MOX-Report No. 53/2015

Universal kriging of functional data: trace-variography vs cross-variography? Application to forecasting in unconventional shales

Menafoglio, A; Grujic, O.; Caers, J.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Universal kriging of functional data: trace-variography vs cross-variography? Application to forecasting in unconventional shales

Alessandra Menafoglio[1*], Ognjen Grujic[2] and Jef Caers[2]

[1]MOX-Department of Mathematics, Politecnico di Milano, Italy
[2]School of Earth, Energy & Environmental Sciences, Stanford University, USA
[*]alessandra.menafoglio@polimi.it

Abstract

In this paper we investigate the practical and methodological use of universal kriging of functional data to predict unconventional shale production in undrilled locations from known production data. In universal kriging of functional data, two approaches are considered: 1) estimation by means of cokriging of functional components (Universal Cokriging, UCok), requiring cross-variography and 2) estimation by means of trace-variography (Universal Trace-Kriging, UTrK), which avoids cross-variogram modeling. While theoretically, under known variogram structures, such approaches may be quite equivalent, the practical application yields marked differences. We investigate these difference by means of a real field application in the Barnett shale play and by a Monte Carlo study inspired from such real field application. We find that, for the studied cases, in terms of sum of squared errors (SSE), UTrK outperforms UCok. We speculate that the main reason lies in the robustness of estimating experimental trace-variography over the cross-variography and the possible loss of information induced by the functional decomposition required for cokriging.
Keywords: Geostatistics; functional data; trace-variogram; shale gas; unconventional resources

1 Introduction

Functional data analysis (FDA, Ramsay and Silverman, 2005) has gained renewed attention in the modeling of phenomenon that can be regarded as statistical observations displaying systematic variation. In particular in terms of time series, FDA has been considered as an alternative to multivariate analysis, where in FDA the data is seen a single functional object with an underlying smooth dynamic that drives variation in time. While applications have primordially been in bio-informatics (see Ullah and Finch, 2013, for a recent review),

FDA has been gaining attention both in development of theory and application in the Earth & Environmental Sciences, for example in climate science (Besse et al., 2000), water resources (Josset et al., 2015; Satija and Caers, 2015), environmental science (Henderson, 2006; Yan et al., 2015; Sancho et al., 2015), oceanography (Nerini et al., 2010), land use (Besse et al., 2005) and geology (Manté and Stora, 2012; Menafoglio et al., 2014).

Particular to the application in the Earth Sciences is the spatial context and the need for spatial models for functional data, as has historically been developed in geostatistics (Matheron, 1969; Cressie, 1993). A recent body of theoretical work has been published, extending ordinary kriging to the functional case (e.g., Delicado et al., 2010; Nerini et al., 2010; Menafoglio et al., 2014; Menafoglio and Petris, 2015, and references therein). However, in several practical applications, there is a need to address phenomena that require non-stationary approaches in space. To address such need, universal kriging of spatial functional data has been proposed by Caballero et al. (2013); Menafoglio et al. (2013). An alternative approach to deal with non-stationarity is proposed by Ignaccolo et al. (2014).

In this work, we present a timely and economically important application of functional data, namely to the modeling and forecasting in unconventional shale resources. The term "unconventional" emanates from the way such resources are exploited: a sand-water mixture is injected into horizontal wells, fracturing nearly impermeable shale formations enabling production of commercially significant hydrocarbon volumes. Shale production can be considered as one of the driving factors for low oil/gas prices in 2015 (Mănescu and Nuño, 2015) which has put financial pressures on further resource development ("The Shale Industry Could Be Swallowed By Its Own Debt", Bloomberg news, July 18, 2015). As a consequence, technical innovation is required for such resources to remain competitive with conventional exploitation which tend to have lower costs. In addition, better modeling, understanding and more optimal drilling practices will lead to lesser environmental impact (see, e.g., Vidic et al., 2013). Part of such technical innovation lies in understanding the impact of the geological and hydraulic fracturing factors on production, which drives the spatial variability of production in wells. Due to the complexity involved, data scientific approaches are preferred over physical modeling approaches (Mohaghegh, 2011, 2013; Kormaksson et al., 2015; Grujic et al., 2015). Production rates in wells start from an initial peak in production right after hydraulic fracturing followed by a long multi-month decline. In this paper, we focus on modeling the spatial distribution of production decline rates only. We consider data from the prolific Barnett Shale and our data set contains 922 wells drilled over the basin. Such dataset consists of functional data (decline rate) varying over space (geographic coordinated).

The first aim of this paper is to investigate the use of universal kriging of functional data to the spatial interpolation of gas production rate curves (GPRC) which is required to estimate production for undrilled location. Our second aim is to compare two methods: 1) estimation by means of cokriging of the components over a functional basis (Universal Cokriging, UCok), which is similar to the strategy of Nerini et al. (2010) and requires cross-variography, and 2) estimation by means of trace-variography (Universal Trace-Kriging, UTrK), which follows the approach of Menafoglio et al. (2013) and avoids cross-variogram modeling. Our comparison is with respect to the Barnett shale data set and a Monte Carlo study designed based on that dataset.

In that vein, the paper structure is as follow. We first review both methods as well as provide an overview of the theoretical and methodological implications for comparison. Then we apply and evaluate both methods to the Barnett shale and provide and in-depth Monte Carlo comparison.

2 Methodology

In this Section, we pursue the functional and geostatistical approach to the analysis of GPRCs, and explore two alternative methods for their spatial prediction.

We call $(\Omega, \mathfrak{F}, \mathbb{P})$ a probability space, Ω denoting a space of events, $\mathfrak{F}$ a σ-algebra, and $\mathbb{P}$ a probability measure. We indicate by $x_{\boldsymbol{s}_1}, ..., x_{\boldsymbol{s}_n}$ the observed GPRCs[1] at a set of given locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ in an Euclidean spatial domain $D \subseteq \mathbb{R}^2$. As in classical geostatistics, we assume the data to be a partial observation of a random field

$$\{X_{\boldsymbol{s}}, \boldsymbol{s} \in D\}, \tag{1}$$

on $(\Omega, \mathfrak{F}, \mathbb{P})$, where the index $\boldsymbol{s}$ indicates a location in D. As the observations are curves, the random field (1) is assumed to be valued in an infinite-dimensional (functional) space. Specifically, throughout this work we assume that, for any location $\boldsymbol{s} \in D$, the element $X_{\boldsymbol{s}}$ is a random element of the space $L^2(T)$ of squared-integrable real-valued functions on the time interval $T = [1, 60]$. The space $L^2(T)$ (or L^2 for short) is a separable Hilbert space if equipped with the usual inner product $\langle f, g \rangle = \int_T f(t)g(t)dt$ and the induced norm $\|f\| = \sqrt{\langle f, f \rangle}$, for $f, g \in L^2(T)$.

Following Menafoglio et al. (2013), we assume that process (1) is non-stationary, and represent the its element $X_{\boldsymbol{s}}$, at a generic location $\boldsymbol{s} \in D$, as sum of its mean $m_{\boldsymbol{s}}$, called drift, and a zero-mean stochastic residual $\delta_{\boldsymbol{s}}$ that is stationary in the sense that is specified below, i.e.,

$$X_{\boldsymbol{s}} = m_{\boldsymbol{s}} + \delta_{\boldsymbol{s}}.$$

Here, the mean $m_{\boldsymbol{s}}$ of $X_{\boldsymbol{s}}$ can be defined point-wise as $m_{\boldsymbol{s}}(\cdot) = \mathbb{E}[X_{\boldsymbol{s}}(\cdot)]$, and is a non-random element of $L^2(T)$. Hereafter in this work, we assume that the mean is non-constant in space, and we model its spatial variation through a functional linear model in $L^2(T)$ of the form

$$m_{\boldsymbol{s}}(t) = \sum_{l=0}^{L} a_l(t) f_l(\boldsymbol{s}), \quad t \in T, \boldsymbol{s} \in D, \tag{2}$$

where $\{a_l, l = 0, ..., L\}$ are functional coefficients in $L^2(T)$, independent of the spatial location, while $\{f_l(\boldsymbol{s}), l = 0, ..., L\}$ are known scalar regressors depending on $\boldsymbol{s} \in D$. Finally, we assume that the residuals form a zero-mean random field $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ on $(\Omega, \mathfrak{F}, \mathbb{P})$, with stationary spatial covariance function. The latter is defined as the function that maps any increment between locations in D, $(\boldsymbol{s}_1 - \boldsymbol{s}_2) \in \mathbb{R}^2$, into the cross-covariance operator on $L^2(T)$ between the elements of the process at those locations:

$$C(\boldsymbol{s}_1 - \boldsymbol{s}_2)x = \mathbb{E}[\langle \delta_{\boldsymbol{s}_1}, x \rangle \delta_{\boldsymbol{s}_2}], \quad x \in L^2(T).$$

[1] To preserve the positivity of GPRCs, the model and the subsequent procedures can be applied to the log-transformed GPRCs.

Under these assumptions, and given the random observations $X_{\boldsymbol{s}_1}, ..., X_{\boldsymbol{s}_n}$ at the sampled locations, we aim to predict the unobserved element $X_{\boldsymbol{s}_0}$ of process (1) at a target location $\boldsymbol{s}_0 \in D$. To this end, in the next Subsections we will introduce and explore two alternative approaches to Kriging functional data, namely the Universal Cokriging based on a projection strategy (UCok), and the Universal Trace-Kriging (UTrK) method of Menafoglio et al. (2013).

2.1 Universal Cokriging of functional data

We here present an extension to the non-stationary setting of the Kriging predictor for functional data proposed by Giraldo (2009) and Nerini et al. (2010).

Throughout this Subsection, we consider an orthonormal set $\{e_k, 1 \leq k \leq K\}$ in $L^2(T)$, and we assume that each element of process (1) can be represented through the expansion over this set, i.e.,

$$X_{\boldsymbol{s}} = \sum_{k=1}^{K} \xi_k(\boldsymbol{s}) e_k, \quad \boldsymbol{s} \in D, \tag{3}$$

where $\xi_k(\boldsymbol{s}) = \langle X_{\boldsymbol{s}}, e_k \rangle$ is the projection of $X_{\boldsymbol{s}}$ on the k-th element of the basis.

Under representation (3), one can characterize process (1) through the distributional properties of the K-dimensional random field of coefficients $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$, with $\boldsymbol{\xi}(\boldsymbol{s}) = (\xi_1(\boldsymbol{s}), ..., \xi_K(\boldsymbol{s}))^T$. For instance, one can define the drift $m_{\boldsymbol{s}}$ of the process at $\boldsymbol{s}$ in D, through the drift of the field $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$ at the same location, $\boldsymbol{m}^{\boldsymbol{\xi}}(\boldsymbol{s}) = (m_1^{\boldsymbol{\xi}}(\boldsymbol{s}), ..., m_K^{\boldsymbol{\xi}}(\boldsymbol{s}))^T$,

$$m_{\boldsymbol{s}}(\cdot) = \sum_{k=1}^{K} \mathbb{E}[\xi_k(\boldsymbol{s})] e_k(\cdot) = \sum_{k=1}^{K} m_k^{\boldsymbol{\xi}}(\boldsymbol{s}) e_k(\cdot). \tag{4}$$

Note that under model (2), the drift of the coefficients field is described by the linear model

$$m_k^{\boldsymbol{\xi}}(\boldsymbol{s}) = \sum_{l=0}^{L} \alpha_{kl} f_l(\boldsymbol{s}), \quad \boldsymbol{s} \in D, \tag{5}$$

with $\alpha_{kl} = \langle a_l, e_k \rangle$. Indeed, one has

$$\langle m_{\boldsymbol{s}}, e_k \rangle = \sum_{l=0}^{L} \langle a_l, e_k \rangle f_l(\boldsymbol{s}) = m_k^{\boldsymbol{\xi}}(\boldsymbol{s}),$$

where the first equality follows from Eq. (2), and the second equality is obtained by using Eq. (4). We thus represent the elements of the multivariate field $\{\boldsymbol{\xi}(\boldsymbol{s}), \boldsymbol{s} \in D\}$ as

$$\boldsymbol{\xi}(\boldsymbol{s}) = \boldsymbol{m}^{\boldsymbol{\xi}}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}), \tag{6}$$

where $\{\boldsymbol{\epsilon}(\boldsymbol{s}), \boldsymbol{s} \in D\}$ forms a zero-mean K-dimensional random field.

If the assumption of second-order stationarity (in $L^2(T)$) for $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ holds true, then also $\{\boldsymbol{\epsilon}(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is second-order stationary (in $\mathbb{R}^K$). Indeed, under representation (3), one has the following matrix representation of the spatial covariance function C

$$C(\boldsymbol{s}_1 - \boldsymbol{s}_2)x = \sum_{j=1}^{K} \sum_{k=1}^{K} C_{jk}^{\boldsymbol{\xi}}(\boldsymbol{s}_1 - \boldsymbol{s}_2) x_j e_k, \quad \boldsymbol{s}_1, \boldsymbol{s}_2 \in D, \tag{7}$$

where $C^{\boldsymbol{\xi}}_{jk}(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \langle C(\boldsymbol{s}_1 - \boldsymbol{s}_2)e_j, e_k\rangle$. The latter quantity is equivalently found as the covariance between the coefficients $\xi_j(\boldsymbol{s}_1)$ and $\xi_k(\boldsymbol{s}_2)$:

$$C^{\boldsymbol{\xi}}_{jk}(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \mathbb{E}[(\xi_j(\boldsymbol{s}_1) - m^{\boldsymbol{\xi}}_j(\boldsymbol{s}_1))(\xi_k(\boldsymbol{s}_2) - m^{\boldsymbol{\xi}}_k(\boldsymbol{s}_2))].$$

As such, the cross-covariogram of ξ_j and ξ_k is stationary for all $j, k = 1, ..., K$.

Given the observations $X_{\boldsymbol{s}_1}, ..., X_{\boldsymbol{s}_n}$, we aim to predict $X_{\boldsymbol{s}_0}$ via the Best Linear Unbiased Predictor (BLUP) in the sense of Nerini et al. (2010), that is $X^*_{\boldsymbol{s}_0} = \sum_{i=1}^n \Lambda^*_i X_{\boldsymbol{s}_i}$, where $\Lambda^*_1, ..., \Lambda^*_n$ are the operators that solve the constrained optimization problem

$$\min \ \mathbb{E}\left[\left\| X_{\boldsymbol{s}_0} - \sum_{i=1}^n \Lambda_i X_{\boldsymbol{s}_i} \right\|^2\right] \quad \text{s.t.} \quad \mathbb{E}\left[X_{\boldsymbol{s}_0} - \sum_{i=1}^n \Lambda_i X_{\boldsymbol{s}_i}\right] = 0, \tag{8}$$

among all the linear Hilbert-Schmidt operators $\Lambda_1, ..., \Lambda_n$ on $L^2(T)$. We call Universal Cokriging (UCok) predictor the solution $X^*_{\boldsymbol{s}_0}$ of problem (8).

As shown by Nerini et al. (2010) in the stationary setting, finding the UCok predictor is equivalent to determine an optimal estimate of the coefficients vector $\boldsymbol{\xi}(\boldsymbol{s}_0)$ at the target location $\boldsymbol{s}_0$ by solving the following Cokriging problem

$$\min \ \mathbb{E}\left[\left\| \boldsymbol{\xi}(\boldsymbol{s}_0) - \sum_{i=1}^n \mathbb{L}_i \boldsymbol{\xi}(\boldsymbol{s}_i) \right\|^2_{\mathbb{R}^d}\right] \quad \text{s.t.} \quad \mathbb{E}\left[\boldsymbol{\xi}(\boldsymbol{s}_0) - \sum_{i=1}^n \mathbb{L}_i \boldsymbol{\xi}(\boldsymbol{s}_i)\right] = \mathbf{0}, \tag{9}$$

among all the matrices of weights $\mathbb{L}_1, ..., \mathbb{L}_n$ in $\mathbb{R}^{K\times K}$. This follows from the observation that (i) the operators Λ_i, $i = 1, ..., n$, admit a matrix representation through $\mathbb{L}_i$, $i = 1, ..., n$, analogous to that of C in Eq. (7); (ii) by exploiting expression (3) and Parsival identity, one has that the objective functionals in Eqs. (8) and (9) coincide; and (iii) developing the unbiasedness constraint in Eq. (8) in the light of Eq. (3) one gets

$$\mathbb{E}\left[X_{\boldsymbol{s}_0} - \sum_{i=1}^n \Lambda_i X_{\boldsymbol{s}_i}\right] = \sum_{k=1}^K \mathbb{E}\left[\xi_k(\boldsymbol{s}_0) - \left[\sum_{i=1}^n \mathbb{L}_i \boldsymbol{\xi}(\boldsymbol{s}_i)\right]_k\right] e_k,$$

which is the null function if and only if $\mathbb{E}\left[\boldsymbol{\xi}(\boldsymbol{s}_0) - \sum_{i=1}^n \mathbb{L}_i \boldsymbol{\xi}(\boldsymbol{s}_i)\right]$ is the null vector in $\mathbb{R}^K$.

Therefore, the UCok predictor $X^*_{\boldsymbol{s}_0}$ of $X_{\boldsymbol{s}_0}$ can be found in this setting as $X^*_{\boldsymbol{s}_0} = \sum_{k=1}^K \xi^*_k(\boldsymbol{s}_0)e_k$, where $\boldsymbol{\xi}^*(\boldsymbol{s}_0) = \sum_{i=1}^n \mathbb{L}^*_i \boldsymbol{\xi}(\boldsymbol{s}_i)$ and the optimal matrices of weights are found by solving the Universal Cokriging system (Chilès and Delfiner, 1999)

$$\begin{pmatrix} \mathbb{C}_{11} & \cdots & \mathbb{C}_{1n} & \mathbb{F}_{10} & \cdots & \mathbb{F}_{1L} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}_{n1} & \cdots & \mathbb{C}_{nn} & \mathbb{F}_{n0} & \cdots & \mathbb{F}_{nL} \\ \mathbb{F}_{10} & \cdots & \mathbb{F}_{n0} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{F}_{1L} & \cdots & \mathbb{F}_{nL} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mathbb{L}_1 \\ \vdots \\ \mathbb{L}_n \\ \mathbb{Z}_0 \\ \vdots \\ \mathbb{Z}_L \end{pmatrix} = \begin{pmatrix} \mathbb{C}_{01} \\ \vdots \\ \mathbb{C}_{0n} \\ \mathbb{F}_{00} \\ \vdots \\ \mathbb{F}_{0L} \end{pmatrix}, \tag{10}$$

where $\mathbb{C}_{ij}$ is the cross-covariance matrix between $\boldsymbol{\xi}(\boldsymbol{s}_i)$ and $\boldsymbol{\xi}(\boldsymbol{s}_j)$, $i, j = 1, ..., n$; $\mathbb{C}_{i0}$ is the cross-covariance matrix between $\boldsymbol{\xi}(\boldsymbol{s}_0)$ and $\boldsymbol{\xi}(\boldsymbol{s}_i)$, $i = 1, ..., n$; $\mathbb{F}_{il} = diag(f_l(\boldsymbol{s}_i), ..., f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{K\times K}$ and $\mathbb{Z}_l$, $l = 0, ..., L$ are the matrices of Lagrange multipliers.

2.2 Universal Trace-Kriging

In this Subsection, we recall another approach to Kriging, following the work of Menafoglio et al. (2013). Here, we get rid from the assumption of the basis representation (3) by defining a different measure of spatial dependence.

We call trace-covariogram of the residual field $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, the real-valued function C_{tr} defined, for $\boldsymbol{s}_1, \boldsymbol{s}_2$ in D and in the previous assumptions, as

$$C_{tr}(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \mathbb{E}[\langle \delta_{\boldsymbol{s}_1}, \delta_{\boldsymbol{s}_2} \rangle].$$

The trace-covariogram defines a global measure of spatial dependence, in the sense that, for any fixed increment $(\boldsymbol{s}_1 - \boldsymbol{s}_2)$ in $\mathbb{R}^2$, it is the trace of the corresponding cross-covariance operator $C(\boldsymbol{s}_1 - \boldsymbol{s}_2)$. The trace-covariogram plays the same role as the univariate covariogram, but in the infinite-dimensional setting. Here, the corresponding trace-variogram is defined as

$$2\gamma_{tr}(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \mathbb{E}[\|\delta_{\boldsymbol{s}_1} - \delta_{\boldsymbol{s}_2}\|^2],$$

and describes the expected increment in the value of the (functional) process for a given increment in the spatial domain. On these bases, Menafoglio et al. (2013) define and explore global notion of stationarity for functional random fields, weaker than those considered so far. For sake of clarity in the following comparisons, we here keep the same stationarity assumption on the residual field as those introduced before.

To predict the unobserved element $X_{\boldsymbol{s}_0}$, given the observations $X_{\boldsymbol{s}_1}, ..., X_{\boldsymbol{s}_n}$, Menafoglio et al. (2013) consider a Kriging predictor in the form of a linear combination of the observations. We call Universal Trace-Kriging (UTrK) predictor the linear unbiased predictor $X^{*tr}_{\boldsymbol{s}_0} = \sum_{i=1}^n \lambda_i^* X_{\boldsymbol{s}_i}$, whose weights $\lambda_1^*, ..., \lambda_n^*$ solve

$$\min \ \mathbb{E}\left[\left\| X_{\boldsymbol{s}_0} - \sum_{i=1}^n \lambda_i X_{\boldsymbol{s}_i} \right\|^2\right] \quad \text{s.t.} \quad \mathbb{E}\left[X_{\boldsymbol{s}_0} - \sum_{i=1}^n \lambda_i X_{\boldsymbol{s}_i}\right] = 0, \tag{11}$$

over all the scalar weights $\lambda_1, ..., \lambda_n$ in $\mathbb{R}$. The authors then prove that problem (11) is well-posed even if one relies only upon the global definitions of spatial dependence introduced above. Indeed, the weights $\lambda_1^*, ..., \lambda_n^*$ are found by solving

$$\begin{pmatrix} C_{tr}(\mathbf{0}) & \cdots & C_{tr}(\boldsymbol{h}_{1n}) & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ C_{tr}(\boldsymbol{h}_{n1}) & \cdots & C_{tr}(\mathbf{0}) & f_1(\boldsymbol{s}_n) & \cdots & f_L(\boldsymbol{s}_n) \\ f_1(\boldsymbol{s}_1) & \cdots & f_1(\boldsymbol{s}_n) & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_L(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_n) & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \zeta_0 \\ \vdots \\ \zeta_L \end{pmatrix} = \begin{pmatrix} C_{tr}(\boldsymbol{h}_{10}) \\ \vdots \\ C_{tr}(\boldsymbol{h}_{n0}) \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{pmatrix}, \tag{12}$$

with $\boldsymbol{h}_{ij} = \boldsymbol{s}_i - \boldsymbol{s}_j$, and $\zeta_0, ..., \zeta_L$ the Lagrange multipliers associated with the $(L+1)$ unbiasedness constraints.

2.3 Methodological comparison

Even though the methods devised in Subsections 2.1 and 2.2 do achieve the same purpose, namely the geostatistical prediction of functional data, they are intrinsically different. We here compare the different perspectives they are grounded

on, to underline the methodological strengths and weaknesses of the two approaches to deal with real studies.

We first note that the derivation of the UCok predictor relies upon the assumption that the elements of the process admit expansion (3), for a given order K and an orthonormal set $\{e_k, 1 \leq k \leq K\}$. In FDA, this is a viable assumption, as most analyses in this field employ a preprocessing step based on an expansion over a truncated functional basis (e.g., Fourier basis). This step enables one to smooth the actual discrete observations, by removing from the measured curves the effects that are chiefly due to the measurement error. Non-orthonormal bases may be employed in the smoothing procedure (e.g., B-splines). Nevertheless, one can always perform a change of basis to map the observations on an orthonormal set, and accordingly represent the process via expansion (3).

Even though expansion (3) does not imply a substantial loss of generality in most real studies, the dimension of the expansion is influential indeed on the analysis. Indeed, as noted by Menafoglio and Petris (2015) in the stationary setting, even in the ideal case of known drift and covariance structure, the parameter K controls the dimension of system (10) (i.e., the number $K(n+L+1)$ of equations and unknowns), hence the problem complexity. Thus, in a real case study, one may need to employ a dimensionality reduction method prior to the geostatistical analysis. For instance, one can perform Functional Principal Component Analysis (FPCA, Ramsay and Silverman, 2005) as detailed in the Appendix, or the Functional Singular Value Decomposition (SVD, Yang et al., 2011). In all these cases, part of the information is inevitably lost as a consequence of the dimensionality reduction, and cannot be employed for prediction purposes. This marks the first difference between Universal Cokriging and Universal Trace-Kriging methodologies, as the latter does not require to express the observations through a basis representation.

Furthermore, the dimensionality of the Trace-Kriging system does not depend on the representation of the data. Indeed, system (12) is a linear system of $(n+L+1)$ equations and the same number of unknowns, independently on the possible truncated basis expansion employed to represent the data. Notice that this is possible because of the simple form of the UTrK predictor, as opposed to the more complex form of the UCok predictor. This is both a weakness and a strength of Trace-Kriging. On one hand, the UCok predictor is more general and, in principle, able to achieve a better prediction quality than the UTrK predictor. On the other hand, the simple form of the UTrK allows to exploit the entire information embedded into the data, without the need to reduce the dimensionality of the dataset prior to the geostatistical analysis.

In the context of Gaussian stationary random fields and under the representation (3), with $K \leq n$, a formal relation between the Cokriging and the Trace-Kriging predictors has been established by Menafoglio and Petris (2015). In this framework, the authors prove that the Cokriging and Trace-Kriging approaches lead almost surely to the same results, provided that the spatial covariance function C is known. In most applications, the spatial dependence is actually not known *a priori* and one needs to infer C from available data (or basis coefficients). To this end, different viewpoints on the Kriging problem induce different ways to estimate the spatial covariance. If one is willing to solve the Universal Cokriging system (10), the covariograms and cross-covariogram of the coefficients will be the target estimates. To this end, a Linear Model of

Coregionalization (LMC) can be introduced for the vector of coefficients, and a parametric semivariogram structure can be fitted to the empirical estimates. Note that the dimension K of the representation (3) directly reflects on the number of variogram and cross-variogram structures that one needs to estimate to solve Eq. (10). In contrast, to solve system (12) one will only estimate the trace-covariogram, or the trace-variogram as usually preferred. Estimating the latter follows the same line as in finite dimension. First an empirical estimate is computed from (estimated) residuals as

$$\gamma(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(\boldsymbol{h})} \|\delta_{\boldsymbol{s}_i} - \delta_{\boldsymbol{s}_j}\|^2, \tag{13}$$

where $\boldsymbol{h}$ denotes the lag, $N(\boldsymbol{h})$ the set of couples at lag $\boldsymbol{h}$ and $|N(\boldsymbol{h})|$ the number of elements of set $N(\boldsymbol{h})$. Second, a valid variogram model is fitted to the empirical estimate. Here, the well-known one-dimensional parametric families, such as spherical or Matérn, can be employed. Note that estimator (13) requires to compute a number of integrals (recall: $\|f-g\|^2 = \int_T (f(t)-g(t)^2 dt$, for f, g in $L^2(T)$). These can be computed in terms of basis coefficients whenever a basis representation of the kind (3) is employed, or via quadrature schemes otherwise.

To estimate the drift, the residuals and the variogram structure, one can employ very similar strategies in both the discussed approaches. In this work, we estimate the drift $\boldsymbol{m}^{\boldsymbol{\xi}}$ via Generalized Least Squares (GLS), and employ the classical iterative algorithm to jointly estimate the drift via GLS and the variograms/cross-variograms of the corresponding residuals. Similarly, we estimate via GLS the drift $m_{\boldsymbol{s}}$, as follows (see Menafoglio et al., 2013). We call Σ the global covariance matrix $\Sigma_{ij} = \mathbb{E}[\langle \delta_{\boldsymbol{s}_i}, \delta_{\boldsymbol{s}_j} \rangle]$, $\mathbb{F}$ the design matrix $\mathbb{F}_{il} = f_l(\boldsymbol{s}_i)$, $i = 1, ..., n$, $l = 0, ..., L$, $\boldsymbol{X}$ the vector of functional observations $\boldsymbol{X} = (X_{\boldsymbol{s}_1}, ..., X_{\boldsymbol{s}_n})^T \in L^2 \times \cdots \times L^2$ and we introduce the following matrix notation: $[\mathbb{A}\boldsymbol{f}]_i = \sum_{j=1}^n \mathbb{A}_{ij} \boldsymbol{f}_j$, for $\mathbb{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{f} \in L^2 \times \cdots \times L^2$. The GLS estimator of the drift at the observed locations, $\boldsymbol{m} = (m_{\boldsymbol{s}_1}, ..., m_{\boldsymbol{s}_n})^T$, is (Menafoglio et al., 2013)

$$\widehat{\boldsymbol{m}} = \mathbb{F}(\mathbb{F}^T \Sigma^{-1} \mathbb{F})^{-1} \mathbb{F}^T \Sigma^{-1} \boldsymbol{X}. \tag{14}$$

Similarly as in the classical setting, $\widehat{\boldsymbol{m}}$ can be computed by resorting to an iterative algorithm: having initialized $\widehat{\boldsymbol{m}}$ (e.g., to the ordinary least square estimate obtained by setting Σ to a multiple of the identity matrix in Eq. (14)), at each step the residuals are estimated by difference from the GLS estimate of the drift, and then the trace-variogram is estimated from the latest update of the residuals estimate. We refer to Menafoglio et al. (2013) for the algorithmic details.

In the next Section we compare the results of applying these different strategies to the dataset of GPRCs available at the Barnett Shale field site.

3 Application

3.1 Dataset Description

In this study we use gas production rate curves (GPRC) from 922 wells drilled in the Barnett shale, one of the most prolific and the most developed unconventional gas reservoirs in North America. This dataset was compiled from the

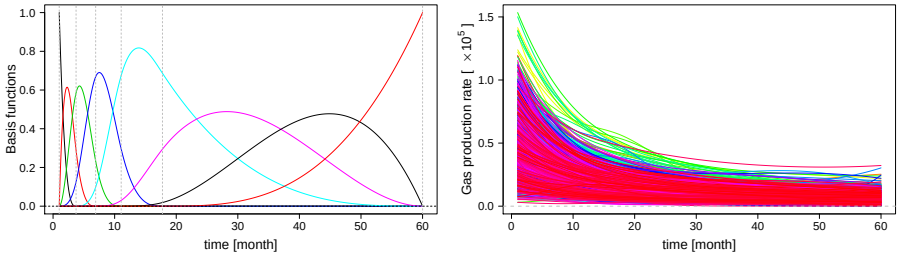

Figure 1: Left: final basis system found by the GCV on GPRCs; Right: Smoothed GPRCs. Gas rates are given in MMCF per month.

"drillinginfo.com" (further DI), an online oil and gas data repository. At the time when this dataset was prepared, DI did not provide information about well specific hydraulic fracturing parameters, which would have enabled us to search for wells with similar completions. Therefore, we decided to query for wells whose lateral length was anywhere between 1800 and 2300 feet and were owned by the same company (operator), with an assumption that the number of hydraulic fractures was the same or at least very similar across all wells. We considered wells drilled after 2005, which had at least 5 years of production history (60 months), immediately following the peak gas rate. As a part of pre-processing, all data entries preceding the peak gas rates (about 3 months) were discarded. Such pre-processing approach is very common in unconventional reservoir data analyses (see Patzek et al., 2013), since during that time period wells mostly produce flow-back water that comes as a consequence of hydraulic fracturing.

Most functional data analyses start with a basis expansion, or data fitting. In ideal conditions[2], GPRCs are smooth and monotonically decreasing positive curves. Obviously, in such setting there is a complete absence of periodicity and B-Spline basis system comes as a natural choice for representing such data. To honor the data positivity, we elected to perform basis expansion (Ramsay and Silverman, 2005) on the log-transformed observations, with a smoothing penalty on the second derivative. Preliminary data analysis revealed that most variation in GPRCs occurred during the first 12 months of production history. Therefore, we decided to place the knots of B-spline basis functions irregularly over analyzed time domain (60 months), with higher placement density over the first 12 months (Figure 1 Left). Finally, the number of basis functions for this dataset was set to $n = 10$, and the best smoothing penalty on the second derivative ($\lambda = 10^3$) was found with generalized cross validation (GCV, Ramsay and Silverman, 2005). Figure 1 shows final B-spline basis system (left) with resulting smoothed GPRCs (right).

3.2 Results

We first analyzed the smoothed dataset according to the approach devised in Subsection 2.2. For the analyses that follow we considered the log-transformed GPRCs (further log-GPRCs) to honor the positivity constraint. Hereafter we display the results in the original scale to ease their interpretation.

[2]Uninterrupted production with constant bottom hole or well head pressures.

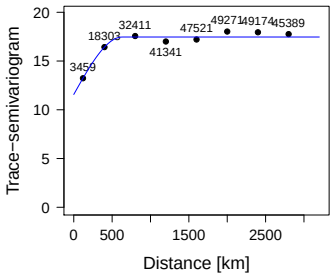

Figure 2: Estimated trace-variogram of log-GPRCs at the Barnett shale: empirical estimate (symbols), calibrated model (solid line). Numbers indicate the number of couples of locations upon which the corresponding empirical estimate is based.

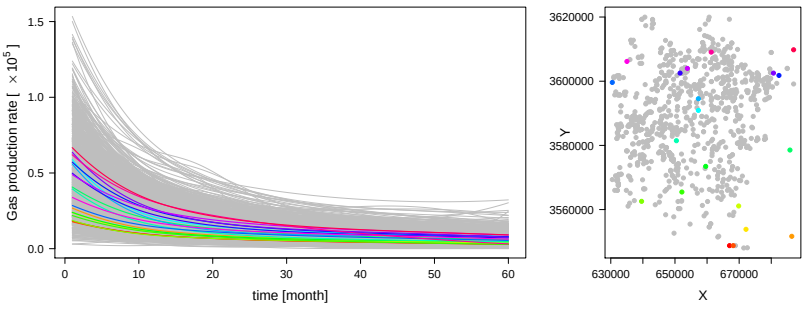

Figure 3: Prediction by UTrK of GPRCs for 20 random locations at the Barnett shale. Left: smoothed data (grey lines) and predictions (colored lines). Right: sampled location (grey symbols) and target locations (colored symbol).

Based on preliminary analyses at the site, we selected for the drift term the set of linear regressors in the spatial coordinates: $f_0(\boldsymbol{s}) = 1; f_1(\boldsymbol{s}) = x; f_2(\boldsymbol{s}) = y$, $\boldsymbol{s} = (x, y)$ denoting a location in D. Following the strategy detailed in Subsection 2.3, we jointly estimate the drift and the trace-variogram, fitting to the latter a spherical model with nugget. Figure 2 shows the empirical trace-variogram along with the fitted model.

Based on the calibrated trace-variogram model, we perform a UTrK prediction over a uniform spatial grid of 10^4 locations at the field site. Figure 3 reports a subsample of 20 predicted GPRCs, for a randomly selected set of 20 location of the prediction grid. Figure 4 reports the maps of the predicted functional field, taken for the time instants $t = 1, 12, 24, 48$ months (colors are given on a non-uniform color scale). Note that, unlike multivariate techniques, a functional prediction allows to obtain simultaneous kriging maps for any instant point $t \in [1, 60]$ (months).

To compare these predictions with those obtained through UCok, we performed FPCA to reduce the dimensionality of the dataset and represent the log-GPRCs over an orthonormal system (see the Appendix for the modeling details of FPCA). We select the first two FPCs, which together explain 94% of

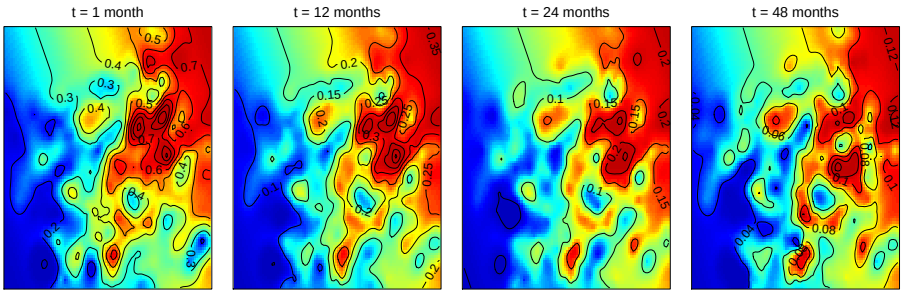

Figure 4: Prediction maps obtained with UTrK at the Barnett shale, for $t = 1, 12, 24, 48$ months. Colors are given on a non-uniform scale. Gas rate reported at contour lines are meant up to a factor 10^5.

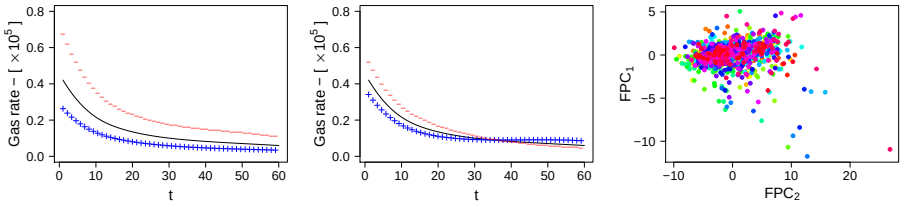

Figure 5: FPCA of log-GPRCs. Left: sample mean of log-GPRCs plus/minus the first FPC, represented in the original scale. Center: sample mean of log-GPRCs plus/minus the second FPC, represented in the original scale. Right: scores along the first two FPCs.

the data variability. To ease the interpretation of FPCs, Figure 5 reports, in the original space, the sample mean of the log-GPRCs plus/minus the FPCs. Visual inspection of Figure 5 suggests that the first FPC can be interpreted in terms of the amplitude of the peak gas and the overall production rate, high scores being associated with low peak gas and production rates. The second FPC is instead interpreted in terms of a contrast between the production rate in the first 35 months and the further production rate. Here, high scores correspond to curves with early production rate lower than the mean, and further production rate higher than the mean.

Based on the results of FPCA, we geostatistically analyzed the scores along the selected components, reported in Figure 5 (right panel). Consistent with the previous assumption, we consider for the drift the set of modified linear spatial regressors $f_1(\boldsymbol{s}) = x - \frac{1}{n}\sum_{i=1}^{n} x_i$; $f_2(\boldsymbol{s}) = y - \frac{1}{n}\sum_{i=1}^{n} y_i$, with $\boldsymbol{s} = (x, y)$ in D (see the appendix for the details). Variograms and cross-variograms of the residuals – referred to the multivariate model (5) for the scores – are estimated by fitting a LMC, based on a spherical model with nugget. Figure 6 reports the calibrated multivariate model. On this basis, we performed the UCok of the scores on the same spatial grid introduced before, obtaining the results displayed in Figure 7. Here we represent the same set of curves reported in Figure 3. Finally, Figure 8 displays the maps of the predicted gas production rate, at the time instants $t = 1, 12, 24, 48$ months (colors are given on a non-uniform color

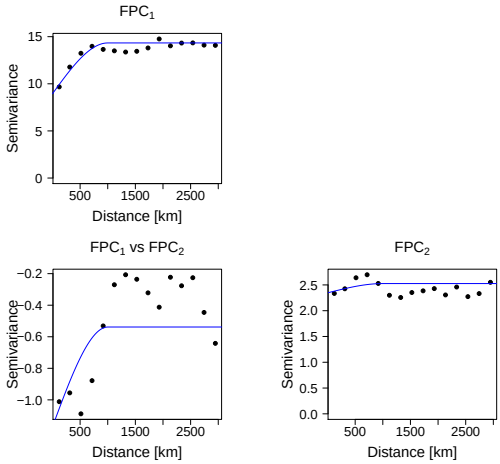

Figure 6: Estimated variograms and cross-variograms of the scores along the first two FPCs.

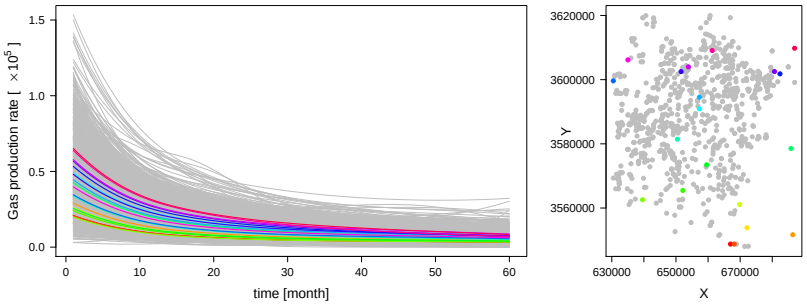

Figure 7: Prediction by UTrK of GPRCs for 20 random location at the Barnett shale. Left: smoothed data (grey lines) and predictions (colored lines). Right: sampled location (grey symbols) and target locations (colored symbol).

scale).

The predictions obtained via UTrK and UCok appears overall consistent. They both show increasing values of the gas production rate in direction S-W to N-E, justifying *a posteriori* the introduction of the drift term. From the application viewpoint, the results suggest the presence of a more productive region in the North-Eastern part of the study area. From field development perspective, this indicates "sweet spots" that would be most favourable for future reservoir development.

3.3 Monte Carlo Study

To assess the quality of predictions obtained through UTrK and through UCok, we performed a Monte Carlo study based on the same dataset analysed before. We randomly split the dataset in (a) a training set of $\kappa\%$ of data and (b) a test set of $(100-\kappa)\%$ of data, with $\kappa = 25, 50, 75$. For each training set, we applied

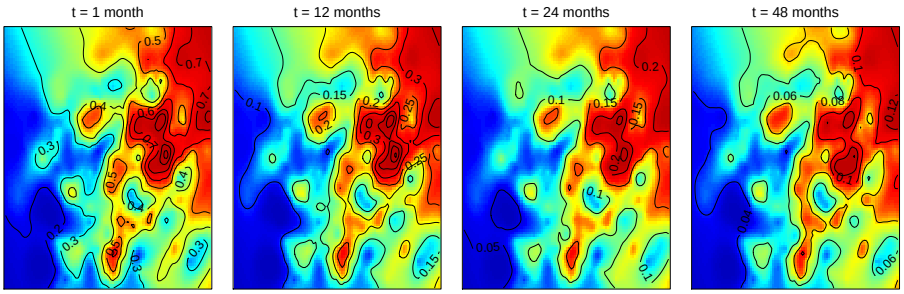

Figure 8: Prediction maps obtained with UTrK at the Barnett shale, for $t = 1, 12, 24, 48$ months. Colors are given on a non-uniform scale. Gas rate reported at contour lines are meant up to a factor 10^5.

the estimation procedures as illustrated in Subsection 3.2, and predicted the data in the test set by UTrK and UCok. To assess the quality of the predictions, we defined the sum of squared error (SSE) of UTrK for the i-th element $x_{\boldsymbol{s}_i}$ of the test set as:

$$SSE_i^{(UTrK)} = \|X_{\boldsymbol{s}_i}^{*tr} - x_{\boldsymbol{s}_i}\|^2.$$

Here, $X_{\boldsymbol{s}_i}^{*tr}$ denotes the predictor of the GPRC at $\boldsymbol{s}_i$, obtained by taking the exponential of the UTrK predictor from the log-GPRCs. Analogously, we defined the SSE related to UCok as

$$SSE_i^{(UCok)} = \|X_{\boldsymbol{s}_i}^{*} - x_{\boldsymbol{s}_i}\|^2,$$

with $X_{\boldsymbol{s}_i}^{*}$ the exponential of the UCok predictor at $\boldsymbol{s}_i$ from the log-GPRCs.

An overall index of prediction performance on a given test set can be then obtained as the mean or the median of $SSE_i^{(UTrK)}$, $SSE_i^{(UTrK)}$ over the elements of the test set. To appreciate the magnitude of the error with respect to the amplitude of the data, we normalized the SSEs by the average squared norm of the data in the training set, as suggested by Menafoglio et al. (2013). We refer to the normalized indices as relative SSEs (RSSEe). SSEs evaluated on log-GPRCs are in agreement with those in the original scale and are thus omitted from the description below.

To provide a Monte Carlo estimate of the RSSEs, we replicated the experiment over 100 randomly selected training/test sets, for each value of the parameter κ in $\{25, 50, 75\}$, i.e, we repeated the experiment for 100 random training sets with 25% (or 50% or 75%) of the data to predict the elements of the corresponding test sets composed by the remaining 75% (or 50% or 25% respectively) of the data.

Figure 9 reports the boxplots of the mean/median $RSSE^{(UTrK)}$ and $RSSE^{(UCok)}$ estimated via Monte Carlo simulation. To ease the comparison, Table 1 reports the mean, median and standard deviation of the estimated indices, assessed via Monte Carlo. Simulations show that for both UTrK and UCok the prediction quality increases as the number of data in the training set increases. This reflects the fact that κ is associated with the amount of information available to perform predictions. Moreover, for any given κ, UTrK and UCok performances are almost equivalent, with slightly better results for UTrK. This is confirmed from the graphical inspection of Figure 9.

		Median	Mean	Std. dev.
Mean $RSSE^{(Cok)}$	$\kappa = 25$	0.153	0.153	0.019
	$\kappa = 50$	0.136	0.137	0.012
	$\kappa = 75$	0.132	0.133	0.014
Median $RSSE^{(UCok)}$	$\kappa = 25$	0.066	0.066	0.006
	$\kappa = 50$	0.060	0.060	0.005
	$\kappa = 75$	0.056	0.057	0.007
Mean $RSSE^{(UTrK)}$	$\kappa = 25$	0.150	0.151	0.017
	$\kappa = 50$	0.134	0.136	0.011
	$\kappa = 75$	0.129	0.130	0.014
Median $RSSE^{(UTrK)}$	$\kappa = 25$	0.065	0.065	0.005
	$\kappa = 50$	0.059	0.059	0.005
	$\kappa = 75$	0.056	0.057	0.006

Table 1: Distribution of the RSSE indices for UCok and UTrK in the original scale, assessed via Monte Carlo simulation.

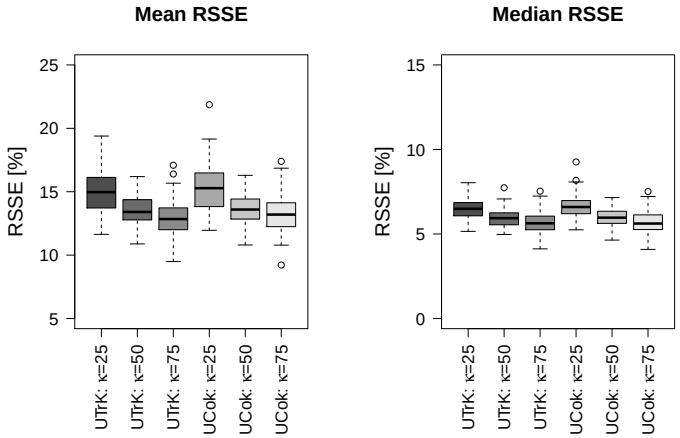

Figure 9: Boxplots of RSSE indices for UCok and UTrK.

		Median	Mean	Std. dev.
Mean $RSSE^{(UCok)}$	$\kappa = 25$	0.170	0.528	3.198
	$\kappa = 50$	0.148	0.163	0.080
	$\kappa = 75$	0.141	0.144	0.019
Median $RSSE^{(UCok)}$	$\kappa = 25$	0.070	0.070	0.007
	$\kappa = 50$	0.063	0.063	0.006
	$\kappa = 75$	0.060	0.060	0.009

Table 2: Distribution of the RSSE indices for UCok based of FSVD in the original scale, assessed via Monte Carlo simulation.

Even though UCok could potentially provide improved results with respect to UTrK – due to its generality – no gain seems to be obtained when increasing the problem complexity. We recognize at least two reasons for this: (i) the preprocessing step by FPCA and (ii) the fitting of a multivariate variogram model to the scores. In fact, the simplicity of the UTrK predictor is likely to be the key for its slightly better performance over the UCok predictor.

For sake of completeness, we report in Table 2 and Figure 10 the results obtained by projecting the log-GPRCs over the basis of the first two right functional singular vectors (FSVs), computed numerically. These are the equivalent, in the functional case, of the right singular vector obtained in the well-known singular value decomposition (SVD) of a multivariate dataset. In this context and unlike FPCA, the procedure detailed in Subsection 2.1 can be applied without modifications, e_k being represented by the k-th right FSVs.

We first notice that a significant number of upper outliers affects the results related to the mean $RSSE^{(UCok)}$, for $\kappa = 25$. This is probably caused by an amplification of the error due to the exponential transformation, as only one outlier is recorded when SSE is evaluated on a log-scale. Besides this, simulations show that UTrK outperforms UCok in all the tested scenarios under FSVD preprocessing. In fact, marked differences are recorded between the performances of UCok under FPCA and of UCok under FSVD preprocessing. This is likely to be due to the fact that UCok under FPCA preprocessing is applied to centered observations, i.e., those obtained by subtracting the sample mean of the dataset from the observations (see the Appendix). In this sense, the entire information within the sample mean is kept in UCok prediction, as the latter is added to the UCok prediction obtained from centered data. In contrast, FSVD is applied to non-centered observations: here the dimensionality of the non-centered observations is reduced. Intuitively, given K, FPCA exploits one dimension (that of the sample mean) more than FSVD, at the same expense.

These results underline the fact that, whenever a dimensionality reduction is performed prior to apply UCok, the prediction results may be strongly influenced by the preprocessing step and the chosen dimension K. Such problem is overcome when using UTrK, as no preprocessing is required.

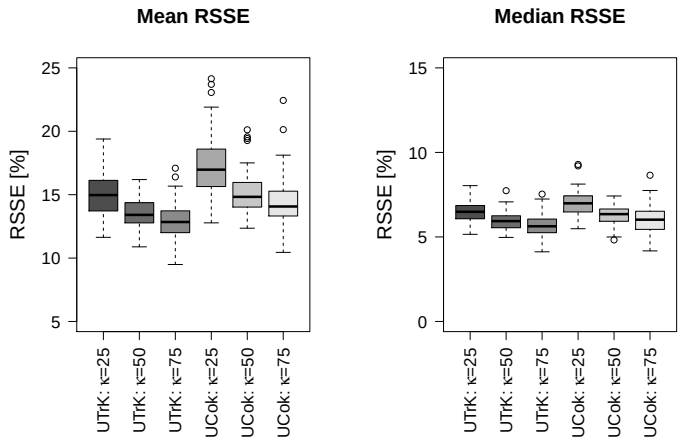

Figure 10: Boxplots of RSSE indices for UCok based on FSVD and UTrK.

4 Conclusions

In this work, we considered two approaches to the spatial prediction of gas production rate curves (GPRCs) in unconventional reservoirs: (1) Universal Cokriging (UCok) and (2) Universal Trace-Kriging (UTrK). We analysed the strengths and weaknesses of these methodologies both theoretically and from the application viewpoint, through an extensive Monte Carlo study on field data.

Our study shows that the investigated approaches lead to consistent results on available data. Nevertheless, UTrK proved preferable in most scenarios tested though Monte Carlo simulation. Here we showed that the dimensionality reduction operated on the data prior to the geostatistical analysis with UCok approach can be highly influential on the quality of the results. In this context, the simplicity of UTrK allows to avoid such preprocessing, and seems to be the key of its better performances over UCok.

Anyway, we note that the theoretical study of UCok deserves attention from the methodological viewpoint. Indeed, UCok appears better suited than UTrK to extensions related to the form of the drift term. Indeed, one might want to consider a more complex functional linear model than that in Eq. (2). This would allow to incorporate geological or production variables, possibly time-varying (i.e., functional regressors), in the geostatistical analysis.

Appendix: Functional Principal Component Analysis

We here consider the Functional Principal Component Analysis (FPCA Ramsay and Silverman, 2005, Chapter 8) as a dimensionality reduction method for GPRCs, possibly log-transformed. FPCA is a methodology aiming to identify a reduced space to optimally represent a set of observations. Given a target dimension K, FPCA determines the system of K orthonormal directions

$\{e_k, 1 \leq k \leq K\}$ that best represents the variability of the data set around its mean.

As in the multivariate setting, the k-th Functional Principal Component (FPC) is the eigenfunction associated with the k-th largest eigenvalue of the (zero-lag) covariance operator $C(\mathbf{0})$, that is $C(\mathbf{0})e_k = \rho_k e_k$, $\rho_1 > ... > \rho_k$. Note that, in the assumptions of Section 2, all the data are featured by the same covariance operator $C(\mathbf{0})$.

The proportion of the variability explained by the first K Functional Principal Components (FPCs) can be measured through the ratio between the partial and the total sum of the eigenvalues of $C(\mathbf{0})$: $\sum_{k=1}^{K} \rho_k / \sum_{k=1}^{\infty} \rho_k$. To perform dimensionality reduction, one can then consider the projection of the data along the first K FPCs, $\{e_k, 1 \leq k \leq K\}$, where K is set as to explain a given amount of the total variability (e.g., 90% or 95%).

If spatial covariance C is unknown, one can introduce the empirical FPCA, that is based on the eigen-decomposition of the empirical zero-lag covariance operator, defined for $x \in L^2(T)$ as

$$\widehat{C}(\mathbf{0})\, x = \frac{1}{n} \sum_{i=1}^{n} \langle X_{\boldsymbol{s}_i} - \overline{X}, x \rangle (X_{\boldsymbol{s}_i} - \overline{X}),$$

where $\overline{X} = \sum_{i=1}^{n} X_{\boldsymbol{s}_i}$ is the sample mean (see, e.g., Horváth and Kokoszka, 2012, Chapter 2.17). The coefficients for the basis representation are then obtained as $\widetilde{\xi}_k(\boldsymbol{s}_i) = \langle X_{\boldsymbol{s}_i} - \overline{X}, e_k \rangle$, $i = 1, ..., n$, $k = 1, ..., K$. Notice that, if the mean were spatially constant, $\overline{X}$ would estimate the mean of the process, and $\widetilde{\xi}_k(\boldsymbol{s}_i)$, $i = 1, ..., n$ would be (approximately) zero-mean. In the non-stationary assumptions of Section 2, $\widetilde{\xi}_k(\boldsymbol{s}_i)$ is not zero-mean, but approximately follows a model of the form (5), as we show below.

We call $\widetilde{X}_{\boldsymbol{s}_i}$, $i = 1, ..., n$, the modified dataset obtained by centering the $X_{\boldsymbol{s}_i}$ with respect to the sample mean of the dataset, i.e., $\widetilde{X}_{\boldsymbol{s}_i} = X_{\boldsymbol{s}_i} - \overline{X}$. Under models (2) and (3), for the modified process $\widetilde{X}_{\boldsymbol{s}_i}$ one has

$$\widetilde{X}_{\boldsymbol{s}} = \sum_{k=1}^{K} \sum_{l=0}^{L} a_{kl} \left(f_l(\boldsymbol{s}) - \frac{1}{n} \sum_{i=1}^{n} f_l(\boldsymbol{s}_i) \right) + \sum_{k=1}^{K} \epsilon_k(\boldsymbol{s}) e_k - \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \epsilon_k(\boldsymbol{s}_i) e_k. \quad (15)$$

For a sample size sufficiently large with respect to K and a moderate spatial dependence (see Horváth and Kokoszka, 2012, Chapter 18), the last term becomes negligible, since $\{\epsilon_k(\boldsymbol{s})\}$ is zero-mean. By noting that, given the regressors, $\sum_{i=1}^{n} f_l(\boldsymbol{s}_i)$ is a known constant, the following approximate model is obtained from expression (15)

$$\widetilde{X}_{\boldsymbol{s}} \approx \sum_{k=1}^{K} \sum_{l=0}^{L} a_{kl} \widetilde{f}_l(\boldsymbol{s}) + \sum_{k=1}^{K} \epsilon_k(\boldsymbol{s}) e_k. \quad (16)$$

The latter term has the same form as Eq. (6), but with modified regressors. Notice in particular that model (16) is without intercept.

Therefore, when the dimensionality reduction is performed via the empirical FPCA, one can proceed as follows: (i) project the $\widetilde{X}_{\boldsymbol{s}_i}$, $i = 1, ..., n$ over the first K eigenfunctions and compute the scores $\widetilde{\xi}_k(\boldsymbol{s}_i)$, $i = 1, ..., n$, $k = 1, ..., K$; (ii) perform the geostatistical analysis/prediction of $\widetilde{\xi}_k(\boldsymbol{s}_i)$ and obtain the UCok

prediction $\widetilde{X}^*_{\boldsymbol{s}_0}$ at the target location $\boldsymbol{s}_0$ as described in Subsection 2.1; (iii) compute the final prediction by adding to $\widetilde{X}^*_{\boldsymbol{s}_0}$ the sample mean $\overline{X}$: $X^*_{\boldsymbol{s}_0} = \widetilde{X}^*_{\boldsymbol{s}_0} + \overline{X}$.

Acknowledgements Authors would like to thank âĂIJdrillinginfo.comâĂİ for generously providing academic license/access to their website/database. Grujic and Caers would also like to acknowledge Stanford Center for Reservoir Forecasting (SCRF) consortium (2013-2015) for financially supporting this research.

References

Besse, P. C., H. Cardot, R. Faivre, and M. Goulard (2005). Statistical modelling of functional data. *Applied Stochastic Models in Business and Industry 21*(2), 165–173.

Besse, P. C., H. Cardot, and D. B. Stephenson (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics 27*(4), 673–687.

Caballero, W., R. Giraldo, and J. Mateu (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment 27*, 1553–1563.

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

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

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

Giraldo, R. (2009). *Geostatistical Analysis of Functional Data*. Ph. D. thesis, Universitat Politècnica da Catalunya, Barcellona.

Grujic, O., C. Da Silva, and J. Caers (2015). Functional approach to data mining, forecasting, and uncertainty quantification in unconventional reservoirs. In *SPE 174849, SPE Annual Technical Conference & Exhibition*.

Henderson, B. (2006). Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics 17*(1), 65–80.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer.

Ignaccolo, R., J. Mateu, and R. Giraldo (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment 28*(5), 1171–1186.

Josset, L., D. Ginsbourger, and I. Lunati (2015). Functional error modeling for uncertainty quantification in hydrogeology. *Water Resources Research 51*(2), 1050–1068.

Kormaksson, M., M. R. Vieira, and B. Zadrozny (2015). A data driven method for sweet spot identification in shale plays using well log data. In *SPE Digital Energy Conference and Exhibition. Society of Petroleum Engineers*.

Manté, C. and G. Stora (2012). Functional PCA of measures for investigating the influence of bioturbation on sediment structure. In *20th International Conference on Computational Statistics*, pp. 531–542.

Matheron, G. (1969). Le krigeage universel. vol. 1. Technical report, Fontainebleau: Cahiers du Centre de Morphologie Mathématique, École des Mines de Paris.

Menafoglio, A., A. Guadagnini, and P. Secchi (2014). A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment 28*(7), 1835–1851.

Menafoglio, A. and G. Petris (2015). Kriging for Hilbert-space valued random fields: The operatorial point of view. *Journal of Multivariate Analysis*. . DOI:10.1016/j.jmva.2015.06.012.

Menafoglio, A., P. Secchi, and M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics 7*, 2209–2240.

Mohaghegh, S. (2011). Modeling, history matching, forecasting and analysis of shale reservoirs performance using artificial intelligence. In *SPE 143875, SPE Digital Energy Conference*.

Mohaghegh, S. (2013). A critical view of current state of reservoir modeling in shale assets. In *SPE 165713, SPE Eastern Regional Meeting*.

Mănescu, C. B. and G. Nuño (2015). Quantitative effects of the shale oil revolution. *Energy Policy*. . DOI:10.1016/j.enpol.2015.05.015.

Nerini, D., P. Monestiez, and C. Manté (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis 101*(2), 409–418.

Patzek, T. W., F. Male, and M. Marder (2013, December). Gas production in the barnett shale obeys a simple scaling theory. *PNAS 110*(49), 19731–âĂŞ19736.

Ramsay, J. and B. Silverman (2005). *Functional data analysis* (Second ed.). Springer, New York.

Sancho, J., C. Iglesias, J. Piñeiro, J. Martínez, J. Pastor, M. Araújo, and J. Taboada (2015). Study of water quality in a spanish river based on statistical process control and functional data analysis. *Mathematical Geosciences*, 1–24. . DOI:10.1007/s11004-015-9605-y.

Satija, A. and J. Caers (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources 77*, 69 – 81.

Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology 13*(43), 1–12.

Vidic, R. D., S. L. Brantley, J. M. Vandenbossche, D. Yoxtheimer, and J. D. Abad (2013). Impact of shale gas development on regional water quality. *Science 340*(6134).

Yan, F., L. Liu, Y. Li, Y. Zhang, M. Chen, and X. Xing (2015). A dynamic water quality index model based on functional data analysis. *Ecological Indicators 57*, 249 – 258.

Yang, W., H. G. Müller, and U. Stadtmüller (2011). Functional singular component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 303–324.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

49/2015 Ghiglietti, A.; Ieva, F.; Paganoni, A.M.
Statistical inference for stochastic processes: two sample hypothesis tests

50/2015 Grillo, A.; Guaily, A.; Giverso, C.; Federico, S.
Non-Linear Model for Compression Tests on Articular Cartilage

51/2015 Ballarin, F.; Faggiano, E.; Ippolito, S.; Manzoni, A.; Quarteroni, A.; Rozza, G.; Scrofani, R.
Fast simulations of patient-specific haemodynamics of coronary artery bypass grafts based on a POD–Galerkin method and a vascular shape parametrization

52/2015 Giverso, C.; Scianna, M.; Grillo, A.
Growing Avascular Tumours as Elasto-Plastic Bodies by the Theory of Evolving Natural Configurations

47/2015 Colombo, M. C.; Giverso, C.; Faggiano, E.; Boffano,C.; Acerbi, F.; Ciarletta, P.
Towards the personalized treatment of glioblastoma: integrating patient-specific clinical data in a continuous mechanical model

48/2015 Ambrosi, D.; Pettinati, V.; Ciarletta, P.
Active stress as a local regulator of global size in morphogenesis

46/2015 Giverso, C.; Verani, M.; Ciarletta P.
Emerging morphologies in round bacterial colonies: comparing volumetric versus chemotactic expansion

45/2015 Lange, M.; Palamara, S.; Lassila, T.; Vergara, C.; Quarteroni, A.; Frangi, A.F.
Improved hybrid/GPU algorithm for solving cardiac electrophysiology problems on Purkinje networks

44/2015 Antonietti, P.F.; Houston, P.; Smears, I.
A note on optimal spectral bounds for nonoverlapping domain decomposition preconditioners for hp-Version Discontinuous Galerkin methods

43/2015 Deparis, S.; Forti, D.; Gervasio, P.; Quarteroni, A.
INTERNODES: an accurate interpolation-based method for coupling the Galerkin solutions of PDEs on subdomains featuring non-conforming interfaces