applying our procedure, proves to be much more satisfactory than the stationary interpolations in terms both of global prediction error as well as of local behavior: Universal Kriging prediction is precise and flexible, besides being simple and easier to compute with respect to the most sophisticated stationary methods existing in the literature.



(a) 1st January (b) 1st August

Figure 18: Comparison of the results obtained with Universal Kriging and Ordinary Kriging, using in both cases the geodesic distance. Upper panels show contours maps, lower panels represent the associated 3D plots.

6 Conclusions and further research

In this work, a new kriging methodology for non-stationary spatially dependent functional data has been developed. On one hand, the theoretical effort has been spent for the formulation of a coherent framework, based on minimal assumptions. On the other hand, the developed algorithms aimed at making our theoretical results applicable on real data, through reliable and efficient procedures.

The development of inferential tools for spatially dependent functional data is still one of the most challenging topic to be addressed: the significance of regressors coefficients should be tested during drift model selection and kriging confidence bands should be provided together with point-wise prediction. To this end, a possible immediate perspective is given by the extension to the georeferenced functional case of non-parametric resampling methods like the bootstrap –e.g., [Efron and Tibshirani, 1993] and more recently, in the field of FDA, [Ferraty et al., 2010]–, which would allow to avoid distributional assumptions by means of a computer-intensive technique.

Developing statistical models and inferential procedures for general Hilbert spaces, instead of working out ad hoc techniques for the L^2 space, opens broad perspectives of research: indeed, it may allow the integration of the kriging methodology, which is in fact an interpolation technique, with the physical model underlying the observed phenomenon. In this direction, more complex linear models (e.g., FDA Total Model [Ramsay and Silverman, 2005]) would be worth investigating in order to model more precisely the drift term, possibly including more complex regressors which might influence or drive the physical system.

7 Appendix: Proofs

Proof. (*Proposition 7*) Consider the auxiliary optimal problem:

$$\min_{\widehat{\boldsymbol{a}}_{\vec{l}} \in H^{L+1}} \widetilde{\Phi}^{OLS}(\widehat{\boldsymbol{a}}_{\vec{l}}) \tag{39}$$

where:

$$\widetilde{\Phi}^{OLS}(\widehat{\boldsymbol{a}}_{\vec{l}}) = \|\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{a}}_{\vec{l}}\|^2_{H^n}, \tag{40}$$

with $\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} = \Sigma^{-1/2}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}}$, whose components are uncorrelated, and $\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}} = \Sigma^{-1/2}\mathbb{F}_{\vec{\boldsymbol{s}}}$.

It is easily seen the equivalence of the estimation problems (23) and (39), as $\Phi^{GLS}(\widehat{\boldsymbol{a}}_{\vec{l}}) = \widetilde{\Phi}^{OLS}(\widehat{\boldsymbol{a}}_{\vec{l}})$.

Assume that Σ is known and denote with $\widetilde{V}$ the closed subspace of H^n generated by linear combination of $\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}$ columns with coefficients in H and let $\widetilde{V}^{\perp}$ be its orthogonal complement:

$$\widetilde{V} = \{\widetilde{\boldsymbol{v}} \in H^n : \widetilde{\boldsymbol{v}} = \widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}} a_{\vec{l}}, \quad a_{\vec{l}} \in H^L\}, \tag{41}$$

$$\widetilde{V}^{\perp} = \{\widetilde{\boldsymbol{w}} \in H^n : \langle \widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{v}} \rangle_{H^n} = 0, \quad \forall \widetilde{\boldsymbol{v}} \in \widetilde{V}\}. \tag{42}$$

The estimator $\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} = \widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{a}}_{\vec{l}}$ of $\widetilde{m}_{\vec{\boldsymbol{s}}}$ –mean vector of $\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}$– minimizing $\widetilde{\Phi}^{OLS}$ is the projection of $\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}$ on $\widetilde{V}$, while the residual vector $\widehat{\widetilde{\boldsymbol{\delta}}}_{\vec{\boldsymbol{s}}} = \widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}}$ is the projection of $\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}$ on $\widetilde{V}^{\perp}$:

$$\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} = P_{\widetilde{V}}\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} \tag{43}$$

$$\widehat{\widetilde{\boldsymbol{\delta}}}_{\vec{\boldsymbol{s}}} = P_{\widetilde{V}^{\perp}}\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} \tag{44}$$

If $\text{rank}(\mathbb{F}_{\vec{\boldsymbol{s}}}) = L+1$ and $\text{rank}(\Sigma) = n$, then $\text{rank}(\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}) = L+1$, which ensures the existence and uniqueness of the projections (43) and (44).

Moreover, the projection (43) can be explicitly computed pre-multiplying $\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}$ for the orthogonal projection matrix $\widetilde{\mathbb{H}} = \widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}(\widetilde{\mathbb{F}}^T_{\vec{\boldsymbol{s}}}\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}})^{-1}\widetilde{\mathbb{F}}^T_{\vec{\boldsymbol{s}}}$, deriving directly the following linear expressions:

$$\begin{aligned} \widehat{\boldsymbol{a}}^{GLS}_{\vec{l}} &= (\widetilde{\mathbb{F}}^T_{\vec{\boldsymbol{s}}}\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}})^{-1}\widetilde{\mathbb{F}}^T_{\vec{\boldsymbol{s}}}\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}; \\ \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} &= \widetilde{\mathbb{H}}\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} = \widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}}(\widetilde{\mathbb{F}}^T_{\vec{\boldsymbol{s}}}\widetilde{\mathbb{F}}_{\vec{\boldsymbol{s}}})^{-1}\widetilde{\mathbb{F}}^T_{\boldsymbol{s}}\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}}. \end{aligned} \tag{45}$$

Applying the inverse transformation, expressions (25) and (26) can be finally obtained. □

Proof. (*Proposition 8*) Let $\widehat{\boldsymbol{a}}_{\vec{l}}$ be a generic linear estimator of the coefficients $a_{\vec{l}}$:

$$\widehat{\boldsymbol{a}}_{\vec{l}} = \mathbb{A}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} + \boldsymbol{b}, \tag{46}$$

with $\mathbb{A} \in \mathbb{R}^{L+1,n}$, $\boldsymbol{b} \in H^{L+1}$. The unbiasedness condition translates into the constraints:

$$\mathbb{A}\mathbb{F}_{\vec{\boldsymbol{s}}} = \mathbb{I}_n \tag{47}$$

$$\boldsymbol{b} = \boldsymbol{0}, \quad q.o. \tag{48}$$

where $\mathbb{I}_n$ is the identity matrix in $\mathbb{R}^n$, $\mathbf{0} \in H^{L+1}$ is the vector of $L+1$ identically zero functions.

By definition of optimality of $\widehat{\boldsymbol{a}}_{\vec{l}}^{BLUE}$, for every other linear unbiased estimator $\widehat{\boldsymbol{a}}_{\vec{l}}$, the matrix:

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}^{BLUE}),$$

is positive semi-definite, or equivalently:

$$\boldsymbol{x}^T(\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}^{BLUE}))\boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

For a generic linear estimator (46), under the unbiasedness constraints (48), the variance-covariance matrix is:

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}}) = \mathbb{A}\Sigma\mathbb{A}^T;$$

and, by the inequality [Shumway and Dean, 1968]:

$$\boldsymbol{\alpha}\mathbb{C}^{-1}\boldsymbol{\alpha}^T \geq \boldsymbol{\alpha}\mathbb{D}(\mathbb{D}^T\mathbb{C}\mathbb{D})^{-1}\mathbb{D}^T\boldsymbol{\alpha}^T,$$

that holds for $\mathbb{C} \in \mathbb{R}^{n,n}$ semidefinite positive, $\mathbb{D} \in \mathbb{R}^{n,L+1}$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$, a lower bound for $\boldsymbol{x}^T \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\vec{l}})\boldsymbol{x}$ can be obtained by setting $\mathbb{C} = \Sigma^{-1}$, $\alpha = \boldsymbol{x}^T\mathbb{A}$ and $\mathbb{D} = \mathbb{F}_{\vec{\boldsymbol{s}}}$:

$$\boldsymbol{x}^T\mathbb{A}\Sigma\mathbb{A}^T\boldsymbol{x} \geq \boldsymbol{x}^T\mathbb{A}\mathbb{F}_{\vec{\boldsymbol{s}}}(\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}^T\mathbb{A}^T\boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

The lower bound is reached for:

$$\mathbb{A}^{BLUE} = (\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}.$$

Hence, the optimal linear estimator is:

$$\widehat{\boldsymbol{a}}_{\vec{l}}^{BLUE} = (\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}})^{-1}\mathbb{F}_{\vec{\boldsymbol{s}}}^T\Sigma^{-1}\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} \equiv \widehat{\boldsymbol{a}}_{\vec{l}}^{GLS},$$

that in particular minimizes the mean square errors MSE_l simultaneously for every $l = 0, ..., L$:

$$MSE_l = \mathbb{E}[\|\widehat{\boldsymbol{a}}_l - \boldsymbol{a}_l\|^2] = (\mathbb{A}\Sigma\mathbb{A}^T)_{ll},$$

subject to the unbiasedness constraints (47) and (48).

As a consequence, by linearity, $\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}$ is the BLUE for the drift. □

Proof. (*Decomposition of variance* (29)) Let Σ^{GLS} be the $n \times n$ covariance matrix of the estimator $\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}$, $\Sigma^{GLS} = \mathbb{E}[\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}^T]$ and consider the following matrix notations:

$$\begin{aligned} gg^T &= (\langle g_i, g_j \rangle), \quad g = (g_1, ..., g_n) \in H^n \\ \mathbb{E}[\mathbb{A}] &= (\mathbb{E}[\mathbb{A}_{ij}]), \quad \mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^n, \end{aligned}$$

Then:

$$
\begin{aligned}
\Sigma := \mathrm{Cov}(\boldsymbol{\chi}_{\vec{\boldsymbol{s}}}) &= \mathbb{E}[(\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})(\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})^T] = \\
&= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})^T \Sigma^{T/2}] = \\
&= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} \pm \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} \pm \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})^T \Sigma^{T/2}] = \\
&= \Sigma^{1/2}\mathbb{E}[(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}})(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}})^T + (\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})(\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})^T]\Sigma^{T/2} = \\
&= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}})(\widetilde{\boldsymbol{\chi}}_{\vec{\boldsymbol{s}}} - \widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}})^T \Sigma^{T/2}] + \mathbb{E}[\Sigma^{1/2}(\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})(\widehat{\widetilde{\boldsymbol{m}}}_{\vec{\boldsymbol{s}}} - \widetilde{m}_{\vec{\boldsymbol{s}}})^T \Sigma^{T/2}] = \\
&= \mathbb{E}[(\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - \widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}})(\boldsymbol{\chi}_{\vec{\boldsymbol{s}}} - \widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}})^T] + \mathbb{E}[(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})^T] = \\
&= \mathbb{E}[\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}\widehat{\boldsymbol{\delta}}_{\vec{\boldsymbol{s}}}^T] + \mathbb{E}[(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}} - m_{\vec{\boldsymbol{s}}})^T] = \\
&= \Sigma^{GLS} + \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\vec{\boldsymbol{s}}}).
\end{aligned}
$$

□

References

M. Armstrong and P. Diamond. Testing variogram for positive-definiteness. *Mathematical geology*, 16(4):407–421, 1984.

L. Arnold. *Random Dynamical Systems*. Springer, second edition, 2003.

S. Banerjee. On geodetic distance computations in spatial modeling. *Biometrics*, 61:617–625, 2005.

D. Bosq. *Linear Processes in Function Spaces*. Springer, New York, 2000.

J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York, 1999.

N. Cressie. *Statistics for Spatial data*. John Wiley & Sons, New York, 1993.

F. Curriero. On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology*, 38:907–926, 2006.

P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data. *Environmetrics*, 21(3-4):224–239, 2010.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis : theory and practice*. Springer, New York, 2006.

F. Ferraty, I. Van Keilegom, and F. Vieu. On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics*, 37 (2):286–306, 2010.

R. Giraldo. *Geostatistical Analysis of Functional Data*. PhD thesis, Universitat Politècnica da Catalunya, Barcellona, 2009.

R. Giraldo, P. Delicado, and J. Mateu. Geostatistics for functional data: An ordinary kriging approach. Technical report, 2008a. Universitat Politècnica de Catalunya, http://hdl.handle.net/2117/1099.

R. Giraldo, P. Delicado, and J. Mateu. *Point-wise kriging for spatial prediction of functional data*, chapter 22, pages 135–142. Functional and operatorial statistics. Proceedings of the first international workshop on functional and operatorial statistics. Springer, Toulouse, France, 2008b.

R. Giraldo, P. Delicado, and J. Mateu. Geostatistics for functional data: An ordinary kriging approach. *Environmental and Ecological Statistics*, 2010a. In press.

R. Giraldo, P. Delicado, and J. Mateu. Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(1):66–82, 2010b.

R. Giraldo, P. Delicado, and J. Mateu. *geofd: a package for prediction for functional data*. PhD thesis, Universitat Politecnica de Catalunya, 2010c. URL `http://code.google.com/p/geofd`.

O. Gromenko, P. Kokoszka, L. Zhu, and J. Sojka. Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Annals of Applied Statistics*, 6(2):669–696, 2012.

S. Hörmann and P. Kokoszka. Consistency of the mean and the principal components of spatially distributed functional data. In F. Ferraty, editor, *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, pages 169–175. Physica-Verlag HD, 2011.

C. Huang, H. Zhang, and S. M. Robeson. On the validity of commonly used covariance and variogram functions on the sphere. *Mathematical Geosciences*, 43(6):721–733, 2011.

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings National Institute of Science,* India, 2(1):49–55, 1936.

P. Monestiez and D. Nerini. A cokriging method for spatial functional data with applications in oceanology. In S. Dabo-Niang and F. Ferraty, editors, *Functional and operatorial statistics*. Springer, 2008.

Natural Resources of Canada website. 2012.

J. Ramsay and B. Silverman. *Functional data analysis*. Springer, New York, second edition, 2005.

R. H. Shumway and W. Dean. Best linear unbiased estimation for multivariate stationary processes. *Technometrics*, 10(3):523–534, 1968.

D. Stanley. *Canada's Maritime provinces*. Marybirnong: Lonely Planet Pubblications, first edition, 2002.

Y. Yamanishi and Y. Tanaka. Geographically weighted functional multiple regression analysis: a numerical investigation. *Journal of Japanese Society of Computational Statistics*, (15):307–317, 2003.

K. Zhu. *Operator Theory in Function Spaces*. American Mathematical Society, second edition, 2007.