SUBSURFACE MODELING WITH FUNCTIONAL DATA

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ENERGY RESOURCES ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ognjen Grujić
November 2017

This dissertation is online at: http://purl.stanford.edu/tg998gb5075

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jef Caers, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tapan Mukerji

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Celine Scheidt

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

Unconventional shale reservoirs are currently one of the most developed energy resources in the world. In the US alone, more than 1.1 million shale wells are currently in production and the estimated US shale oil reserves are at around 4.29 trillion barrels. The development of shale reservoirs is very rapid in the US. Some companies drill more than 400 horizontal shale wells per year, and this trend is likely to increase in the coming years. In such setting, quick uncertainty quantification and forecasting is of paramount importance. Conventional approaches to uncertainty quantification and forecasting were mostly found impractical in shales due poorly understood production mechanisms, high temporal requirements of uncertainty quantification studies and also because of the poor collection of scientific data. Shale reservoirs are mostly developed by small to mid size companies that usually cannot afford the collection of highly sophisticated scientific data or advanced geomodeling and simulation softwares. Moreover, transport in shales and upscaling of shale reservoir properties are very active research areas without well established practical work-flows and software. For these reasons, many reservoir modelers have adopted the so called data-driven approaches for reservoir data analysis and production forecasting. Commonly employed data-driven methodologies include regression methods for the analysis of production data and neural network based models for production forecasting. What most of the approaches employed up to date have in common is that they work with scalar outputs (i.e. 3, 6, 9 months of cumulative production or peak oil/gas), spatial correlations and spatial trends are rarely analyzed, and many of the commonly employed approaches are incapable of properly quantifying uncertainty in forecasts. In this dissertation, we take a different approach to data-driven reservoir data analysis and forecasting. Firstly, we start from a perspective that shale production profiles need to be analyzed as a whole, or as curves, and not as some discretized components of

cumulative production. Secondly, we embrace the existence of spatial correlations between production profiles. It is quite intuitive that similarly completed wells would produce similar amounts of hydrocarbons if drilled reasonably close (without interference). The question is at what distance does this "similarity" disappear and how does that affect forecasts, recoverable reserve estimation and ultimate decision making? Here, we develop methodologies for interpretation, forecasting and uncertainty quantification of spatially correlated reservoir production curves or functions. The methodologies are based on the tools of relatively recently established statistical discipline called functional data analysis (FDA) and the advances in its sub-discipline, geostatistics for functional data. We demonstrate that the well known geostatististical tools such as variograms and sequential Gaussian simulation can be efficiently used for shale reservoir data analysis and forecasting. The developed methodologies are demonstrated in two unconventional reservoir case studies. The first case study analyzed gas production from 900 horizontal wells completed in the Barnett shale, while the second case study used Anadarko Petroleum Company (APC) provided dataset with 189 wells that produced oil and gas.

The previously mentioned methodologies are capable of analyzing and forecasting single variate functional data (i.e. oil rates curves only). However, the APC dataset contained wells that produced multivariate functional data, oil and gas rates over time. When analyzing such dataset the question of multivariate functional data analysis and forecasting naturally arises. The encounter of this question motivated the development of a methodology capable of analyzing and forecasting multivariate functional data. The developed methodology is based on regression trees, a well known machine learning technique, and it represents a contribution to both fields of Earth sciences and functional data analysis. The methodology is also demonstrated on the APC dataset.

Functional data is not only observed in unconventional reservoir data analysis and forecasting. One also encounters functional data in conventional reservoir modeling. For example, flow simulation curves computed with conventional reservoir simulators also represent functional data. Proper numerical uncertainty quantification requires consideration of a large number of modeling parameters with wide ranges. Exhaustive exploration of such high dimensional spaces is computationally demanding and rarely achievable in practice. For this reason, modelers often employ statistical emulators

that aim to interpolate or emulate the reservoir simulation solution at unexplored portions of the input space. Statistical emulators require a certain number of training runs computed with high fidelity reservoir simulators and in most applications up to date work with scalar outputs (i.e. EUR). Since reservoir simulator outputs are functional in nature, one can develop statistical emulators with the aforementioned methodologies we developed for shale reservoir forecasting. We explore this application in the last, sixth chapter of this dissertation.

Another problem in conventional uncertainty quantification studies is with the use of proxy models. Proxy models are numerical models of lower fidelity and high speed that are commonly used to quickly explore the high dimensional input spaces. Their stand-alone solution is often considered noisy and sub-optimal. For this reason, modelers often employ machine learning to model the discrepancies (errors) between the proxies and their high fidelity counterparts, or they employ co-kriging based schemes that fit a statistical emulator that aggregates both proxy and high fidelity solutions in estimating unevaluated high fidelity solutions. In chapter 6, we develop novel functional co-kriging methodologies for building statistical emulators that aggregate functional responses produced by proxies and high fidelity numerical models. The methodologies are applied and compared in three case studies along with the emulators constructed with the techniques we previously used in shale reservoir modeling.

Acknowledgements

Looking back at the past six years of my life, I can confidently say that obtaining a PhD was the most difficult thing I have ever done. I can also confidently say that it was also the most rewarding experience in my life. I have learned a lot, grown a lot, and had an extraordinary opportunity to meet and work with some of the most brilliant people in the industry. While I wrote this dissertation all by myself, many have helped me in my work and in getting to where I am today. Here, I would like to express my most sincere gratitude towards all of them.

I would like to thank my adviser professor Jef Caers for numerous meetings and discussions. I am very grateful for everything I have learned from Jef and for his support during some of the most difficult moments of this journey. I would also like to thank professor Tapan Mukerji who was my adviser during the first year of my PhD and who served on my reading committee. Tapan is one of the kindest people I know and I am very grateful for everything I have learned from him. I am thankful for all his great comments on my work throughout the years, during my defense and on this dissertation. I am thankful to professors Erik Dunham and Roland Horne for serving on my committee and for giving valuable suggestions and comments on my dissertation. Many thanks to Celine Scheidt for serving on my reading committee, for all great comments on an earlier version of this dissertation, for always being friendly and positive, for all her help with many aspects of my research, and on top of everything for being a great friend! Since 2015 I have been having a wonderful collaboration with Alessandra Menafoglio of Politecnico di Milano. Alessandra brought mathematical rigor into my life and thought me the nuts and bolts of geostatistics for functional data. I am truly blessed to have Alessandra as a friend and as a colleague! This dissertation and the ideas therein would not have been possible without the real dataset provided by Anadarko Petroleum Corporation. In

that regard, I am grateful to Carla da Silva and P.K. Pande for many discussions about the dataset and for the great collaboration we had throughout the years.

My research has benefited a lot from my industry experiences. I did three internships with Chevrons Energy Technology Company (ETC), and one internship jointly with Ar2Tech and Streamsim. During these internships, I had a chance to work with some of the greatest minds of our industry. I learned a lot from Herve Gross, Brad Mallison, Andrew Oghena, Robert Fitzmorris, Alex Boucher, and Marko Thiele. I am forever thankful to all of them for being great colleagues and even better mentors, and for everything they had thought me. My dissertation would not have been nearly this good if it weren't for the great opportunities of working with them!

One of the most important things I have learned during my PhD journey was how to fail and be totally OK with it. This was probably the most painful element of this learning experience. Thankfully, I made many friends who helped me endure some of the most difficult moments of this journey. I am forever grateful to Priscilla Ribeiro for being a dear friend and being there for me when I was about to present my research proposal. At the time, it seemed like all my battles were lost. Priscila was there to encourage me to keep going, push through, and guess what? I made it :). I am equally grateful to my dear friend Karine Levonyan for many great discussions, numerous coffee breaks and endless support throughout the years. Having Priscila and Karine as friends is a true blessing!

Many thanks to my officemates Francois Hamon, Erik Nesvold and Matthieu Rousset. Sharing an office with them was a great experience. I thank them for many great research discussions and for exceptionally great friendship.

I enjoyed a great friendship, collaboration and discussions with my ERE friends. In particular, I would like to thank: Orhun Aydin, Addy Satija, Nicola Castelletto, Julio Hoffiman, Andre Jung, Ahinoam Pollack, Lijing Wang, and Christin Strandli.

I had a wonderful opportunity to study with: Rohisha Adke, Boxiao Li, Jiaoming Ouyang, Elnur Alyev, and Crystal Shi. I thank them all for being great friends and for many study hours we spent together.

I am very thankful to the ERE staff. Particularly: Joanna Sun, Sandy Costa, Eiko Rutherford, Thuy Nguyen, and Rachael Madison for always being extremely helpful and friendly and always there to answer all my questions and assist me with whatever I needed.

I am wholeheartedly thankful to my mom Nada, my grandma Savka, my brother Filip, and the Zorić family for always supporting me and being there for me.

Many thanks to Mrs Doina Jikich and Mr Sinisha Jikich for encouraging me to go to grad school in the US. Without that initial encouragement, I don't think I would have ever made it to Stanford!

I greatly appreciate the support of my friends: Eduard Tuchfeld, Jovana Ćirković, Jill Stoffers, Peter Jesse, Michael and Borja Dorsch, Sandra Suarez and Jose Palomares.

Last but certainly not least, I am thankful beyond what words can express to my wife Karla Vargas who lived through every good and bad moment of this journey with me and stood by me no matter what. I am truly blessed for having Karla in my life and very grateful for all her love and support.

November 2017

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Unconventional reservoirs, or organically rich shales, are nowadays one of the most widely developed energy resources in the world. What was once considered a vast but untappable resource (due to very low permeability), today it is developed at such a rapid pace that some developers drill more than 400 wells per year. This rapid development is a result of years of scientific and industry research that started in 1976 with the Eastern Shales Gas Project (ESGP) that was initiated and funded by the US Department of Energy (DOE). The project lasted until 1992 and it consisted of many real reservoir experiments that identified horizontal drilling with hydraulic fracturing as one of the most effective engineering techniques to unlock the great potential of organically rich shales. While the project initially started with an objective to unlock the potential of Devonian gas shales in the Appalachian basin, it eventually motivated the development of other oil and gas rich shales throughout the US (figure 1.1), and around the world (China, Argentina, etc.). A conservative estimate of the worlds shale oil resources is around 6 trillion barrels (World Energy Council [2016]). Around 80% of the worlds shale reserves are in teh US (World Energy Council [2016]). The Green River basin alone, is estimated to contain around 3.7 trillion barrels, while Devonian shales in the Appalachian basin are estimated to contain around 189 billion barrels of oil (World Energy Council [2016]).

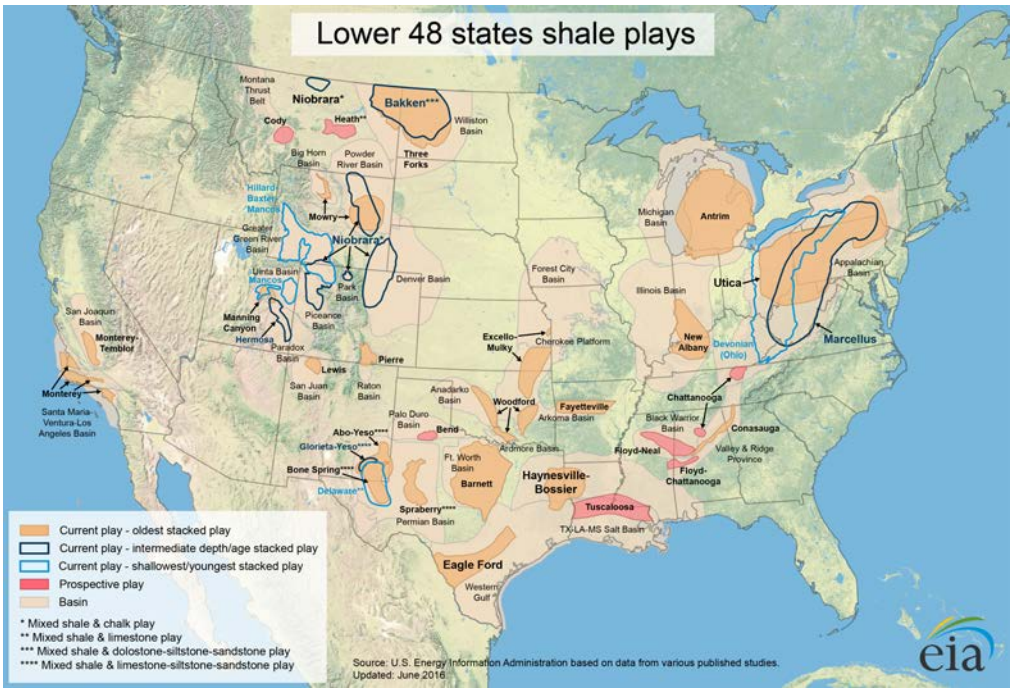


Figure 1.1: *North American shale plays* [1]

1.1 Geological Properties of Shales

Organically rich shales are finely grained sedimentary rocks that were deposited in calm, oxygen poor oceanic environments (Roen et al. [1996]). Low oxygen content enabled the preservation of trapped organic matter until subsequent burial due to sediment influx. Historically, in petroleum geology, organically rich shales were often referred to as source rocks. This was mainly due to the fact that when exposed to high temperatures the organic matter contained in organically rich shales converts into hydrocarbons and migrates into shallower sandstones where it forms conventional hydrocarbon reservoirs. Hydrocarbons contained in shales were generated and trapped in situ, in other words no migration ever happened. For this reason shales are often refereed to as self-sourcing reservoirs. Mineralogical composition of shales varies both between and within shale plays. In table 1.1 we are showing mineralogical composition of core samples extracted from four shale plays in the US. Within shale

[1] Source EIA (https://www.eia.gov/oil_gas/rpd/shale_gas.pdf)

and between shale variations in mineral content are quite significant. Permeability of shales is on the scale of nano Darcys, they are almost entirely impermeable rocks. Hydrocarbons in shales can be free or chemically bounded (sorbed gas) to the rock. Some shales have very well developed networks of natural fractures (i.e. parts of Marcellus shale) while others have very few natural fractures. This high heterogeneity of shales imposes unique challenges on geologists and geomodelers since best practices established in one shale play are rarely transferable to other shale plays.

Table 1.1: *Core sample properties of four shale plays (Modified from Sone and Zoback [2013])*

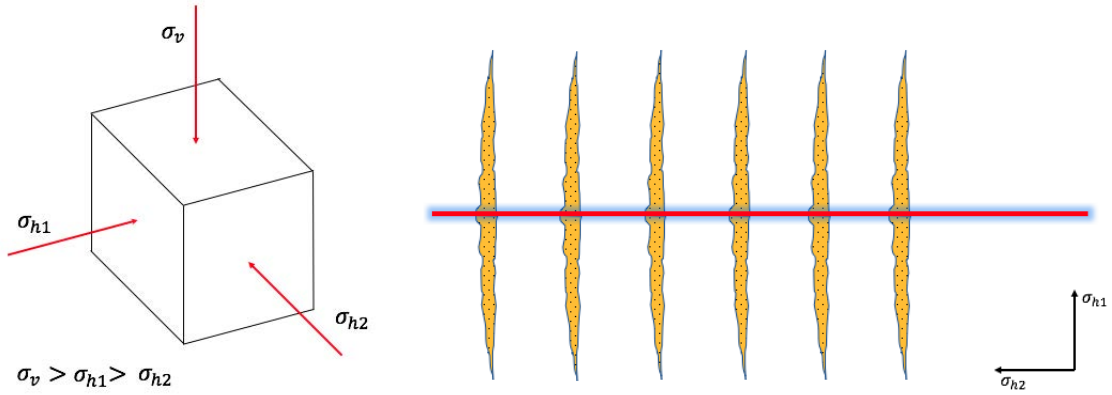
Shale Play	QFP Vol.%	Carbonate Vol.%	Clay Vol.%	Kerogen Vol.%	Porosity %	Description
Barnett-1	50-52	0-3	36-39	9-11	4-9	Silt grains in clayey matrix
Barnett-2	31-53	37-60	3-7	2-3	1-2	Carbonate rock
Haynesville-1	32-35	20-22	36-39	8	6	Silt grains in clayey matrix
Haynesville-2	23-24	49-53	20-22	4	3-4	Silt grains in clayey/calcerous matrix
Eagle-Ford-1	22-29	46-54	12-21	9-11	5-7	Calcerous biotic grains in clayey/calcerous matrix
Eagle-Ford-2	11-18	63-78	6-14	4-5	5-7	Calcerous biotic grains in clayey calcerous matrix
Fort. St. John	54-60	3-5	32-39	4-5	5-6	Silt grains in clayey matrix

QFP = quartz, feldspar, pyrite

1.2 Hydraulic Fracturing

Given their very low permeability, in shales, production with vertical wells is rarely economical. In some Devonian shales in the Appalachian basin economical production was established with vertical wells due to highly developed networks of natural fractures that provided high permeability flow paths (i.e. Lower Huron Shale). In other basins dense fracture networks are not very common and in such cases economical production can be established only by means of horizontal drilling with multi-stage hydraulic fracturing. Hydraulic fracturing is an operation in which a mixture of sand, water, gelling agents and chemicals is pumped in sequence into a well under high pressure until the surrounding rock breaks, creating an artificial fracture. Hydraulic

fracture orientation depends on the orientation of the maximum principal stresses in the subsurface. Hydraulic and natural fractures tend to propagate perpendicularly to the minimum principal stress (Zoback [2007]). At large depths, minimum principal stress is one of the two horizontal principal stresses hence hydraulic fractures are always vertical. To maximize the contact of hydraulic fractures with the reservoir, horizontal wells are drilled in the direction of minimum principal stress (so that hydraulic fractures propagate perpendicular to the horizontal well, 1.2).



(a) (b)

Figure 1.2: *a) - Principal stresses acting on unit volume in the subsurface. b) - Top cross section of a horizontal well with hydraulic fractures drilled along the direction of the minimum principal stress.*

The principal stresses are measured in early stages of drilling by means of micro frac testing (Proskin et al. [1990]) and they do not represent an uncertain parameter in reservoir development. However, mechanical properties of shales can vary throughout the reservoir due to spatially varying mineralogical composition of the rock. It was found (Altamar and Marfurt [2014], Wang and Gale [2009], Ren et al. [2014]) that brittleness of shales depends on the mineralogical content. In particular, high clay content makes the rock ductile while large content of quartz, feldspar or dolomite makes the rock brittle. Ren et al. [2014] states that hydraulic fracturing in ductile shales produces planar hydro fractures, while fracing in brittle shales produces dense networks of artificial fractures. Spatial distribution of mechanical reservoir properties is uncertain and it determines the success of a hydraulic fracturing job and the magnitude of subsequent production. It is important to mention that in some shales

hydraulic fracturing with water based mixture is detrimental to hydrocarbon production due to swelling of clay minerals (Karpiński and Szkodo [2015]). In such shales, foam based fracing mixtures are commonly used, however in early stages of reservoir development a significant amount of experimentation always takes place.

1.3 The Need for Data Driven Modeling

While the technology to develop unconventional reservoirs is relatively new, the questions that reservoir engineers need to answer are the same as in conventional reservoirs. Where to drill the next well? And what production to expect? Conventionally these questions are answered through some sort of forecasting and uncertainty quantification studies. One of the most widely adopted uncertainty quantification approaches in conventional reservoir engineering is reservoir modeling with flow simulation (RMFS, Caers [2011]). RMFS approach starts by developing complex earth models that incorporate all available reservoir data (well logs, seismics, well test, pvt, production) and then uses such earth models to conduct flow simulation studies and forecast new wells. This approach to reservoir characterization and forecasting is impractical for unconventional reservoirs due to the following:

1. Flow mechanisms of shales are not yet well understood hence there is no consensus on the best and the most appropriate flow and transport modeling technique. Early approaches to flow simulation in shales considered the well-known dual porosity models. However, recent research has shown that dual porosity models are inappropriate for unconventional reservoir modeling since transport occurs across multiple scales (Yan et al. [2016]). It was also found that Darcys law becomes inapplicable in the case when pore sizes are at the level of nano-meters (Guo et al. [2015]) that is very common in shales. Several reservoir simulation techniques have been recently developed to deal with these problems. Particularly interesting are multiscale methods that simulate transport in shale reservoirs through organic and inorganic components of the matrix (Yan et al. [2016]). What all of the recently developed methods have in common is the fact that they are highly complex and that they require a significantly more sophisticated static reservoir modeling compared to conventional reservoirs[2].

[2]There are many more parameters to consider compared to conventional reservoir simulation

Moreover, upscaling of reservoir properties and flow simulation of micro effects to a larger scale is an active area of research, hence currently available methods are not yet ready for a wide practical use.

2. RMFS is highly sophisticated and as such it requires large teams of geologists, geo-modelers, geo-mechanics experts, flow simulation engineers, and decision analysts. Such large teams are rarely available to small to mid-sized companies that are the main developers of unconventional resources in the US.

3. RMFS uncertainty quantification technique falls short in meeting the demands of rapid reservoir development since one solid RMFS study can take anywhere from a couple of months to a couple of years. This obviously becomes highly impractical when 400 wells are being drilled per year.

As an alternative to RMFS modelers have turned towards the so-called data driven approaches. Data driven approaches rely on some sort of machine or statistical learning technique for interpretation and forecasting of reservoir production. Statistical methods are often fast to develop and capable of incorporating many different types of information. Commonly used data driven approaches in unconventional reservoir data interpretation include regression analyses (LaFollette and Holcomb [2011]), neural network based sensitivity analyses (Shelley et al. [2012], Nejad et al. [2015]), and production clustering analyses (Esmaili and Mohaghegh [2013]).
Neural networks appear to be one of the most popular forecasting approaches. Applications to date used neural networks to develop single well, single time, forecasting models without explicitly taking into consideration spatial correlations between the wells (Mohaghegh et al. [2011], Cao et al. [2016], Grujic and Mohaghegh [2010]). Other approaches to reservoir forecasting rely on some form of decline curve analysis (Arps [1945], Duong [2010]) applied on existing wells and then aim to use regression or neural networks to forecast new wells by forecasting its decline curve parameters.

studies

1.4 Motivation and Key Contributions

One thing that is commonly overlooked in nowadays performed unconventional production data analyses and forecasting studies is the fact that from a statistical perspective production data represents functions or curves. The shape and the magnitude of these production curves depends on the location of the well, its completion parameters (i.e. the number of fractures, lateral length, etc.), and its operating conditions. Proper analysis of production curves requires looking at the entire production curve as a whole and not some segments of it, like peak production or 3, 6, 9 months of cumulative production, as it is commonly done in the industry. Another very important, but often overlooked aspect in production data analysis and forecasting studies is the presence and the magnitude of spatial correlations between wells. Knowing the extent of spatial correlations (i.e. variogram ranges) can inform us about the main trends in the reservoir quality and also improve the results of production forecasting and uncertainty quantification of new wells.

Statistical discipline that develops techniques for the analysis of functional data is called functional data analysis (FDA, Ramsay and Silverman [2005]). A special branch of this discipline deals with the analysis and forecasting of spatially correlated functional data such as unconventional reservoir production curves. In this dissertation, we explore the applicability of geostatistics for functional data for the analysis and forecasting of unconventional wells. We develop work-flows and methodologies to analyze, forecast and quantify uncertainty in production curves in unconventional reservoirs. The developed methodologies are demonstrated on two real unconventional reservoir datasets: the Barnett shale dataset with 934 horizontal gas wells and a dataset provided to us by Anadarko Petroleum Corporation (APC) with 188 horizontal oil and gas wells.

Most nowadays developed unconventional reservoirs produce more than one fluid (i.e. oil and gas). This means that engineers need to analyze multivariate curves, that we know are spatially correlated. Motivated by this problem we developed a novel methodology for the analysis and forecasting of multivariate production curves based on regression trees, a statistical learning technique. While our developments aimed at unconventional reservoir data analysis and uncertainty quantification, the developed methodology represents a contribution to the field of functional data analysis and as

such it is applicable to problems in other branches of science.

Production curves also occur in numerical uncertainty quantification studies in conventional reservoirs. Proper uncertainty quantification requires exploration of high dimensional input spaces that is always computationally expensive. To save computational time engineers often develop statistical or machine learning surrogate models or proxies. Up to date proxies were developed for scalar outputs (i.e. EUR) and not the entire simulated production curves. The methods developed for the analysis and forecasting of unconventional production curves are also applicable for construction of statistical emulators for computer experiments that produce curves as outputs. In this dissertation, we also explore this application, and in addition, we develop a novel method for constructing emulators for computer codes of multiple levels of fidelity[3] that all produce time series as outputs. This development is also novel and it represents a contribution to both Earth science and functional geostatistics fields.

1.5 Dissertation Outline

In chapter 2, we outline the basics of functional data analysis mostly based on the book Functional Data Analysis by Ramsay and Silverman [2005]. Topics covered include: functional data smoothing from raw observations to functions, functional principal component analysis, and regression for functional data. In addition to this literature review, we outline a practical technique for smoothing of curves that have different length, such as unconventional produciton curves[4].

In chapter 3, we review the techniques of geostatistics for functional data. We start by reviewing the recently developed universal trace kriging methodology by Menafoglio et al. [2013], co-kriging for functional data by Nerini et al. [2010] and we also propose an extension to Nerini et al. [2010] method for non-stationary spatially correlated functional data. The chapter concludes with a real Barnett shale unconventional reservoir case study with 834 multi-stage fractured wells of similar horizontal length. Work presented in chapter 3 is a result of collaboration with Alessandra Menafolgio of Politecnico di Milano and it was recently published (Menafoglio et al.

[3]For example, finely gridded reservoir models and coarsely gridded reservoir models. Another example are models with reduced physics.

[4]Since wells start on different dates, production curves often have different length

[2016b]) in the journal of Spatial Statistics.

In chapter 4, we propose a methodology based on the techniques presented in chapter 3 to model a real reservoir data with a variable number of completion parameters. We demonstrate that the techniques from chapter 3 can be used without much change to forecast and interpret unconventional reservoir production data when wells have different completion parameters. The methodologies were applied on a real unconventional reservoir case study with 188 horizontal wells that produced oil and gas and had 26 well specific parameters (covariates). The case study considered oil production curves only. Parts of the work presented in chapter 4 were presented at the SPE Annual Technical Conference and Exhibition (ATCE) in 2015 (Grujic et al. [2015]), and at the Petroleum Geostatistics conference in Biarritz, France in 2015.

In chapter 5, we develop a regression tree based methodology for forecasting and interpretation of multivariate functional data. The developments were motivated by the need for an uncertainty quantification technique capable of forecasting both oil and gas rates simultaneously. The methodology was applied on both oil and gas rates of the Anadarko dataset considered in the previous chapter.

In chapter 6, we explore new avenues of research that were opened by the work outlined in the previous three chapters. In particular, we explore the use of geostatistics for functional data for emulation of computer codes that produce functional outputs. In addition, we develop two novel techniques for incorporation of secondary data into functional modeling workflows outlined previously and we conclude the chapter with an outline for future research directions. The work presented in chapter 6 was done in collaboration with Alessandra Menafoglio and it was recently submitted for publication in the journal of Stochastic Environmental Research and Risk Assessment (SERRA)(Grujic et al. [2017]).

In addition to the research presented in this dissertation, three R software packages were developed. The packages implement all of the methods discussed in this dissertation and are freely available in the public domain. The list of packages along with the web-links to their repositories is given below:

- The *DGSA* package that implements the DGSA sensitivity analysis method (Fenwick et al. [2014]). The package can be found at: www.github.com/ogru/DGSA
- The *fdagstat* package that implements the methods outlined in chapters 3,4 and 6. The package can be found at: www.github.com/ogru/fdagstat
- The *fTree* package that implements the functional regression trees methodologies outlined in chapter 5. The package can be found at: www.github.com/ogru/fTree

Chapter 2

Functional data analysis

In many fields of science and technology it is pretty common for data to come in a form of curves or surfaces, or in other words, data that varies over a continuum (i.e. space or time). Such data is often referred to as "functional" to emphasize its dependence on the continuum. Examples of functional data are temperature and precipitation measurements as a function of time in meteorology, oil and gas production as a function of time in petroleum engineering, well logs as a function of depth in geology, micro resonance images in medical research, etc. A statistical discipline that deals exclusively with interpretation, analysis, and prediction of functional data is called "Functional data analysis" or FDA in short. In this chapter, we will review the basic concepts of FDA that will serve as building blocks for the developments presented in subsequent chapters where we deal with the functional data that occurs in subsurface engineering. The developments presented below are mostly based on the seminal books by Ramsay and Silverman [2005], Ferraty and Vieu [2006], and Horváth and Kokoszka [2012].

2.1 An Infinite Dimensional Hilbert Space Framework for Functional Data

In statistics of scalars we analyze random variables that belong to the space of real numbers $\mathbb{R}$. While in multivariate statistics we analyze vectors as finite dimensional random variables that take values in finite dimensional Euclidean vector space $\mathbb{R}^n$.

Given that functional data varies over a continuum its dimensions are essentially infinite. For example, if we consider continuous curves that vary over time $\mathcal{X}(t), t \in T$ it is obvious that such data can be evaluated at infinitely many time steps that are all within T. Therefore, proper analysis of functional data requires an appropriate mathematical framework that would honor the infinite dimensional nature of the data. The most basic tools in FDA were developed around the idea that functional data comes as samples of a functional random variable, which was defined by Ferraty and Vieu [2006] as follows

Definition 2.1.: A random variable $\boldsymbol{\mathcal{X}}$ is called functional random variable if it takes values in an infinite dimensional space (or functional space). An observation (or a sample) of $\boldsymbol{\mathcal{X}}$ is denoted with $\mathcal{X}$.

Note that in this definition Ferraty and Vieu [2006] "implicitly make the following identification $\boldsymbol{\mathcal{X}} = \{\boldsymbol{\mathcal{X}}(t), t \in T\}$ and $\mathcal{X} = \{\mathcal{X}(t), t \in T\}$". In further text we will use the two notations interchangeably.

There are many infinite dimensional spaces (Hilbert, Banach, Sobolev,...) around which one can develop statistical frameworks for the analysis of functional data. The most basic developments in FDA assume that functional data takes values in an infinite dimensional $L^2(T)$ space, which is a separable Hilbert[1] space endowed with the inner product (Horváth and Kokoszka [2012]):

$$\langle \mathcal{X}_i(t), \mathcal{X}_j(t) \rangle = \int_T \mathcal{X}_i(t)\mathcal{X}_j(t)dt$$

This inner product induces the following norm:

$$\|\mathcal{X}(t)\| = \sqrt{\langle \mathcal{X}(t), \mathcal{X}(t) \rangle} = \sqrt{\int_T \mathcal{X}^2(t)dt}$$

Two additional assumptions are made in the $L^2(T)$ framework. The assumption of

[1]Separable vector spaces have a countable orthonormal basis. Hilbert space is a complete inner product space that can be also viewed as a generalization of the Euclidean vector space into high dimensions (Wik). A finite dimensional Euclidean space $\mathbb{R}^n$ is also a separable Hilbert space, by definition.

integrability of functions $\mathbb{E}[\|\mathcal{X}\|] < \infty$, and the assumption of square integrability of functions $\mathbb{E}[\|\mathcal{X}\|^2] < \infty$.
If the $L^2(T)$ space is also equipped with Borel's σ algebra then the assumption of integrability implies that there exists a unique mean function $\mu \in L^2(T)$ such that $\mathbb{E}[\langle f, \mathcal{X}\rangle] = \langle f, \mu\rangle, \quad \forall f \in L^2(T)$ (Horváth and Kokoszka [2012]). The mean function has the following important property:

$$\mu(t) = \mathbb{E}[\mathcal{X}(t)] \quad for \quad almost \quad all \quad t \in T \tag{2.1}$$

Implying that within $L^2(T)$ framework we can estimate the mean function from N available samples of $\boldsymbol{\mathcal{X}}(t)$, with the following intuitive relation (Horváth and Kokoszka [2012]):

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{X}(t), \quad \forall t \in T$$

From the assumption of square integrability follows the definition of covariance function[2]:

$$c(s,t) = \mathbb{E}\left[\left(\boldsymbol{\mathcal{X}}(t) - \mu(t)\right)\left(\boldsymbol{\mathcal{X}}(s) - \mu(s)\right)\right], \quad s,t \in T \tag{2.2}$$

Which is in practice estimated from a sample of N curves with the following relation (Horváth and Kokoszka [2012]):

$$\hat{c}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N} \left(\mathcal{X}_i(t) - \mu(t)\right)\left(\mathcal{X}_i(s) - \mu(s)\right)$$

In this dissertation, we will adopt this $L^2(T)$ framework for analyzing functional data. In other words, in all our developments we will assume that the analyzed functions are realizations of a random process that takes values in $L^2(T)$ separable Hilbert space equiped with the inner product and the induced norm, and also that functions are square integrable.

[2]Operators on Hilbert spaces are beyond the scope of this dissertation. Interested readers should consider Horváth and Kokoszka [2012] for a detailed dissection of the topic.

2.2 From Discrete Observations to Functions

All FDA analyses start from observations of functional data that are often given in pairs $(y_{ij}, t_{ij}), t \in T$. Here $i = 1, 2, ..., N$ represents realizations of functional data and $j = 1, 2, ..., M_i$ represents the index of observations for $i - th$ function. Sampling along T can be coincident or it can vary across realizations which makes the analysis and comparison of realizations difficult. In addition, the data are often observed with noise that adds another level of analytic difficulty. The main assumption behind the methods of FDA is that observations of functional data were generated by a smooth process corrupted by noise.

$$y_{ij} = \mathcal{X}_i(t_{ij}) + \epsilon \tag{2.3}$$

Therefore, the first question of FDA is how to estimate the smooth and continuous function $\mathcal{X}_i(t)$ from its raw and noisy observations. In FDA, this is achieved by means of basis expansion, a non-parametric curve fitting procedure. Basis expansion starts from a selection of an appropriate basis system that consists of a finite number of analytic basis functions. The analytic basis functions span the entire domain T. The type of the basis system is data dependent. When functional data are non-periodic, the most commonly used analytic basis functions are B-splines (Boor [1978]), while in the case of periodicity in the data, periodic (trigonometric) basis functions are often used. Hence, we distinguish two most commonly used basis systems, B-Spline and Fourier basis systems (figure 2.1).

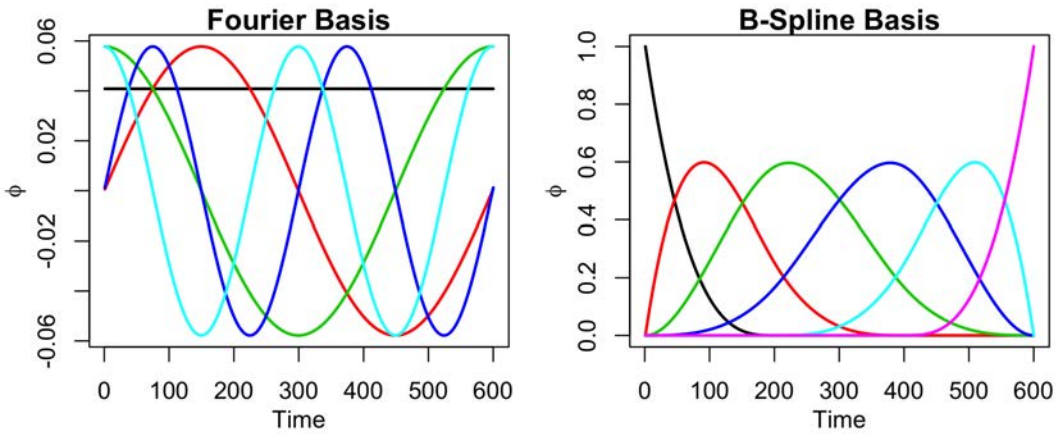


Figure 2.1: *An example of basis systems. Left - Fourier Basis, Right - B-Spline Basis*

After selecting the basis system one has to expand it to the observed data. This is achieved by scaling the basis functions with appropriate set of coefficients and summing up the scaled products

$$\hat{\mathcal{X}}_i(t) = \sum_{k=1}^{K} c_{ik} b_k(t) \tag{2.4}$$

This procedure is commonly referred to as data smoothing since it results in a smooth function fitted to noisy observations. The most appropriate coefficients for basis expansion are found such that the following least squares objective functional is minimized

$$\underset{c_{i1},c_{i2},\ldots,c_{ik}}{\operatorname{argmin}} \sum_{j=1}^{T} \left(\mathcal{X}(t_j) - \sum_{k=1}^{K} c_{ik} b_k(t_j) \right)^2 \tag{2.5}$$

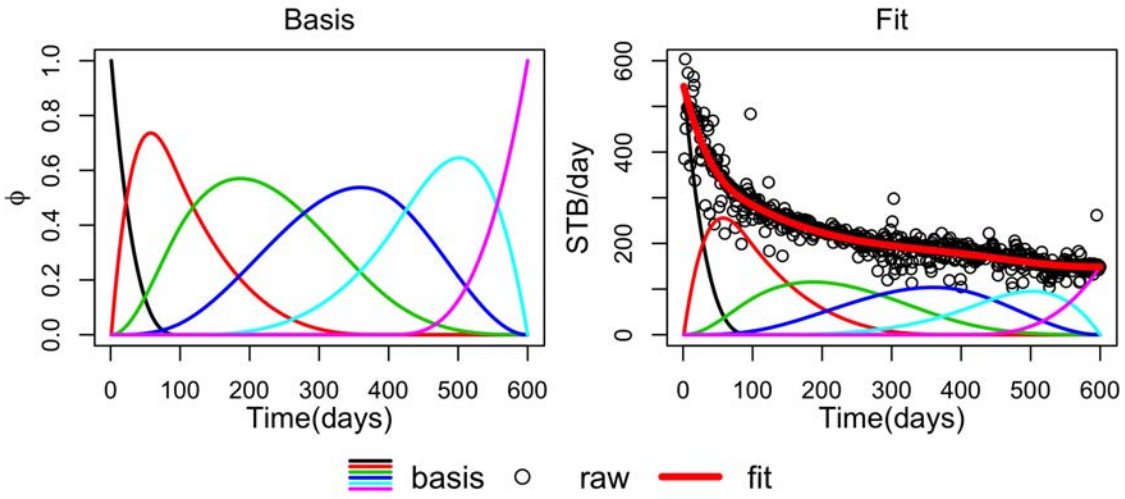


Figure 2.2: *An example of basis expansion. Left - Selected B-spline basis system, Right - Basis expansion and the resulting fit.*

Since smoothing is achieved by scaling of analytic basis functions whose derivatives are known, the derivatives of the resulting fit are also known[3]. This opens doors to deeper derivative analyses of the data that are unavailable to multivariate statistics, and also to a more complex fitting criterion that imposes smoothness penalty on the

[3]The derivative of a sum is a sum of the derivatives.

second derivative of the fit

$$\underset{c_{i1},c_{i2},\ldots,c_{ik}}{\operatorname{argmin}} \sum_{j=1}^{T} \left(\mathcal{X}(t_j) - \sum_{k=1}^{K} c_{ik} b_k(t_j) \right)^2 + \lambda \int_T \hat{\mathcal{X}}''(t) dt \tag{2.6}$$

The quality of the fit depends on the number of basis functions and the magnitude of the smoothing penalty (λ). Too many basis functions might result in over-fitting, while too high smoothing penalty might cause the fit to be overly smooth and miss important features in the data. An example of the influence of these two parameters on the resulting fit is given in figure 2.3.

Ramsay and Silverman [2005] outline a generalized cross validation (GCV) procedure for estimation of the number of basis functions and the smoothing penalty. The idea is simply to remove a part of the observations, perform basis expansion and compute the error between the removed observations and the corresponding smooth values. The number of basis functions and the value of the regularization coefficient that yields the lowest mean cross validation error are deemed the best. In our experience these are just tools to help users narrow down ranges for fitting the parameters. The final decision on how many basis functions to use and the degree of smoothing is problem dependent and always at the discretion of the analyst.

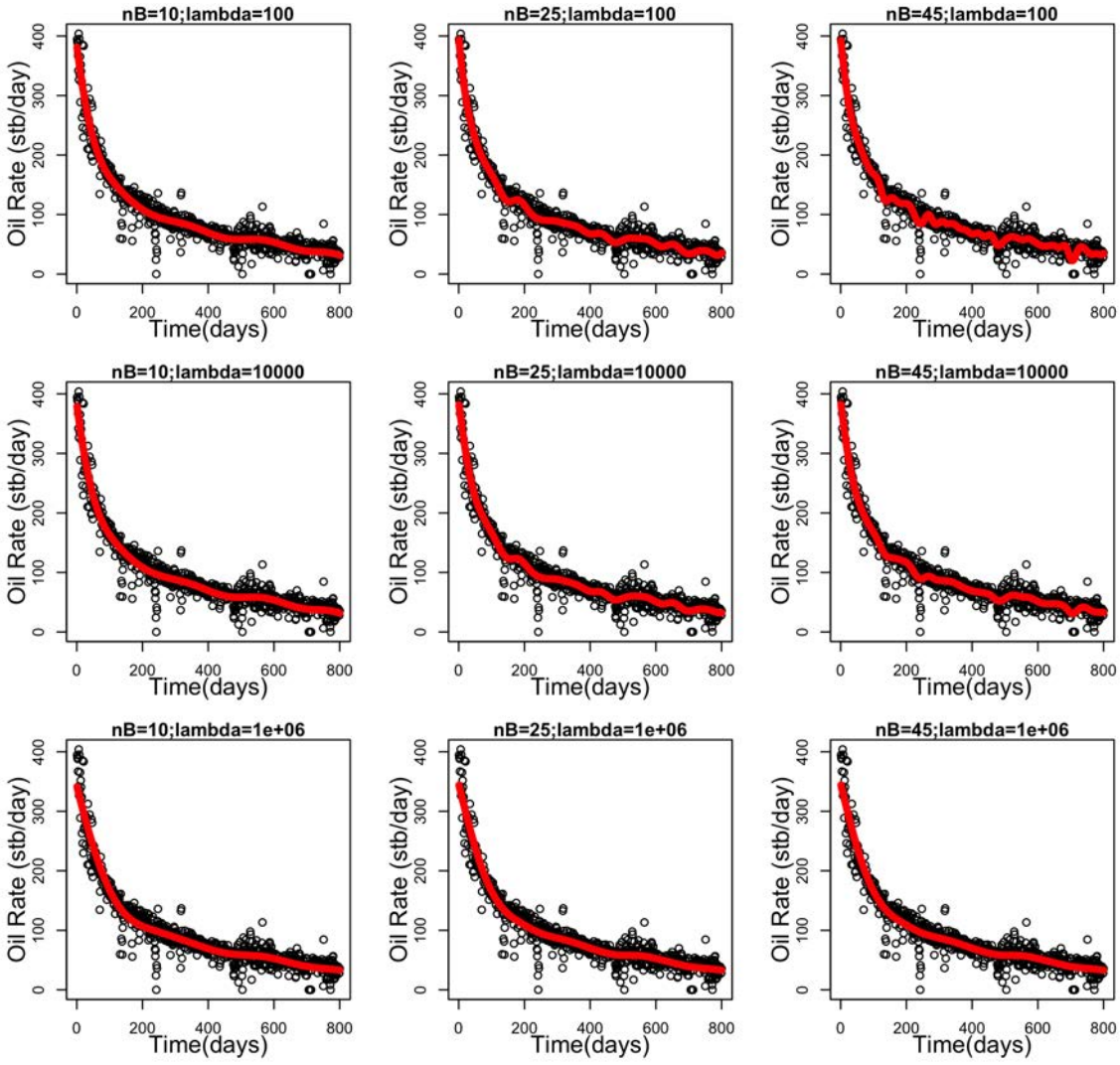


Figure 2.3: *The influence of the number of basis functions and the smoothing penalty on basis expansion. (nB - the number of basis functions; lambda - the value of the smoothing penalty (equation 2.6))*

2.3 Functional Principal Component Analysis

Functional data analysis community invented a functional version of the principal component analysis. The technique is analogous to the conventional (multivariate) principal component analysis, and it represents an irreplaceable component of every

functional data analysis study. The developments presented below start from a computational perspective since in that way it is much easier to understand functional principal component analysis and see the many analogies it has with conventional principal component analysis.

Let $\{\mathcal{X}_i(t), t \in T\}_{i=1}^N$ be a set of curves and let $\{\boldsymbol{x}_i \in R^m\}_{i=1}^N$ be a set of their "vectorized" equivalents produced by evaluating each curve over a fine equidistant grid $t_1, t_2, ..., t_m$ that spans the entire time domain T. Let $\mu(t)$ and $\boldsymbol{\mu}$ be the mean function and vector respectively, let $\{\boldsymbol{x}_i^c \in R^m\}_{i=1}^N$ be a set of centered vectors produced by subtracting the mean vector from each $\boldsymbol{x}_i$ and let $\boldsymbol{\Sigma}_{m\times m}$ be the empirical variance-covariance matrix of $\boldsymbol{x}_i^c$'s. Note that this variance-covariance matrix is a discretized equivalent of the covariance function (eq (2.2)), $c(t_i, t_j) = \boldsymbol{\Sigma}[i, j]$. By definition, principal component decomposition of the variance-covariance matrix $\boldsymbol{\Sigma}_{m\times m}$ is given with

$$\boldsymbol{\Sigma}_{m\times m} = \boldsymbol{V}_{m\times m}\boldsymbol{\Lambda}_{m\times m}\boldsymbol{V}_{m\times m}^T \tag{2.7}$$

Where the columns of matrix $\boldsymbol{V}$ are orthogonal eigen-vectors (principal components) and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigen values $(\lambda_1, \lambda_2, ..., \lambda_m)$. We will refer to the columns of $\boldsymbol{V}$ as $\boldsymbol{\phi}_j, j \in [1, m]$, and assume that index j corresponds to the decreasing order of eigen values[4].

Notice that the length of the eigen vectors is the same as the length of the fine equidistant grid. This suggest that the coefficients of eigen vectors are time dependent. As a matter of fact every $\boldsymbol{\phi}_j$ is a "vectorized" version of its functional equivalent $\phi_j(t)$ that takes values in the same L^2 Hilbert space as the original data. $\phi_j(t)$'s are commonly referred to as functional principal components or "fpcs" in short. Fpc's are orthogonal, meaning that

$$\langle \phi_j(t), \phi_k(t) \rangle = \int_T \phi_j(t)\phi_k(t)dt = \langle \boldsymbol{\phi}_j, \boldsymbol{\phi}_k \rangle = 0 \quad \forall j, k = [1, 2, ..., m]; \quad j \neq k \tag{2.8}$$

and as such they form an ortho-normal basis in $L^2(T)$.

Analogously to the conventional multivariate principal component analysis fpca also makes use of the principal component scores that are given with

$$\xi_i^k = \langle \mathcal{X}(t) - \mu(t), \phi_k(t) \rangle = \int_T ((\mathcal{X}(t) - \mu(t))\phi_k(t))\, dt \tag{2.9}$$

[4] As returned by many software implementations of eigen value decomposition (i.e. *prcomp* in R).

and in practice computed with $\xi_i^k \sim \langle \boldsymbol{x}_i, \boldsymbol{\phi}_k \rangle$. Principal component scores are by definition uncorrelated $\left(cov\left(\xi_i^k, \xi_i^l\right) = 0, \forall k, l \quad s.t. \quad k \neq l\right)$.
As in multivariate principal component analysis functional principal components describe variance in functional data. More specifically, the variance of principal component scores of the k-th fpc equals the k-th eigen value

$$\lambda_k = \frac{1}{N-1} \sum_{i=1}^{N} (\xi_i^k)^2 \tag{2.10}$$

An interesting property of the fpcs is that they can be used to "reconstruct" every function in the ensemble

$$\hat{\mathcal{X}}_i(t) = \mu(t) + \sum_{j=1}^{k} \xi_i^j \phi_j(t), \quad k \leq m \tag{2.11}$$

Given that this reconstruction is achieved by multiplying each eigen function with a scalar (that is very similar to basis expansion) functional principal components are often viewed as an ideal basis system that most adequately represents a given ensemble of curves. In practice, we often work with a truncated fpc basis system formed out of k leading principal components that describe most of the variance in the data. The most commonly used criterion for selecting k is the fraction of variance explained (FVE)

$$FVE_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \tag{2.12}$$

where we select the lowest k for which FVE_k is higher or equal to some user selected threshold (often 0.95). Working with a truncated basis of functional principal components adds additional layer to data smoothing. Higher-order fpcs often describe small variations in functional data that can be attributed to noise. Discarding these higher order fpcs in functional data reconstruction (eq. 2.11) results in smoother curves than was the case with basis expansion that preceded the fpca. In practice, basis expansion, fpca and fpc reconstruction (eq. 2.11) are done in sequence, iteratively, until acceptable functional representations of the data are achieved.

Another interesting property of functional principal components is that they are interpretable. Ramsay and Silverman [2005] suggested plotting fpcs as perturbation

around the mean function

$$\mu(t) \pm \sqrt{\lambda_j}\phi_j(t)$$

to better understand the modes and the degrees of variation that they describe. One example of functional principal component analysis is given in figure (2.4) with the fpcs plotted as perturbations around the mean. We can see that the first fpc describes the overall variation in temperature data. The second component appears to describe the summer and winter variation jointly. The third and the fourth fpcs are more difficult to interpret.

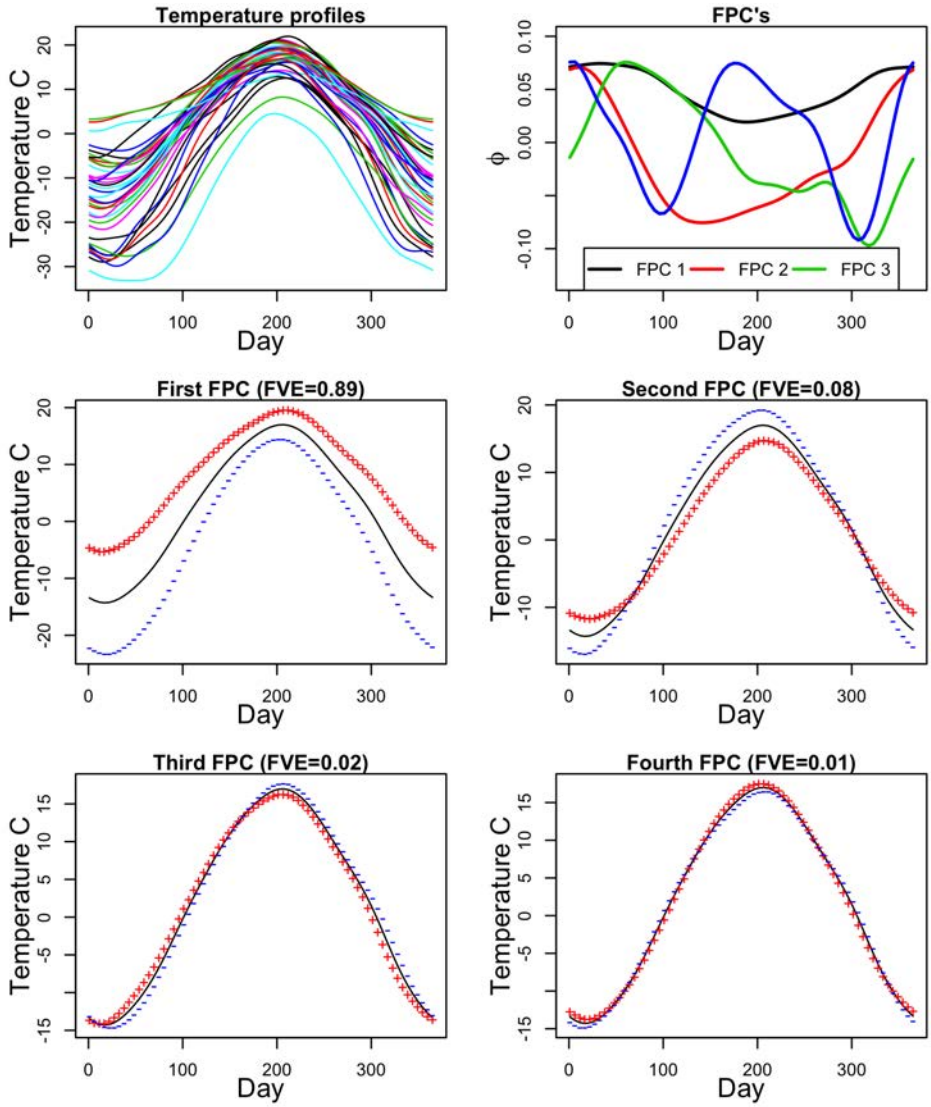


Figure 2.4: *An example of fpca*[5]*. Top row - smooth functional data and associated functional principal components. Middle and bottom rows - FPCS as perturbation (+-) around the mean(black).*

As mentioned before, functional principal components form an ortho-normal basis in $L^2(T)$ Hilbert space (or R^n) that best describes the functional data in terms of variance. While such basis is by definition the best, it is not unique. There are other

[5]This dataset and the analysis were adapted from Ramsay and Silverman [2005]. The data and the functions that produce these plots are a part of the *fda* R package (Ramsay et al. [2009]).

orthogonal bases of the same dimension that describe the same amount of variance in the data (Ramsay and Silverman [2005]). Ramsay and Silverman [2005] suggested that the components of some of those other bases might be more interpretable (than the original fpcs) when plotted as perturbations around the mean function. All such orthogonal bases can be obtained from the basis of fpcs with a simple orthogonal rotation. For example, consider an orthogonal rotation matrix $\boldsymbol{R}$ ($\boldsymbol{R}^T\boldsymbol{R} = I$) and a matrix of fpcs stacked in columns $\boldsymbol{V}$. The rotated functional principal components can be computed as follows[6]

$$\boldsymbol{V}_r = \boldsymbol{R}\boldsymbol{V}^T \tag{2.13}$$

The question is how to find $\boldsymbol{R}$ that produces the most interpretable components in $\boldsymbol{V}_r$? While there are many different criteria to consider, practice has shown that the VARIMAX (Kaiser [1958]) rotation criterion produces the most optimal results (Ramsay and Silverman [2005]). The idea behind VARIMAX is to rotate the fpcs such that they align, as much as possible, with the directions of the time steps in R^n. Consider an fpc vector $\boldsymbol{\xi_i} = [\xi_i(t_1), \xi_i(t_2), ..., \xi_i(t_n)]$. To say that $\boldsymbol{\xi_i}$ is perfectly aligned with the direction of the $j-th$ time step implies that $\xi_i(t_j) = \pm 1$ and $\xi_i(t_k) = 0, \forall k \neq j$. Therefore, the VARIMAX rotation criterion aims to find $\boldsymbol{R}$ that produces $\boldsymbol{V}_r$ whose values are, or very close to: 1, -1 or zero[7]. This is implicitly accomplished by maximizing the variance of the values in $\boldsymbol{V}_r$ hence the name VARIMAX.

VARIMAX rotation is always performed on a truncated basis of fpcs (i.e. for FVE=0.95). One example of VARIMAX rotated fpcs is given in figure 2.5 where we plot the four rotated fpcs as perturbations around the mean on the previously considered temperature dataset. From this analysis, we can see that the first rotated fpc represents winter temperature variations, the second and third represent spring and autumn temperature variations respectively, while the fourth rotated fpc describes summer temperature variations. Clearly, these rotated fpcs are much more interpretable than their unrotated counterparts.

The rotated fpcs describe the same amount of variance in the data as the original fpcs, however, they distribute the variance differently among components (figure 2.5 titles).

[6]Note that here the rotation is performed on raw eigen (unit) vectors and not on loadings as it is commonly done in factor analysis.

[7]That is while maintaining the columns as unit vectors

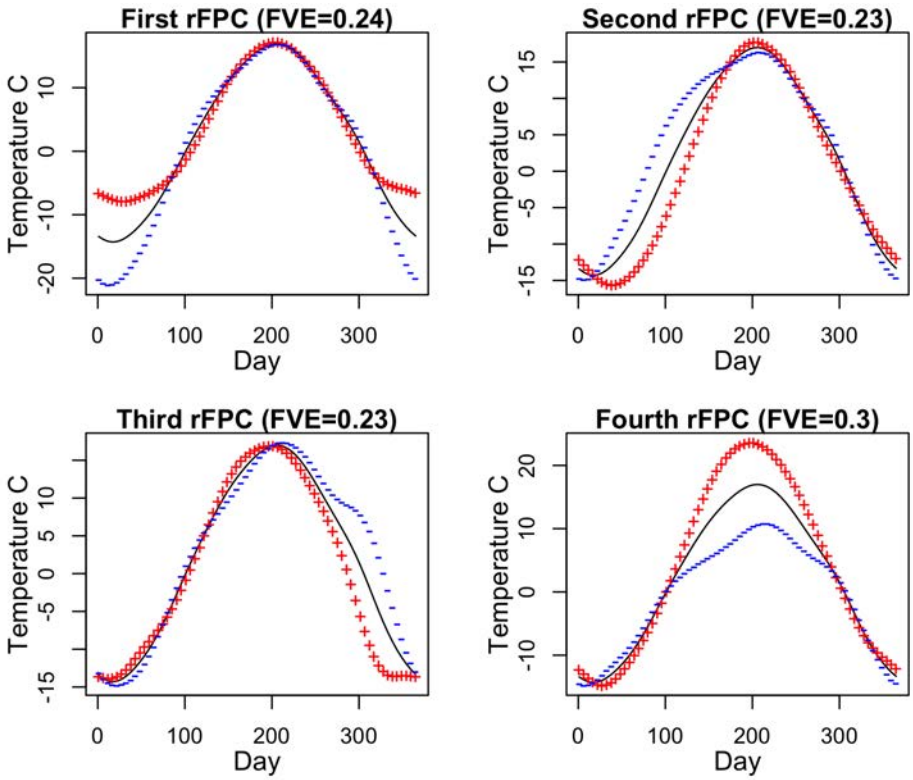


Figure 2.5: *Varimax rotated functional principal components as perturbation (+-) around the mean function.*

2.4 Regression for Functional Data

In many modern applications of FDA, it is quite common for functional data to be observed along with a set of explanatory variables (covariates/predictors). In such setting the question of forecasting of functional data naturally arises. Ramsay and Silverman [2005] outlined several methods for forecasting of functional data, and for forecasting with functional data. Since the former are the most interesting for the problems addressed in this dissertation, in this section, we will review two of the commonly used methods for functional data forecasting. The first method relies on functional principal component analysis, as an intermediate modeling step, while the second method is capable of directly predicting functional data from a given set of covariates. In all our developments, we will refer to observed data as $\{\mathcal{X}_i(t), \boldsymbol{z}_i\}_{i=1}^{N}$

where $t \in T$ and $\boldsymbol{z} \in R^n$, forecasted function as $\mathcal{X}_0(t)$ and associated known vector of explanatory variables (predictors) as $\boldsymbol{z}_0$.

2.4.1 Functional Principal Component Regression

As the name suggests, functional principal component regression relies on functional principal component decomposition of available functional data. Let $\mu(t)$ be the mean function of the observed functional data, $e_k : \{\phi_1(t), \phi_2(t), ..., \phi_k(t)\}$ be a set of ortho-normal functional principal components, and $\xi_i^j, j \in [1, 2, ..., k]$ be associated functional principal component scores. Recall that all observed functions can be reconstructed from the mean, the fpc's and the associated fpc scores (eq. (2.11)). The same relationship can be applied to the forecasted function $\mathcal{X}_0(t)$ as follows:

$$\hat{\mathcal{X}}_0(t) = \mu(t) + \sum_{j=1}^{k} \xi_0^j \phi_j(t) \tag{2.14}$$

If the mean and the fpcs are estimated from all available functional data and assumed constant, the only thing missing in order to fully predict function $\mathcal{X}_0(t)$ are its fpc scores ξ_0^j. The idea of functional principal component regression is to predict $\mathcal{X}_0(t)$ by predicting its functional principal component scores from associated covariates $\boldsymbol{z}_0$. This is achieved by means of multiple regression where predictions are given as linear combinations of the principal component scores of all already observed functions and appropriate regression coefficients

$$\xi_0^j = \sum_{l=0}^{L} f_l(\boldsymbol{z}_0)\beta_l^j \tag{2.15}$$

here $f_l(.)$ represents l-th transformation function[8] and β_l^j represents l-th regression coefficient.

The regression coefficients are estimated from all available training data as a solution

[8] for example $f_0(\boldsymbol{z}) = 1, f_1(\boldsymbol{z}) = z_1, f_2(\boldsymbol{z}) = z_2, ..., f_j(\boldsymbol{z}) = z_1^2, f_{j+1}(\boldsymbol{z}) = z_2^2, ...$

to the following minimization problem

$$\underset{\beta_0^j,\beta_1^j,...,\beta_L^j}{\operatorname{argmin}} \sum_{i}^{N} \left(\xi_i^j - \sum_{l=0}^{L} f_l(\boldsymbol{z}_0)\beta_l^j \right)^2 \tag{2.16}$$

the solution of this optimization problem is found in a closed form with normal equations[9]

$$\boldsymbol{\beta}_j = (\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{Y}_j \tag{2.17}$$

where $\boldsymbol{\beta}_j = \left[\beta_0^j, \beta_1^j, ..., \beta_L^j\right]$, $\boldsymbol{Y}_j = \left[\xi_1^j, \xi_2^j, ..., \xi_N^j\right]$, and

$$\boldsymbol{F} = \begin{bmatrix} f_0(\boldsymbol{z}_1) & f_1(\boldsymbol{z}_1) & \cdots & f_L(\boldsymbol{z}_1) \\ f_0(\boldsymbol{z}_2) & f_1(\boldsymbol{z}_2) & \cdots & f_L(\boldsymbol{z}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\boldsymbol{z}_N) & f_1(\boldsymbol{z}_N) & \cdots & f_L(\boldsymbol{z}_N) \end{bmatrix}$$

To develop a functional regression forecasting model one needs to fit k such multiple regression models, one for each functional principal component. Note that this approach exploits the fact that principal component scores are uncorrelated and as such it cannot be applied directly to varimax rotated functional principal component scores. If forecasting of rotated fpcs is desired one might consider a more advanced form of regression with multiple inputs and multiple outputs called "Currds and Whey" procedure as proposed by Breiman and Friedman [1997].

2.4.2 Functional Regression

Ramsay and Silverman [2005] outlined a predictive procedure called functional regression. The idea is to forecast functions as a linear combination of scalar predictors and associated coefficient functions

$$\mathcal{X}_0(t) = \sum_{l=0}^{L} f_l(\boldsymbol{z_0})a_l(t) \tag{2.18}$$

Where $f_l(.)$ are transformation functions, and $a_l(t)$ are coefficient functions.

[9] for more details on multiple regression please see Hastie et al. [2009].

To fit a regression model one needs to infer coefficient functions $a_l(t)$. This is done in least squares sense by minimizing the following objective functional

$$\underset{a_1(t),a_2(t),...,a_L(t)}{\operatorname{argmin}} \sum_{i}^{N} \int_T \left(\mathcal{X}_i(t) - \sum_{l=0}^{L} f_l(\boldsymbol{z_i}) a_l(t) \right)^2 dt \tag{2.19}$$

In practice this objective functional is minimized by first assuming a common basis system (i.e. B-spline) $b_k(t)(k = 1, 2, ..., K)$ and expressing each $a_l(t)$ in terms of it

$$\underset{c_1^1,c_1^2,...,c_l^k}{\operatorname{argmin}} \sum_{i}^{N} \int_T \left(\mathcal{X}_i(t) - \sum_{l=0}^{L} f_l(\boldsymbol{z_i}) \sum_{k=1}^{K} c_l^k b_k(t) \right)^2 dt \tag{2.20}$$

In this way the problem of inferring coefficient functions is transformed into a problem of inferring the coefficients of their basis expansion. Additionally, the basis and the functions are evaluated over a common fine equidistant grid $t_1, t_2, ..., t_M$, that replaces the integral with a sum

$$\underset{c_1^1,c_1^2,...,c_l^k}{\operatorname{argmin}} \sum_{i}^{N} \sum_{j=1}^{M} \left(\mathcal{X}_i(t_j) - \sum_{l=0}^{L} f_l(\boldsymbol{z_i}) \sum_{k=1}^{K} c_l^k b_k(t_j) \right)^2 = \underset{c_1^1,c_1^2,...,c_l^k}{\operatorname{argmin}} Q \tag{2.21}$$

The solution to this transformed objective functional is sought in a conventional way by setting the partial derivatives of the basis coefficients to zero

$$\frac{\partial Q}{\partial c_1^1} = \frac{\partial Q}{\partial c_1^2} = ... = \frac{\partial Q}{\partial c_2^1} = \frac{\partial Q}{\partial c_2^2} = ... = \frac{\partial Q}{\partial c_l^1} = \frac{\partial Q}{\partial c_l^2} = ... = \frac{\partial Q}{\partial c_l^1} = 0 \tag{2.22}$$

It is easy to show that the solution to this optimization problem is found in a closed form with the following system of normal equations

$$\begin{bmatrix} \mathcal{X}_1(t_1) \\ \mathcal{X}_1(t_2) \\ \vdots \\ \mathcal{X}_1(t_M) \\ \mathcal{X}_2(t_1) \\ \mathcal{X}_2(t_2) \\ \vdots \\ \mathcal{X}_2(t_M) \\ \vdots \\ \mathcal{X}_N(t_1) \\ \mathcal{X}_N(t_2) \\ \vdots \\ \mathcal{X}_N(t_M) \end{bmatrix} = \begin{bmatrix} f_0(\boldsymbol{z}_1)\boldsymbol{b}(t_1), f_1(\boldsymbol{z}_1)\boldsymbol{b}(t_1), \cdots, f_L(\boldsymbol{z}_1)\boldsymbol{b}(t_1) \\ f_0(\boldsymbol{z}_1)\boldsymbol{b}(t_2), f_1(\boldsymbol{z}_1)\boldsymbol{b}(t_2), \cdots, f_L(\boldsymbol{z}_1)\boldsymbol{b}(t_2) \\ \vdots \\ f_0(\boldsymbol{z}_1)\boldsymbol{b}(t_M), f_1(\boldsymbol{z}_1)\boldsymbol{b}(t_M), \cdots, f_L(\boldsymbol{z}_1)\boldsymbol{b}(t_M) \\ f_0(\boldsymbol{z}_2)\boldsymbol{b}(t_1), f_1(\boldsymbol{z}_2)\boldsymbol{b}(t_1), \cdots, f_L(\boldsymbol{z}_2)\boldsymbol{b}(t_1) \\ f_0(\boldsymbol{z}_2)\boldsymbol{b}(t_2), f_1(\boldsymbol{z}_2)\boldsymbol{b}(t_2), \cdots, f_L(\boldsymbol{z}_2)\boldsymbol{b}(t_2) \\ \vdots \\ f_0(\boldsymbol{z}_2)\boldsymbol{b}(t_M), f_1(\boldsymbol{z}_2)\boldsymbol{b}(t_M), \cdots, f_L(\boldsymbol{z}_2)\boldsymbol{b}(t_M) \\ \vdots \\ f_0(\boldsymbol{z}_N)\boldsymbol{b}(t_1), f_1(\boldsymbol{z}_N)\boldsymbol{b}(t_1), \cdots, f_L(\boldsymbol{z}_N)\boldsymbol{b}(t_1) \\ f_0(\boldsymbol{z}_N)\boldsymbol{b}(t_2), f_1(\boldsymbol{z}_N)\boldsymbol{b}(t_2), \cdots, f_L(\boldsymbol{z}_N)\boldsymbol{b}(t_2) \\ \vdots \\ f_0(\boldsymbol{z}_N)\boldsymbol{b}(t_M), f_1(\boldsymbol{z}_N)\boldsymbol{b}(t_M), \cdots, f_L(\boldsymbol{z}_N)\boldsymbol{b}(t_M) \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_0 \\ \boldsymbol{c}_1 \\ \vdots \\ \boldsymbol{c}_L \end{bmatrix} \tag{2.23}$$

Where $\boldsymbol{b}(t_j) = [b_1(t_j), b_2(t_j), ..., b_K(t_j)]$, and $\boldsymbol{c}_l = \left[c_l^1, c_l^2, ..., c_l^K\right]$.
This approach to fitting a functional regression model can be applied to pre-smoothed functions or directly to their raw (noisy) observations. In the case of raw data, one might add an additional smoothing term that penalizes the second derivative of the functional regression coefficients. In this case, the objective functional becomes

$$\underset{a_1(t),a_2(t),...,a_L(t)}{\operatorname{argmin}} \sum_i^N \int_T \left(\mathcal{X}_i(t) - \sum_{l=0}^{L} f_l(\boldsymbol{z_i}) a_l(t) \right)^2 dt + \lambda \int_T a_l''(t) dt \tag{2.24}$$

The solution to this optimization problem is also found in a closed form, however its derivations are beyond the scope of this dissertation. Interested readers should consider chapter 12 of Ramsay and Silverman [2005] for more details.

When functional data is pre-smoothed (i.e. with basis expansion) and the same basis is used for data and the coefficients, then, the coefficient functions found with the procedure outlined above coincide with the coefficient functions estimated by applying conventional multiple regression on each time step $t_1, t_2, ..., t_M$. The procedure is as follows. For one time step (t_j) the values of functional coefficients are sought as

solutions to the following optimization problem

$$\underset{a_1(t_j),a_2(t_j),...,a_L(t_j)}{\operatorname{argmin}} \sum_{i}^{N} \left(\mathcal{X}_i(t_j) - \sum_{l=0}^{L} f_l(\boldsymbol{z_i}) a_l(t_j) \right)^2 \tag{2.25}$$

the solution to this optimization problem is given in closed form

$$\boldsymbol{a}(t_j) = (\boldsymbol{F}^T \boldsymbol{F})^{-1} \boldsymbol{F}^T \boldsymbol{y}_j \tag{2.26}$$

where $\boldsymbol{a}(t_j) = [a_1(t_j), a_2(t_j), ..., a_L(t_j)]$, $\boldsymbol{y}_j = [\mathcal{X}_1(t_j), \mathcal{X}_2(t_j), ..., \mathcal{X}_N(t_j)]$, and

$$\boldsymbol{F} = \begin{bmatrix} f_0(\boldsymbol{z}_1) & f_1(\boldsymbol{z}_1) & \cdots & f_L(\boldsymbol{z}_1) \\ f_0(\boldsymbol{z}_2) & f_1(\boldsymbol{z}_2) & \cdots & f_L(\boldsymbol{z}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\boldsymbol{z}_N) & f_1(\boldsymbol{z}_N) & \cdots & f_L(\boldsymbol{z}_N) \end{bmatrix}$$

It should be emphasized that this piece-wise parameter inference procedure is appropriate only in the case of pre-smoothed functional data due to the fact that it does not explicitly enforce smoothness of functional coefficient functions.

2.5 Curve Completion

In subsurface engineering, it is quite common for functional data to be sparsely or partially observed over analyzed time domain. One such example are hydrocarbon decline curves in unconventional reservoir engineering where the temporal length of hydrocarbon production profiles varies between wells. The resulting fit of basis expansion on such partially observed functional data becomes unstable and often produces sub-optimal results (figure 2.6). This is because basis functions get too many degrees of freedom in time domains where functional data was not observed. Given that all FDA techniques assume that all realizations of functional data were observed (or can be adequately smoothed) over the same time domain, some sort of smoothing with data completion is necessary. In oil and gas engineering this is routinely accomplished with parametric models (i.e. decline curves). While widely used, parametric models are too rigid and often unable to adequately capture all important features

in functional data. As an alternative, in this dissertation we will use a very robust non-parametric curve completion approach based on functional principal component analysis. The method allows us to smooth incomplete curves with the information extracted from complete functional data[10]. In petroleum engineering terms this means that we are learning the parameters of production decline from all available longer producing wells and use it to smooth and complete production profiles of wells with shorter production.

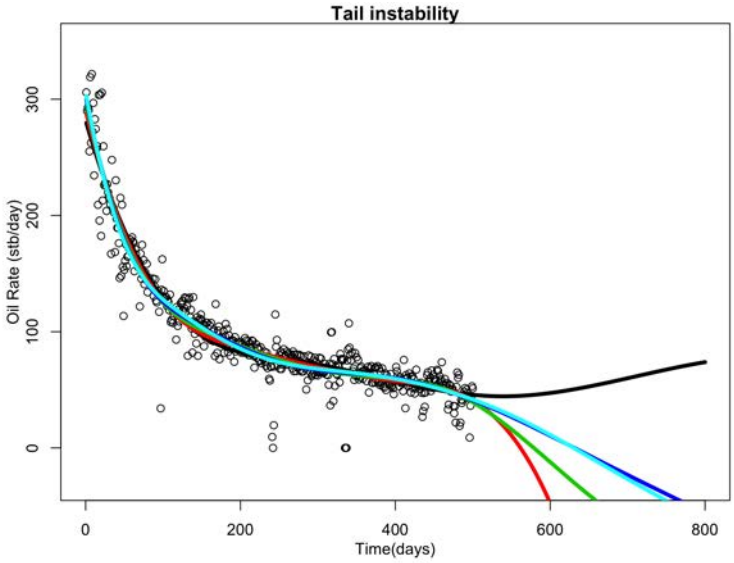


Figure 2.6: *An example of unstable fit resulting from partially observed functional data.*

Let $\{\mathcal{X}_i(t), t \in T\}_{i=1}^{N}$ be a set of smooth functions fully observed on time domain T, let $e_k : \{\phi_1(t), \phi_2(t), ..., \phi_k(t)\}$ be a set of functional principal components, let ξ_i^k be associated functional principal component scores, and let $\{y_j, \tau_j\}_{j=1}^{J}$ be a set of raw observations of curve $\mathcal{Y}(t)$ observed over a subset of the time domain $T(\tau_j \in \mathcal{T} \subset T)$. Here we assume that $\mathcal{X}_i(t)$'s and $\mathcal{Y}(t)$ were generated by the same data generating process.

[10]The method originates from the course notes by Giles Hooker (http://faculty.bscb.cornell.edu/~hooker/FDA2008/Lecture10_handout.pdf).

Recall the fpc reconstruction formula (eq. (2.11))

$$\hat{\mathcal{X}}_i(t) = \mu(t) + \sum_{j=1}^{k} \xi_i^j \phi_j(t) \tag{2.27}$$

and consider its "vectorized" form produced by evaluating $\mathcal{X}_i(t), \mu(t)$ and $\phi_k(t)$'s over a common regular grid $t_1, t_2, ..., t_M$ that spans the entire time domain T.

$$\begin{bmatrix} \mathcal{X}_i(t_1) \\ \mathcal{X}_i(t_2) \\ \vdots \\ \mathcal{X}_i(t_M) \end{bmatrix} = \begin{bmatrix} \mu(t_1) \\ \mu(t_2) \\ \vdots \\ \mu(t_M) \end{bmatrix} + \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_k(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_k(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_M) & \phi_2(t_M) & \cdots & \phi_k(t_M) \end{bmatrix} \begin{bmatrix} \xi_i^1 \\ \xi_i^2 \\ \vdots \\ \xi_i^k \end{bmatrix} \tag{2.28}$$

This equation suggests that when the mean and the fpc's are known, functional principal component scores can be computed with the following relation

$$\begin{bmatrix} \xi_i^1 \\ \xi_i^2 \\ \vdots \\ \xi_i^k \end{bmatrix} = \begin{bmatrix} \phi_1(t_1) & \phi_2(t_1) & \cdots & \phi_k(t_1) \\ \phi_1(t_2) & \phi_2(t_2) & \cdots & \phi_k(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(t_M) & \phi_2(t_M) & \cdots & \phi_k(t_M) \end{bmatrix}^T \left(\begin{bmatrix} \mathcal{X}_i(t_1) \\ \mathcal{X}_i(t_2) \\ \vdots \\ \mathcal{X}_i(t_M) \end{bmatrix} - \begin{bmatrix} \mu(t_1) \\ \mu(t_2) \\ \vdots \\ \mu(t_M) \end{bmatrix} \right) \tag{2.29}$$

This relation can be used to compute functional principal component scores of function $\mathcal{Y}(t)$ from its raw observations (y_j, τ_j) by evaluating $\phi_k(t)$'s and $\mu(t)$ at times where the raw data was observed:

$$\begin{bmatrix} \xi_y^1 \\ \xi_y^2 \\ \vdots \\ \xi_y^k \end{bmatrix} = \begin{bmatrix} `\phi_1(\tau_1) & \phi_2(\tau_1) & \cdots & \phi_k(\tau_1) \\ \phi_1(\tau_2) & \phi_2(\tau_2) & \cdots & \phi_k(\tau_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\tau_J) & \phi_2(\tau_J) & \cdots & \phi_k(\tau_J) \end{bmatrix}^T \left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{bmatrix} - \begin{bmatrix} \mu(\tau_1) \\ \mu(\tau_2) \\ \vdots \\ \mu(\tau_J) \end{bmatrix} \right) \tag{2.30}$$

Once the fpc scores are computed, $\mathcal{Y}(t)$ can be expressed in terms of $\mu(t)$ and $\phi_k(t)$ with equation (2.28), and as such it is fully defined on the entire time domain T.

The previously outlined curve completion procedure is very robust and applicable to most types of functional data. However, in our experience, it requires a pretty

large sample of complete data (> 50) before it starts producing accurate results. This mainly has to do with the estimation of the mean function ($\mu(t)$) and the functional principal components ($\xi_i(t)$). The discussion that follows is concerned with the question of how to use all available raw functional data (complete and incomplete) to better estimate the mean function and the functional principal components.

The estimation of $\mu(t)$. Starting from raw observations of N realizations of functional data (y_{ij}, t_{ij}) one can estimate the mean function by computing the estimates of the mean at every available time step, with all raw observations available at that time step

$$\mu(t_j) = \frac{1}{|N(t_j)|} \sum_{i \in N(t_j)} y_{ij} \tag{2.31}$$

Where $N(t_j)$ is a set of curves for which raw observations are available at time step t_j. After computing the raw estimates of the mean, at all available time steps, one can simply employ basis expansion (curve smoothing) to arrive to its continuous smooth functional estimate, $\mu(t)$.

The estimation of the FPCS. To estimate functional principal components it is sufficient to estimate the corresponding covariance matrix from all available data. This can be achieved by estimating the elements of the covariance matrix, one by one, with the following equation

$$[\mathbf{\Sigma}]_{l,k} = \frac{1}{|N(t_{lk})|} \sum_{i \in N(t_{lk})} (y_{il} - \mu(t_l))(y_{ik} - \mu(t_k)) \tag{2.32}$$

where $[\mathbf{\Sigma}]_{l,k}$ is the l, k-th element of the covariance matrix $\mathbf{\Sigma}$, $N(t_{lk})$ is a set of curves for which raw observations are available at time steps t_l and t_k.

Once the covariance matrix is estimated one can compute the fpcs with the equation (2.7). Note that the fpcs estimated in this way might not be smooth. In such situations, one has to employ curve smoothing (basis expansion) on the estimates of the fpcs.

Once the mean and the functional principal components are estimated from all available data one can proceed to smooth all functions in the ensemble with the curve completion procedure we outlined previously. That is, expand the mean and the fpcs onto raw functional data with equation (2.30).

Chapter 3

Forecasting of Spatially Correlated Functional Data[1]

Spatially correlated functional data is recorded in various fields of science. The most rudimentary example of spatially correlated functional data are daily temperature measurements curves presented in chapter 2. In petroleum engineering, an example of spatially correlated curves are unconventional reservoir production curves. In this chapter, we will review modern techniques for interpolation and interpretation of spatially correlated functional data. In particular, we will review universal trace kriging (Menafoglio et al. [2013]), a recently introduced method for direct interpolation of functional data. We will also review a method by Nerini et al. [2010] for interpolation of stationary functional data and propose and extension for non-stationary case. The reviewed methods are demonstrated and compared on a real unconventional Barnett shale case study with 832 horizontal wells.

3.1 Universal Trace Kriging (UTrK)

Menafoglio et al. [2013] developed UTrK methodology concurrently with Caballero et al. [2013]. The method relies on the definitions of trace variance and trace covariance that we present next.

[1]Research presented in this chapter was conducted in collaboration with Alessandra Menafoglio and it was recently published in Menafoglio et al. [2016b]

3.1.1 Trace Variance

Trace variance was defined by Menafoglio et al. [2013] as follows[2]:

$$\begin{aligned} \mathrm{Var}_t(\boldsymbol{\mathcal{X}}) &= \mathbb{E}\left[\|\boldsymbol{\mathcal{X}} - \mu\|^2\right] = \mathbb{E}\left[\|\boldsymbol{\mathcal{X}}\|^2\right] - \|\mathbb{E}\left[\boldsymbol{\mathcal{X}}\right]\|^2 \\ &= \mathbb{E}\left[\int_T (\boldsymbol{\mathcal{X}}(t) - \mu(t))^2 dt\right] \end{aligned} \tag{3.1}$$

This equation is an infinite dimensional analogue of the total variance, a measure of overall dispersion commonly used in multivariate statistics. Consider a vector $\mathbf{X}$ of length n and its variance covariance matrix $\boldsymbol{\Sigma}$ of size $n \times n$. In multivariate statistics, the total variance is defined as the trace of the variance covariance matrix $\boldsymbol{\Sigma}$

$$trace(\boldsymbol{\Sigma}) = \mathbb{E}[(x_1 - \mu_1)^2] + \mathbb{E}[(x_2 - \mu_2)^2] + ... + \mathbb{E}[(x_n - \mu_n)^2] \tag{3.2}$$

This functional is equal to the expected value of the inner product of $\boldsymbol{X} - \boldsymbol{\mu}$ with itself.

$$\mathbb{E}\left[\langle \boldsymbol{X} - \boldsymbol{\mu}, \boldsymbol{X} - \boldsymbol{\mu} \rangle\right] = \sum_{i=1}^{N} \mathbb{E}\left[(x_i - \mu_i)^2\right] \tag{3.3}$$

3.1.2 Trace Covariance

In a similar manner one can define trace covariance between two random functions

$$\begin{aligned} \mathrm{Cov}_t(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) &= \mathbb{E}\left[\langle \boldsymbol{\mathcal{X}} - \mu_{\mathcal{X}}, \boldsymbol{\mathcal{Y}} - \mu_{\mathcal{Y}} \rangle\right] \\ &= \mathbb{E}\left[\int_T \left(\boldsymbol{\mathcal{X}}(t) - \mu_{\mathcal{X}}(t)\right)\left(\boldsymbol{\mathcal{Y}}(t) - \mu_{\mathcal{Y}}(t)\right) dt\right] \end{aligned} \tag{3.4}$$

This functional is also an infinite-dimensional analogue of the total covariance between two vectors of the same length.

With the definitions of trace variance and trace covariance for functional data Menafoglio et al. [2013] developed universal trace kriging that we review next. The developments presented below closely follow the developments presented in Menafoglio et al. [2013].

[2]The proof of the first line in equation (3.1) can be found in Hsing and Eubank [2015] page 179.

Starting from a set of functions $\{\mathcal{X}_i(t), t \in T\}_{i=1}^N$, that take value in $L^2(T)$ space, observed over a set of spatial locations $\boldsymbol{s}_i \in D \subset R^2$, Menafoglio et al. [2013] seek to predict an unobserved function $\mathcal{X}_0(t)$ at some location $\boldsymbol{s}_0$ as a linear combination of all observed functions

$$\mathcal{X}_0(t) = \sum_{i=1}^{N} \lambda_i \mathcal{X}_i(t) \tag{3.5}$$

The data generating process is assumed non-stationary in D and further decomposed into deterministic drift and second order stationary and spatially correlated functional residual

$$\mathcal{X}_i(t) = m_i(t) + r_i(t) \tag{3.6}$$

The drift term is modeled with the functional regression (chapter 2)

$$m_i(t) = \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) a_l(t) \tag{3.7}$$

Trace covariances between the residuals are assumed to depend on distance in D, for example: $\mathrm{Cov}_t(r_i(t), r_j(t)) = C(\boldsymbol{s}_i, \boldsymbol{s}_j)$.

The weights $\lambda_1, \lambda_2, ..., \lambda_N$ are found by minimizing the mean squared error under unbiasedness constraint

$$\underset{\lambda_1,\lambda_2,...,\lambda_N}{\operatorname{argmin}} \left(\mathbb{E}\left[\|\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t)\|^2 \right] = \mathrm{Var}_t\left[\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t) \right] - \left\| \mathbb{E}\left[\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t) \right] \right\|^2 \right) \tag{3.8}$$

The unbiasedness constraints are developed from the second term on the right of

equation (3.8) as follows:

$$\begin{aligned}\mathbb{E}\left[\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t)\right] &= \mathbb{E}\left[\sum_{i=1}^{N} \lambda_i \mathcal{X}_i(t)\right] + \mathbb{E}\left[\mathcal{X}_0(t)\right] \\ &= \sum_{i=1}^{N} \lambda_i \mathbb{E}\left[\mathcal{X}_i(t)\right] - \mathbb{E}\left[\mathcal{X}_0(t)\right] \\ &= \sum_{i=1}^{N} \lambda_i \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) a_l(t) - \sum_{l=0}^{L} f_l(\boldsymbol{s}_0) a_l(t) \\ &= \sum_{l=0}^{L} a_l(t) \left(\sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right)\end{aligned}$$

Obviously, this quantity will be equal to zero if and only if

$$\sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0) \tag{3.9}$$

Therefore, to find the optimal weights for the best linear unbiased prediction one needs to solve the following constrained optimization problem:

$$\underset{\lambda_1,\lambda_2,\ldots,\lambda_N}{\operatorname{argmin}} \operatorname{Var}_t\left(\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t)\right) \quad s.t. \quad \sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0) \tag{3.10}$$

Analogously to universal kriging of scalars, the constrained optimization problem (3.10) is solved by introducing $L+1$ Lagrange multipliers $(\eta_0, \eta_1, ..., \eta_L)$ that leads to the following objective functional

$$\Phi = \operatorname{Var}_t\left(\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t)\right) + 2\sum_{l=0}^{L} \eta_l \left(\sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) \tag{3.11}$$

Which is further transformed into a more convenient form as follows

$$\begin{aligned}\Phi &= \mathrm{Var}_t\left(\sum_{i=1}^{N}\lambda_i\mathcal{X}_i(t)\right) + \mathrm{Var}_t\left(\mathcal{X}_0(t)\right) - 2\mathrm{Cov}_t\left(\sum_{i=1}^{N}\lambda_i\mathcal{X}_i(t),\mathcal{X}_0(t)\right) + \\ &\qquad 2\sum_{l=0}^{L}\eta_l\left(\sum_{i=1}^{N}\lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\mathrm{Cov}_t(\mathcal{X}_i(t),\mathcal{X}_j(t)) + \mathrm{Var}_t\left(\mathcal{X}_0(t)\right) - 2\sum_{i=1}^{N}\lambda_i\mathrm{Cov}_t\left(\mathcal{X}_i(t),\mathcal{X}_0(t)\right) + \\ &\qquad 2\sum_{l=0}^{L}\eta_l\left(\sum_{i=1}^{N}\lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j C(\boldsymbol{s}_i,\boldsymbol{s}_j) + \mathrm{Var}_t\left(\mathcal{X}_0(t)\right) - 2\sum_{i=1}^{N}\lambda_i C\left(\boldsymbol{s}_i,\boldsymbol{s}_0\right) + \\ &\qquad 2\sum_{l=0}^{L}\eta_l\left(\sum_{i=1}^{N}\lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right)\end{aligned} \tag{3.12}$$

To find the weights λ_i that minimize (3.12), its partial derivatives with respect to λ_i are set to zero:

$$\begin{aligned}\frac{\partial\Phi}{\partial\lambda_i} &= 2\sum_{j=1}^{N}\lambda_j C(\boldsymbol{s}_i,\boldsymbol{s}_j) - 2\mathrm{Var}_t\left(\mathcal{X}(t)\right)) + 2\sum_{l=0}^{L}\eta_l f_l(\boldsymbol{s}_i) = 0 \quad i = 1,2,...,N \\ \frac{\partial\Phi}{\partial\eta_l} &= 2\left[\sum_{i=1}^{N}\lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right] = 0 \quad l = 0,1,...,L\end{aligned} \tag{3.13}$$

Which, after rearranging, leads to the following system of linear equations

$$\begin{aligned}\sum_{j=1}^{N}\lambda_j C(\boldsymbol{s}_i,\boldsymbol{s}_j) + \sum_{l=0}^{L}\eta_l f_l(\boldsymbol{s}_i) &= C(\boldsymbol{s}_i,\boldsymbol{s}_0) \quad i = 1,2,...,N \\ \sum_{i=1}^{N}\lambda_i f_l(\boldsymbol{s}_i) &= f_0(\boldsymbol{s}_0) \quad l = 0,1,...,L\end{aligned} \tag{3.14}$$

Note that this system of equations is analogous to the system of universal kriging equations presented in Chiles and Delfiner [1999].

The trace kriging variance is given by

$$\sigma^2_{UTrK}(\boldsymbol{s}_0) = C(0) - \sum_{i=1}^{N} \lambda_i C(\boldsymbol{s}_i, \boldsymbol{s}_0) - \sum_{l=0}^{L} \eta_l f_l(\boldsymbol{s}_0) \tag{3.15}$$

System (3.14) can be expressed in matrix form as follows:

$$\begin{bmatrix} C(\boldsymbol{s}_1,\boldsymbol{s}_1) & C(\boldsymbol{s}_1,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_1,\boldsymbol{s}_N) & f_0(\boldsymbol{s}_1) & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ C(\boldsymbol{s}_2,\boldsymbol{s}_1) & C(\boldsymbol{s}_2,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_2,\boldsymbol{s}_N) & f_0(\boldsymbol{s}_2) & f_1(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\boldsymbol{s}_N,\boldsymbol{s}_1) & C(\boldsymbol{s}_N,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_N,\boldsymbol{s}_N) & f_0(\boldsymbol{s}_N) & f_1(\boldsymbol{s}_N) & \cdots & f_L(\boldsymbol{s}_N) \\ f_0(\boldsymbol{s}_1) & f_0(\boldsymbol{s}_2) & \cdots & f_0(\boldsymbol{s}_N) & 0 & 0 & \cdots & 0 \\ f_1(\boldsymbol{s}_1) & f_1(\boldsymbol{s}_2) & \cdots & f_1(\boldsymbol{s}_N) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_L(\boldsymbol{s}_1) & f_L(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_N) & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \eta_0 \\ \eta_1 \\ \vdots \\ \eta_L \end{bmatrix} = \begin{bmatrix} C(\boldsymbol{s}_1,\boldsymbol{s}_0) \\ C(\boldsymbol{s}_2,\boldsymbol{s}_0) \\ \vdots \\ C(\boldsymbol{s}_N,\boldsymbol{s}_0) \\ f_0(\boldsymbol{s}_0) \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{bmatrix} \tag{3.16}$$

The same developments apply in the case of constant functional mean that ultimately leads to the ordinary trace kriging system of equations introduced by Giraldo [2009][3].

3.1.3 Parameter Estimation

In universal trace kriging, parameter inference is accomplished with the method of moments. Currently, this is the only applicable parameter inference approach since the concept of density for functional data is not well defined (Delaigle and Hall [2010]). Depending on the modeler's preference, one can choose to work with trace covariances or trace variograms, both were defined by Menafoglio et al. [2014], Giraldo [2009] as follows

Trace covariogram estimator:

$$g_{TR}(h) = \frac{1}{|N(h)|} \sum_{(ij)\in N(h)} \int_T r_i(t) r_j(t) dt \tag{3.17}$$

[3]OTrK is a special case of universal trace kriging obtained for $L = 0$ and $f_0(\boldsymbol{s}_j) = 1, \forall j$.

Trace variogram estimator:

$$\gamma_{TR}(h) = \frac{1}{2|N(h)|} \sum_{(ij)\in N(h)} \int_T (r_i(t) - r_j(t))^2 \, dt \tag{3.18}$$

where $N(h)$ denotes the set of pairs (i, j) such that $h - \Delta h \leq \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq h + \Delta h$.

The properties of the trace variogram are the same as the properties of conventional variogram. It is a negative definite function of spatial distance.

The relationship between trace variogram and trace covariogram is analogous to the relationship between scalar variogram and scalar covariogram:

$$\gamma_{TR}(h) = g_{TR}(0) - g_{TR}(h) \tag{3.19}$$

Universal trace kriging modeling work-flow per Menafoglio et al. [2013] is as follows:

1. Use piece-wise OLS regression to infer a functional regression model and compute functional residuals
2. Compute the empirical trace variogram on the functional residuals for a predefined set of lags with equation (3.18)
3. Fit one of admissible variogram models (Gau, Sph, Matern, etc.) and compute the covariance matrix
4. Fit a piece-wise GLS regression on the raw functional data with the covariance matrix from the previous step and compute functional residuals.
5. Iterate steps 2-4 a few times.
6. Use the final fitted variogram model to forecast new functions at their respective locations.

3.2 Projection Based Approaches for Spatial Interpolation of Functions

An alternative approach to spatial interpolation of functions relies on the expansion (or projection) of functional data onto some sort of basis (i.e. B-splines) and then proceeds to spatially interpolate the coefficients of basis expansion. One of the first approaches of this kind was proposed by Nerini et al. [2010] who used B-spline basis to expand functional data and then employed ordinary co-kriging to spatially interpolate the coefficients of basis expansion. In this section, we will review the method by Nerini et al. [2010] and proceed to develop an extension for non-stationary functional data.

3.2.1 Ordinary co-Kriging of Basis Coefficients of Spatially Correlated Functional Data

Let $\{\mathcal{X}_i(t), t \in T\}_{i=1}^N$ be a set of functions observed over a set of spatial locations $\boldsymbol{s}_i \in D \subset R^2$, let $\boldsymbol{c}_i = \{c_{i1}, c_{i2}, ..., c_{iK}\}$ be a vector of basis expansion coefficients of i-th function, and let's assume that functions have a constant mean $\mu(t)$ in D whose basis expansion coefficients are $c_{\mu 1}, c_{\mu 2}, ..., c_{\mu_K}$. Nerini et al. [2010] proposed to forecast an unobserved function $\mathcal{X}_0(t)$ at location $\boldsymbol{s}_0$ by forecasting its basis expansion coefficients $\{c_{01}^*, c_{02}^*, ..., c_{0k}^*\}$ with the best linear unbiased combination of the basis coefficients of all already observed curves

$$c_{0j}^* = \sum_{i=1}^{N} \lambda_{ij} c_{ij} + \sum_{i=1}^{N} \sum_{k \neq j}^{K} \lambda_{ik} c_{ik} \tag{3.20}$$

While assuming that covariances between the coefficients are a function of spatial distance $Cov(c_{il}, c_{jp}) = C_{lp}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$.

The weights in (3.20) are found by solving the following constrained optimization problem

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \quad \mathbb{E}\left[\left(c_{0j}^* - c_{0j}\right)^2\right] \quad s.t. \quad \mathbb{E}\left[c_{0j}^* - c_{0j}\right] = 0 \tag{3.21}$$

The unbiasedness constraints are developed from the second term in (3.21) as

follows:

$$\begin{aligned}
\mathbb{E}\left[c_{0j}^{*} - c_{0j}\right] &= \mathbb{E}\left[\sum_{i=1}^{N}\lambda_{ij}c_{ij} + \sum_{i=1}^{N}\sum_{k\neq j}^{K}\lambda_{ik}c_{ik}\right] - \mathbb{E}\left[c_{0j}\right] \\
&= \sum_{i=1}^{N}\lambda_{ij}\mathbb{E}\left[c_{ij}\right] + \sum_{i=1}^{N}\sum_{k\neq j}^{K}\lambda_{ik}\mathbb{E}\left[c_{ik}\right] - c_{\mu j} \\
&= c_{\mu j}\sum_{i=1}^{N}\lambda_{ij} + \sum_{i=1}^{N}\sum_{k\neq j}^{K}\lambda_{ik}c_{\mu k} - c_{\mu j} \\
&= c_{\mu j}\left(\sum_{i=1}^{N}\lambda_{ij} - 1\right) + \sum_{k\neq j}^{K}c_{\mu k}\sum_{i=1}^{N}\lambda_{ik}
\end{aligned} \tag{3.22}$$

Obviously, this functional will be equal to zero if and only if

$$\sum_{i=1}^{N}\lambda_{ij} = 1 \quad and \quad \sum_{i=1}^{N}\lambda_{ik} = 0 \quad \forall k \neq j \tag{3.23}$$

Under these unbiasedness constraints, minimizing (3.21) amounts to minimizing the variance of the prediction

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \quad \mathbf{Var}\left[c_{0j}^{*} - c_{0j}\right] \quad s.t. \quad \sum_{i=1}^{N}\lambda_{ij} = 1 \quad and \quad \sum_{i=1}^{N}\lambda_{ik} = 0 \quad \forall k \neq j \tag{3.24}$$

This constrained optimization problem is solved by first developing the first term in

(3.24) and introducing K Lagrange multipliers $\eta_1, \eta_2, ..., \eta_K$

$$\begin{aligned}\Phi(\lambda) = C_{kk}(0) + \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{k'=1}^{K}\sum_{i'=1}^{N}\lambda_{ik}\lambda_{i'k'}C_{k'k}(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) - \\ 2\sum_{k=1}^{K}\sum_{i=1}^{N}\lambda_{ik}C_{kk}(\|\boldsymbol{s}_i - \boldsymbol{s}_0\|) + \\ 2\eta_j\left(\sum_{i=1}^{N}\lambda_{kj} - 1\right) + \\ 2\sum_{\substack{k=1\\k\neq j}}^{K}\eta_k\left(\sum_{i=1}^{N}\lambda_{ik}\right)\end{aligned} \tag{3.25}$$

The solution is found by setting the partial derivatives of (3.25) with respect to the weights to zero. This ultimately leads to the following system of linear equations

$$\begin{aligned}\sum_{k=1}^{K}\sum_{i=1}^{N}\lambda_{ik}C_{k'k}(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) + \eta_{k'} = C_{k'k}(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|), \\ (k' = 1, ..., K; i' = 1, ..., N;); \\ \sum_{i=1}^{N}\lambda_{i1} = 1; \\ \sum_{i=1}^{N}\lambda_{ik} = 0, \quad \forall k \neq 1.\end{aligned} \tag{3.26}$$

This system can be expressed in matrix form as follows:

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1K} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2K} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K1} & \boldsymbol{C}_{K2} & \cdots & \boldsymbol{C}_{KK} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \\ \mathbf{1}^T & \mathbf{0}^T & \cdots & \mathbf{0}^T & 0 & 0 & \cdots & 0 \\ \mathbf{0}^T & \mathbf{1}^T & \cdots & \mathbf{0}^T & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \mathbf{1}^T & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda_1} \\ \boldsymbol{\lambda_2} \\ \vdots \\ \boldsymbol{\lambda_K} \\ \eta_1 \\ \eta_2 \\ \vdots \\ \eta_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{10} \\ \boldsymbol{c}_{20} \\ \vdots \\ \boldsymbol{c}_{K0} \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.27}$$

Where: $\boldsymbol{C}_{pl}$ is a matrix whose (i,m)-th element equals $Cov(c_{ip}, c_{ml}) = C_{pl}(\|\boldsymbol{s}_i - \boldsymbol{s}_m\|)$, $\mathbf{1}$ and $\mathbf{0}$ are a vectors of ones and zeros of length N respectively, $\boldsymbol{\lambda}_k$ is a vector of weights associated with k-th basis coefficients, and $\boldsymbol{c}_{k0}$ is a vector whose p-th element is $Cov(c_{0j}, c_{pk}) = C_{jk}(\|\boldsymbol{s}_0 - \boldsymbol{s}_p\|)$.

In order to forecast one function, one needs to solve K such systems of equations and then simply construct $\mathcal{X}_0^*(t)$ from the forecasted coefficients and the common basis system ($\mathcal{X}_0^*(t) = \sum_{j=1}^{K} c_{0j}^* \phi_j(t)$). The dimensionality of the system (3.27) depends on the number of basis functions used in data smoothing and the number of observed functions. This dimensionality can be significantly reduced by projecting the functional data onto a low dimensional truncated basis of functional principal components[4] and then forecasting the principal component scores instead of forecasting the coefficients of basis expansion. Another advantage of using functional principal components is that in many cases there is a complete absence of spatial cross-correlation (nugget) between functional principal component scores. In such situations, the entire problem of forecasting unobserved functions boils down to ordinary kriging of principal component scores (autokrigeability problem in Wackernagel [2010]). It is a good modeling practice to always investigate for spatial cross correlations between the fpc scores since it is not a general rule that they are equal to zero (see Wackernagel [2010] for a detailed discussion).

3.2.2 Universal co-Kriging of Basis Coefficients of Spatially Correlated Functional Data[5]

Let $\{\mathcal{X}_i(t), t \in T\}_{i=1}^{N}$ be a set of functional data (i.e. oil production curves) observed over a set of spatial locations $\boldsymbol{s}_i, \boldsymbol{s} \in D \subset R^2$, let $b_k : \{\phi_1(t), \phi_2(t), ..., \phi_K(t)\}$ be a basis system consisting of K basis functions and let $\boldsymbol{c}_i = \{c_{i1}, c_{i2}, ..., c_{ik}\}$ be a vector of basis coefficients of i-th function. Here, we assume that the data generating process is non-stationary on D and that it can be decomposed into deterministic drift and

[4] i.e. by using the FVE criterion to select the number of leading fpcs.

[5] The developments presented in this section were published in a slightly different form in Menafoglio et al. [2016b].

second order stationary spatially correlated residual.

$$\mathcal{X}_i(t) = m_i(t) + r_i(t) \tag{3.28}$$

Where the drift term was previously modeled (UTrK) with functional regression $\sum_{l=0}^{L} f_l(\boldsymbol{s}_i)a_l(t)$. The function, the residual and the functional drift coefficients can all be expressed in terms of a common basis system

$$\sum_{k=1}^{K} c_{ik}\phi_k(t) = \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) \sum_{k=1}^{K} b_l^k \phi_k(t) + \sum_{k=1}^{K} c_{ik}^r \phi_k(t) \tag{3.29}$$

Where c_{ik} is the k-th basis coefficient of the i-th function, $f_l(.)$ is a transformation function operating on the vector of spatial location, b_l^k is the k-th basis coefficient of the l-th functional coefficient $(a_l(t))$, and c_{ik}^r is the k-th basis coefficient of the residual of i-th function. The first term on the right side of equation (3.29) can be rearranged as follows

$$\sum_{k=1}^{K} c_{ik}\phi_k(t) = \sum_{k=1}^{K} \phi_k(t) \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) b_l^k + \sum_{k=1}^{K} c_{ik}^r \phi_k(t) \tag{3.30}$$

We can further drop the basis functions from the equation to arrive to the following expression:

$$\begin{bmatrix} c_{i1} \\ c_{i2} \\ \vdots \\ c_{iK} \end{bmatrix} = \begin{bmatrix} \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) b_1^l \\ \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) b_2^l \\ \vdots \\ \sum_{l=0}^{L} f_l(\boldsymbol{s}_i) b_k^l \end{bmatrix} + \begin{bmatrix} c_{i1}^r \\ c_{i2}^r \\ \vdots \\ c_{iK}^r \end{bmatrix} \tag{3.31}$$

This expression implies that non-stationarity in functional data translates into non-stationarity in basis coefficients. This further implies that interpolation of non-stationary functions can be achieved by means of interpolation of non-stationary basis coefficients with universal co-kriging (Chiles and Delfiner [1999]). The developments presented below follow the developments presented in Chiles and Delfiner [1999] pg. 305, for universal co-kriging with algebraically independent drifts.

The objective is to predict function $\mathcal{X}_0(t)$ at location $\boldsymbol{s}_0$ by predicting the coefficients of its basis expansion with a linear combination of basis expansion coefficients of all already observed functions

$$c_{0j}^* = \sum_{i=1}^{N} \lambda_{ij} \xi_{ij} + \sum_{\substack{k=1 \\ k \neq j}}^{K} \sum_{i=1}^{N} \lambda_{ik} c_{ik} \tag{3.32}$$

As before the weights are sought by minimizing the mean squared error in predictions under unbiasedness constraints

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \quad \mathbb{E}\left[\left(c_{0j}^* - c_{0j}\right)^2\right] \quad s.t. \quad \mathbb{E}\left[c_{0j}^* - c_{0j}\right] = 0 \tag{3.33}$$

The unbiasedness constraints are developed from the second term in (3.33) as follows:

$$\begin{aligned}
\mathbb{E}\left[c_{0j}^* - c_{0j}\right] &= \mathbb{E}\left[\sum_{i=1}^{N} \lambda_{ij} c_{ij} + \sum_{i=1}^{N} \sum_{k \neq j}^{K} \lambda_{ik} c_{ik}\right] - \mathbb{E}\left[c_{0j}\right] \\
&= \sum_{i=1}^{N} \lambda_{ij} \mathbb{E}\left[c_{ij}\right] + \sum_{i=1}^{N} \sum_{k \neq j}^{K} \lambda_{ik} \mathbb{E}\left[c_{ik}\right] - \mathbb{E}\left[c_{0j}\right] \\
&= \sum_{i=1}^{N} \lambda_{ij} \sum_{l=0}^{L} b_l^j f_l(\boldsymbol{s}_i) + \sum_{i=1}^{N} \sum_{k \neq j}^{K} \lambda_{ik} \sum_{l=0}^{L} b_l^k f_l(\boldsymbol{s}_i) - \sum_{l=0}^{L} b_l^j f_l(\boldsymbol{s}_0) \\
&= \sum_{l=0}^{L} b_l^j \left(\sum_{i=1}^{N} \lambda_{ij} f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) + \sum_{k \neq j}^{K} \sum_{l=0}^{L} b_l^k \sum_{i=1}^{N} \lambda_{ik} f_l(\boldsymbol{s}_i)
\end{aligned} \tag{3.34}$$

Equation (3.34) will be equal to zero if and only if

$$\begin{aligned}
&\sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\
&\sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{s}_i) = 0, \quad \text{for} \quad j \neq k, \quad \forall l;
\end{aligned} \tag{3.35}$$

The updated constrained optimization problem is as follows

$$\underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \quad \operatorname{Var}\left[c_{0j}^{*} - c_{0j}\right] \quad s.t. \quad \begin{cases} \sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\ \sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{s}_i) = 0, \quad \text{for} \quad j \neq k, \quad \forall l \end{cases} \tag{3.36}$$

This optimization problem is solved in the same manner as before by developing the variance term and introducing $K \times (L+1)$ Lagrange multipliers

$$\begin{aligned} \Phi(\lambda) = C_{kk}(\mathbf{0}) + \sum_{j=1}^{K}\sum_{i=1}^{N}\sum_{j'=1}^{K}\sum_{i'=1}^{N} \lambda_{ji}\lambda_{j'i'}C_{jj'}(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|) - \\ 2\sum_{j=1}^{K}\sum_{i=1}^{N} \lambda_{ji}C_{jk}(\boldsymbol{s}_i - \boldsymbol{s}_0) + \\ 2\sum_{l=0}^{L} \eta_{kl}\left(\sum_{i=1}^{N} \lambda_{ki} f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) + \\ 2\sum_{l=0}^{L}\sum_{\substack{j=1 \\ j\neq k}}^{K} \eta_{jl}\left(\sum_{i=1}^{N} \lambda_{ji} f_l(\boldsymbol{s}_i)\right) \end{aligned} \tag{3.37}$$

After setting the partial derivatives with respect to weights to zero, and rearranging, we arrive to the following system of equations

$$\begin{aligned} \sum_{j=1}^{K}\sum_{i=1}^{N} \lambda_{ji}C_{jj'}(\boldsymbol{s}_i - \boldsymbol{s}_{i'}) + \sum_{l=0}^{L} \eta_{j'l} f_l(\boldsymbol{s}_i) = C_{j'k}(\boldsymbol{s}_{i'} - \boldsymbol{s}_0), \\ (j' = 1, ..., K; i' = 1, ..., N;); \\ \sum_{i=1}^{N} \lambda_{ki} f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\ \sum_{i=1}^{N} \lambda_{ji} f_l(\boldsymbol{s}_i) = 0, \quad j \neq k, \quad \forall l; \end{aligned} \tag{3.38}$$

That can be expressed in matrix form as follows:

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1K} & \boldsymbol{F} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2K} & \mathbf{0} & \boldsymbol{F} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K1} & \boldsymbol{C}_{K2} & \cdots & \boldsymbol{C}_{KK} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F} \\ \boldsymbol{F}^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{F}^T & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F}^T & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda_1} \\ \boldsymbol{\lambda_2} \\ \vdots \\ \boldsymbol{\lambda_K} \\ \boldsymbol{\eta_1} \\ \boldsymbol{\eta_2} \\ \vdots \\ \boldsymbol{\eta_K} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{10} \\ \boldsymbol{c}_{20} \\ \vdots \\ \boldsymbol{c}_{K0} \\ \mathbf{1} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{3.39}$$

Where:

$$\boldsymbol{F} = \begin{bmatrix} f_0(\boldsymbol{s}_1) & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ f_0(\boldsymbol{s}_2) & f_1(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\boldsymbol{s}_N) & f_1(\boldsymbol{s}_N) & \cdots & f_L(\boldsymbol{s}_N) \end{bmatrix}, \boldsymbol{\lambda}_k = \begin{bmatrix} \lambda_{1k} \\ \lambda_{2k} \\ \vdots \\ \lambda_{Nk} \end{bmatrix}, \boldsymbol{\eta_k} = \begin{bmatrix} \eta_{0k} \\ \eta_{1k} \\ \vdots \\ \eta_{Lk} \end{bmatrix}$$

3.2.3 Parameter Estimation

In both stationary and non-stationary cases outlined before, one needs to perform covariance estimation and smoothing by fitting one of the admissable covariance structures (Sph, Mat, Gau,...). Here we review the auto and the cross-covariograms and variograms that are well known in geostatistics.

Covariogram estimator The very well known covariogram estimator was formulated by Matheron as follows:

$$g_{lp}(h) = \frac{1}{N(h)} \sum_{(ij)\in N(h)} (c_{il} - \mu_l)(c_{jp} - \mu_p) \tag{3.40}$$

where $N(h)$ denotes the set of pairs (i, j) such that $h - \Delta h \leq \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq h + \Delta h$, c_{il}, c_{ip} are l-th and p-th coefficients of the i-th function, while μ_l and μ_p are their respective means.

Variogram estimation. A much more widely used approach for parameter inference is by means of auto and cross variography. Variogram estimator is formulated

as follows

$$\gamma_{lp}(h) = \frac{1}{2N(h)} \sum_{(ij)\in N(h)} (c_{il} - c_{jp})^2 \tag{3.41}$$

For $l = p$ this variogram estimator is called auto-variogram estimator, while in the case of $l \neq p$ it is called *pseudo - cross variogram* estimator (Clark et al. [1987]) for distinction from the conventional cross-variogram estimator formulated by Matheron, and given with the following equation

$$\gamma_{lp}(h) = \frac{1}{2N(h)} \sum_{(ij)\in N(h)} (c_{il} - c_{jl})(c_{ip} - c_{jp}) \tag{3.42}$$

The pseudo cross-variogram is always a positive function of spatial distance while the conventional cross-variogram can be both positive and negative. The motivation behind the pseudo-cross variogram definition is that it can be computed on both isotopic and heterotopic data samples while the cross-variogram can be computed only on isotopic data samples (Wackernagel [2010]). In the case of co-kriging of functional basis coefficients (or fpc scores), both cross-variogram estimators are applicable since the data is isotopic by default. This is due to the fact that all basis expansion coefficients are always observed at all locations.

Modeling procedure consists of selecting one of admissable variogram models (Sph, Gau, Exp,...) that are then fitted with the linear model of coregionalization (LMC, Goovaerts [1997]) to the auto and cross variogram estimates computed with equation 3.41 (and/or eq (3.42)) for a pre-selected number of spatial lags.

3.3 Simulation of Functional Data

Simulation of functional data can be achieved by means of sequential Gaussian co-simulation (CO-SGSIM, Verly [1992]) of the functional principal component scores or basis coefficients (Menafoglio et al. [2016a]). Co-simulation maps can be useful for reservoir data interpretation along with the interpretation of functional principal components (or rotated fpcs). For example, high values in fpc scores might indicate higher well productivity or indirectly inform about a secondary variable that is highly correlated with the fpc scores (i.e. total organic content) and the features in functional data described by the scores. Another advantage of co-simulation is that

it produces a large number of functional forecasts at each location that can be used to construct prediction bands. This feature of CO-SGSIM is highly useful in hydrocarbon production forecasting where uncertainty quantification is one of the main modeling objectives.

3.4 Barnett Shale Case Study

In this section, we apply and compare the previously outlined functional interpolation methods on a real unconventional reservoir dataset. The unconventional reservoir in question is the Barnett shale in Eastern Texas, one of the longest producing shale plays in the world. The dataset was obtained from drilling-info website and it contains information on wells horizontal length and monthly gas production rates dating back to year 2007. In this case study, we analyze gas production from horizontal hydraulically fractured wells that have been in production for more than 5 years. As a part of pre-processing, we discarded production data that preceded the "peak" gas rate of each well, since during the early "pre-peak" periods of production wells mostly produce flow back water from hydraulic fracturing operations. The final dataset contained 922 wells that had horizontal length of about 2000 ft. Due to the absence of hydraulic fracturing information, and given similar horizontal length, in our analysis we assumed that all wells were completed in a similar way with a similar number of hydraulic fractures. Well locations and the location of the studied area are shown in figure (3.1).

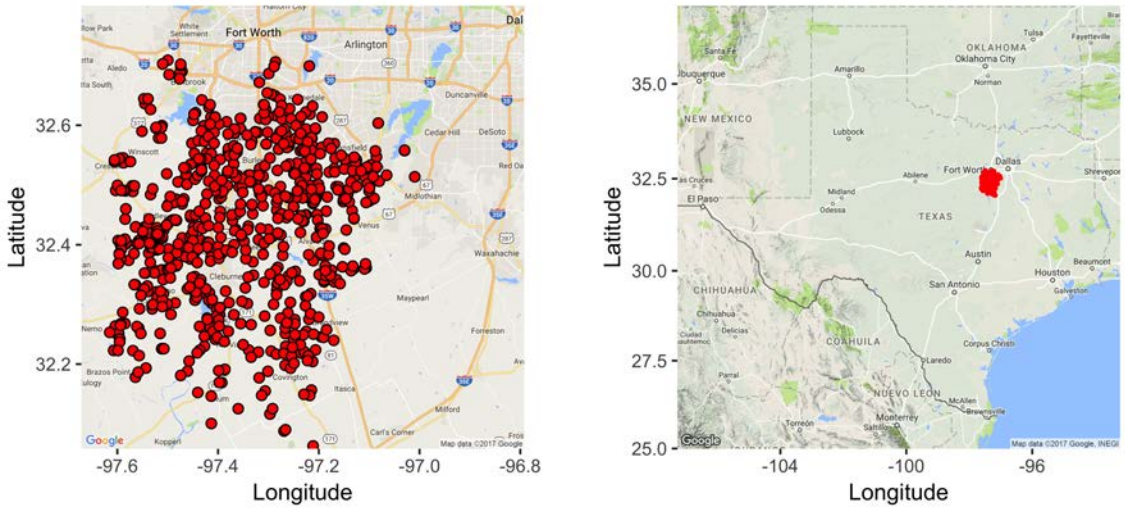


Figure 3.1: *The studied area and well locations*

The first step of functional data analysis is to convert discrete observations of gas production into smooth declining curves. Here we selected a basis system that placed more basis functions at earlier times in order to better "capture" the variations in early production. Basis expansion was performed directly on observed monthly production rates, and it considered only the first 60 post-peak months of production.
Basis expansion is an iterative process. We varied the number of basis functions and the smoothing penalty until visually acceptable smoothly declining fits were produced. Certain wells had erratic behavior where production was not uniformly declining. This erratic behavior was not a consequence of geology or physical processes that take place in the reservoir, but rather a consequence of a change in wells operating conditions[6]. Since no information about the changes in operating conditions was available, basis expansion on such data was very difficult. No amount of smoothing was able to force the fits of such wells to be uniformly declining[7]. In this work, we decided to treat these wells as outliers and we removed them from the study. The final "clean" dataset contained 863 wells. Plots of the basis system and the basis expanded (smooth) functional data are shown in figure 3.2.

[6]For example, a change in wells choke size can increase or decrease wells production.

[7]Alternatively one could consider special smoothing techniques such as monotonically declining fits. These are not analyzed in this dissertation since they are too specialized and as such too difficult to use. Interested readers can consult the documentation of the *fda* R package (Ramsay et al. [2009]

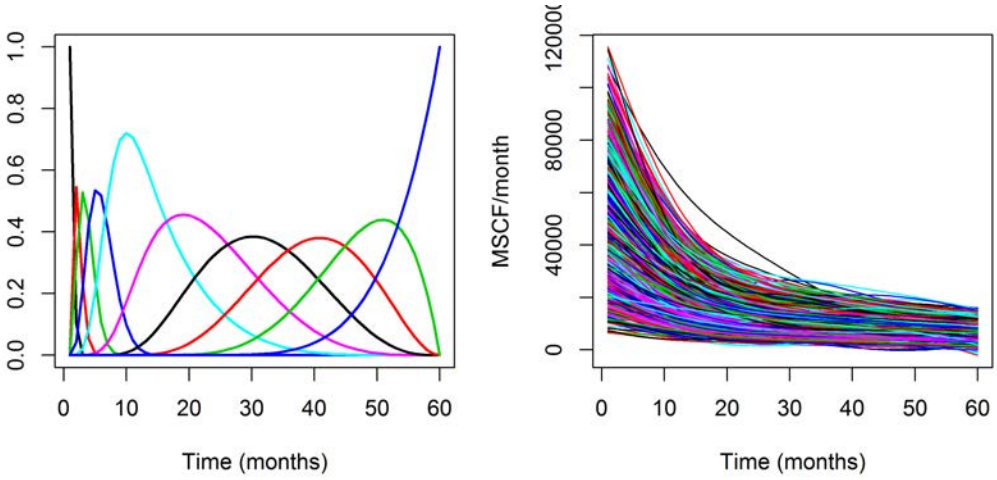


Figure 3.2: *Left - Selected basis system. Right - Basis expanded (smooth) ensemble of curves.*

Smoothed data was first analyzed with universal trace kriging and trace variography. The trace variogram is given in figure 3.3. The model fitted to the empirical variogram is Matern with a range of about 8.6 km. We used this variogram to produce maps of production decline shown in figure 3.4. From these maps we can clearly see that the most prolific area of the reservoir is located in the N-E section with an overall SW-NE trend in reservoir quality. The nugget in figure 3.3 is most likely a consequence of variation in the size of fracturing jobs across the wells (a missing piece of information). The next step in our data analysis was functional principal component analysis (FPCA). FPCA analysis on the entire dataset is given at the top of figure 3.5. We are plotting the first two fpcs since they captured more than 95% of variance in the data. The interpretation of the fpcs is somewhat difficult. We observe that the first fpc describes variation in the overall magnitude of gas production while the second fpc mostly describes variations in later times of production. To improve interpretability of the fpcs, we employed the varimax rotation (see chapter 2). The varimax rotation enabled us to obtain rotated functional principal components that have a higher degree of interpretability (figure 3.5 bottom). From figure 3.5, we observe that the first rotated fpc almost exclusively describes the variation in early gas production, while the second rotated fpc describes production variation in later times. It is well known that in early times unconventional hydrocarbon wells drain existing and artificial fracture networks, while in later times (in the case of gas)

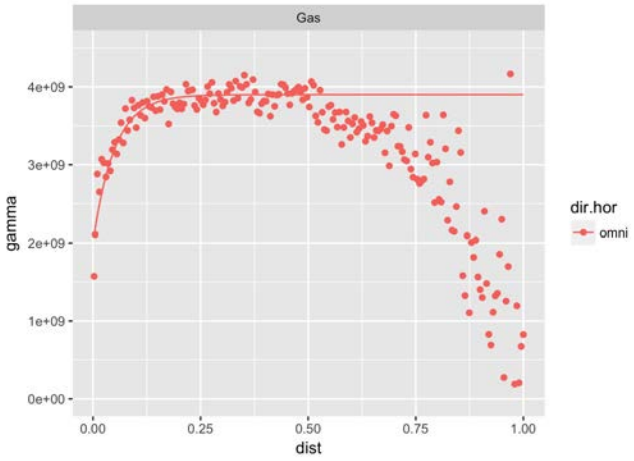


Figure 3.3: *Scaled omni directional trace variogram with a Matern model with a range of 0.12 (8.6km in original scale)*

production is dominated by the amount of sorbed gas in the rock (or reserves). In our case study, we assumed that all wells were completed in the same or similar manner given that they have similar well lengths. In light of this assumption, a high score on the first rotated functional principal component could potentially indicate a high degree of natural fracturing around the well or a high degree of artificial fracturing around the well caused by high brittleness of the rock. In either case, the score on the first rotated fpc describes geological property and as such it is expected to be spatially correlated. The scores of the second rotated fpc are also expected to be highly spatially correlated since they most likely describe the amount of sorbed gas. Both of these hypotheses can be verified by computing the variograms of rotated functional principal component scores. The variograms given in figure 3.6, clearly show the presence of spatial auto and cross correlation in rotated fpc scores. What is interesting to notice is that the cross variogram is always positive[8]. If the assumption of uniform completion holds, in light of the previous interpretation, this would potentially suggest that the brittleness of the rock depends on the amount of sorbed gas (or total organic content). We computed the maps of the first and second rotated fpc that are shown in figure 3.7. These maps are consistent with the universal trace kriging analysis, the North-East portion of the reservoir has high initial production

[8]Cross variogram was computed with conventional formula and as such it can be both positive and negative

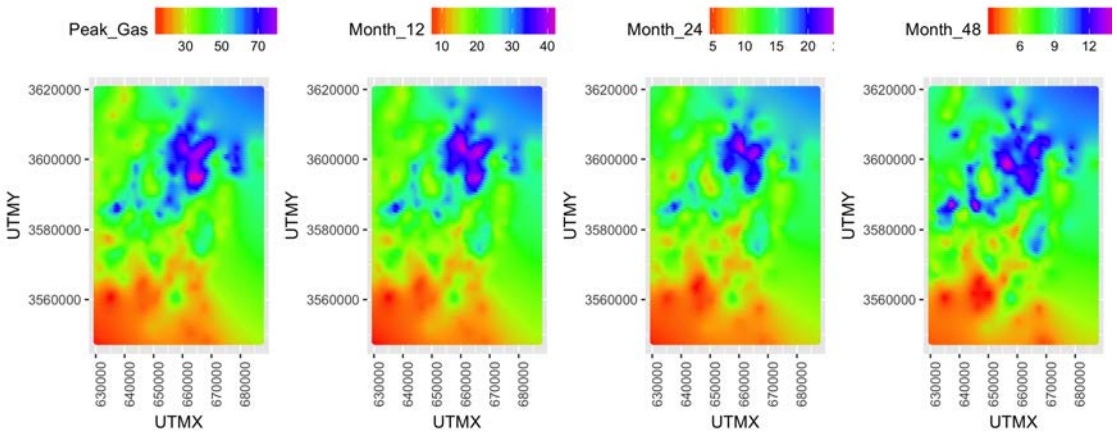


Figure 3.4: *A few maps of gas rate over time produced with universal trace kriging. All maps are in MMSCF*

and high late time production suggesting higher recoverable reserves.

Next we will evaluate the forecasting capabilities of the outlined forecasting methods. We randomly split the dataset into 159 training wells and 704 test wells. We fitted two models on the training data, one universal trace kriging model (UTrK) and one universal co-kriging model on the scores of the varimax rotated fpcs (UCoK.vrmx). The rotated fpcs were re-computed on the sub-setted training data. The fitted variograms had the same range as the variograms used in the previously outlined analysis on the entire dataset, however with different sills due to lower training set size and associated noise.

A few randomly selected forecasts are given in figure 3.8. Visually there is not much difference between the forecasts computed with trace kriging and the forecasts computed with universal co-kriging on rotated fpc scores. To better assess the quality of the forecasts we computed the sum of squared errors (SSE) normalized with the overall trace variance of the data (equation (3.43)). A summary of normalized SSE's is given in table 3.1.

$$nSSE_j = \frac{\|\mathcal{X}_j^*(t) - \mathcal{X}_j(t)\|}{\frac{1}{N}\sum_{i=1}^{N}\|\mathcal{X}_i^*(t) - \mu(t)\|} \tag{3.43}$$

From table 3.1, we can see that the trace-based approach slightly outperformed universal co-kriging of rotated fpcs approach. This is because we were not using all functional principal components but rather the two that captured the most of the

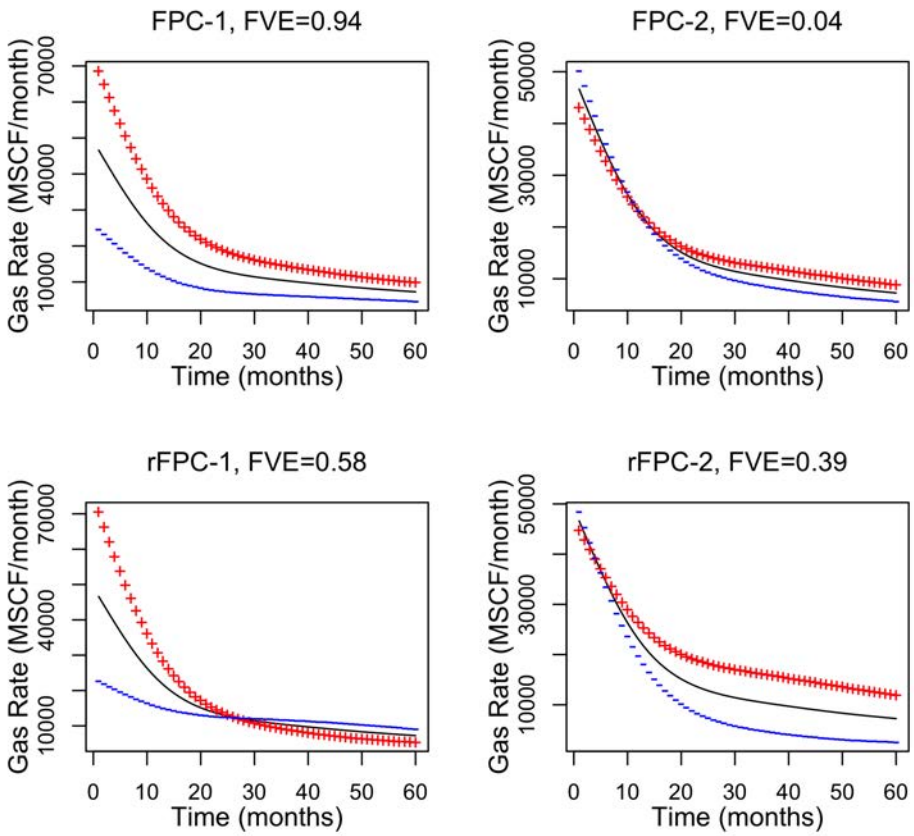


Figure 3.5: *Top - Functional principal components as perturbation about the mean. Bottom - rotated functional principal components as perturbation about the mean*

Table 3.1: *SSE table*

	UTrK	UCoK.vrmx
min	0.0012	0.003
mean	0.697	0.993
median	0.352	0.499
max	7.408	10.271

variance in the data. Therefore, truncation leads to some degree of information loss. To compute prediction bands around our forecasts we employed sequential Gaussian co-simulation on the residuals of the rotated functional principal component scores computed on the training data. A few realizations of co-SGS simulations with added

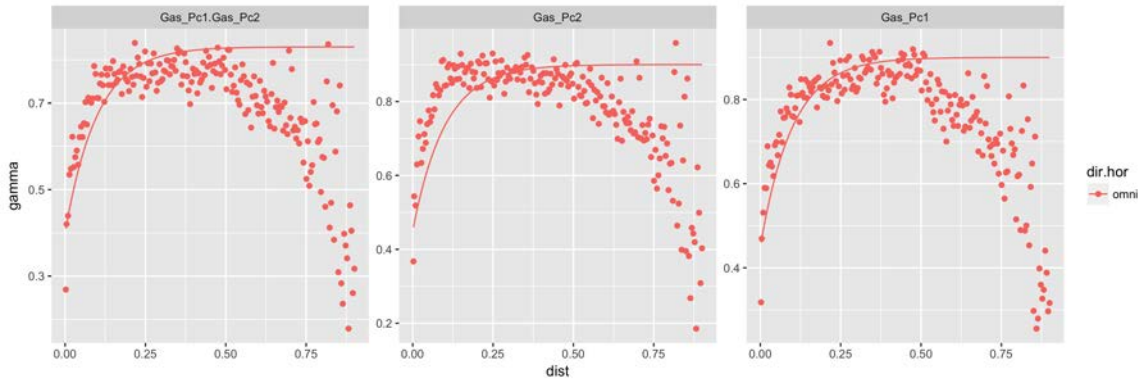


Figure 3.6: *Empirical variograms computed on the rotated fpc scores along with the fits produced with the linear model of coregionalization. The fit is Matern with a range of 0.1 (7.2km in original scale)*

spatial trend are shown in figure 3.9 while the forecasts produced with co-SGS simulations are shown in figure 3.10. We observe that co-SGS forecasts fully enclose the true production data.

3.4.1 Monte Carlo Study

In this section we present a more thorough analysis of the predictive capabilities of the presented forecasting methods. We selected 5 training set sizes, and for each training set size we sampled the dataset 100 times, fitted the two forecasting models and used them to predict the left out wells (test set). On each iteration, we computed the SSE errors between the forecasted and the actual functional data and summarized each test set with the mean and the median of the SSE error. A plot of the mean and the median of each forecasting method for each training set size is given in figure 3.11 while the supporting data is given in table 3.2.

From the results of the Monte Carlo study, we observe that the two methods produce very similar results. However, it is evident that universal trace kriging slightly outperforms the universal co-kriging of fpc scores. As before, this small discrepancy is the result of working with truncated functional principal components instead with all fpc's. It should be noted that universal trace kriging also requires a much smaller modeling effort compared to the universal kriging of fpc scores. We also noticed that the LMC parameter inference becomes increasingly difficult with the increase in the

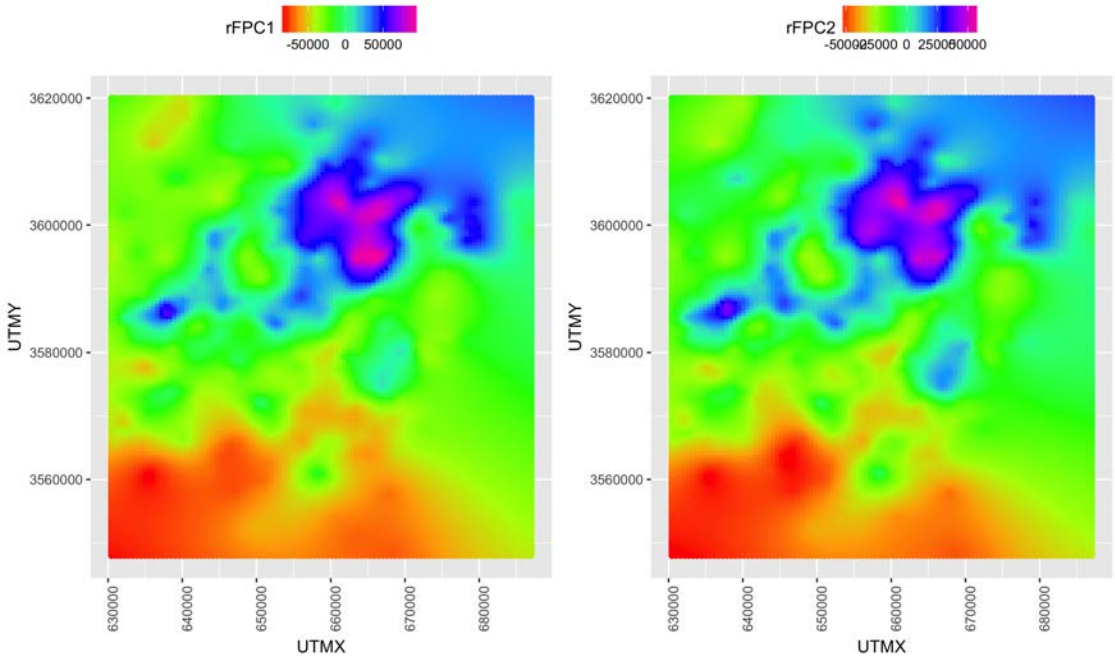


Figure 3.7: *Maps of the first and the second rotated functional principal component*

number of kept fpc's.

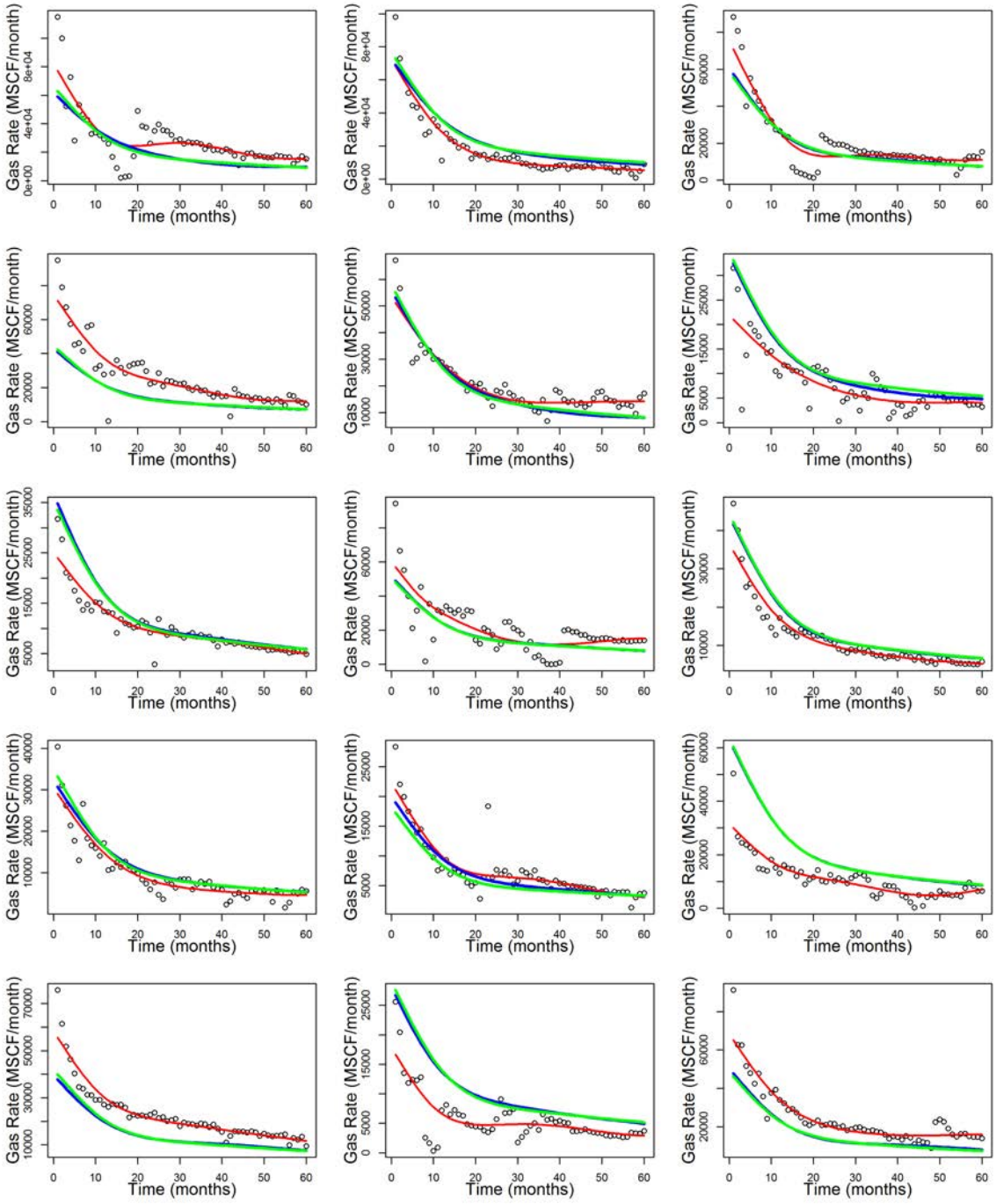


Figure 3.8: *A few forecasts. Black dots - real data, Red line - smoothed real data, Blue line - universal trace kriging forecast, green line - UcoK forecast produced with universal co-kriging on rotated fpc scores*

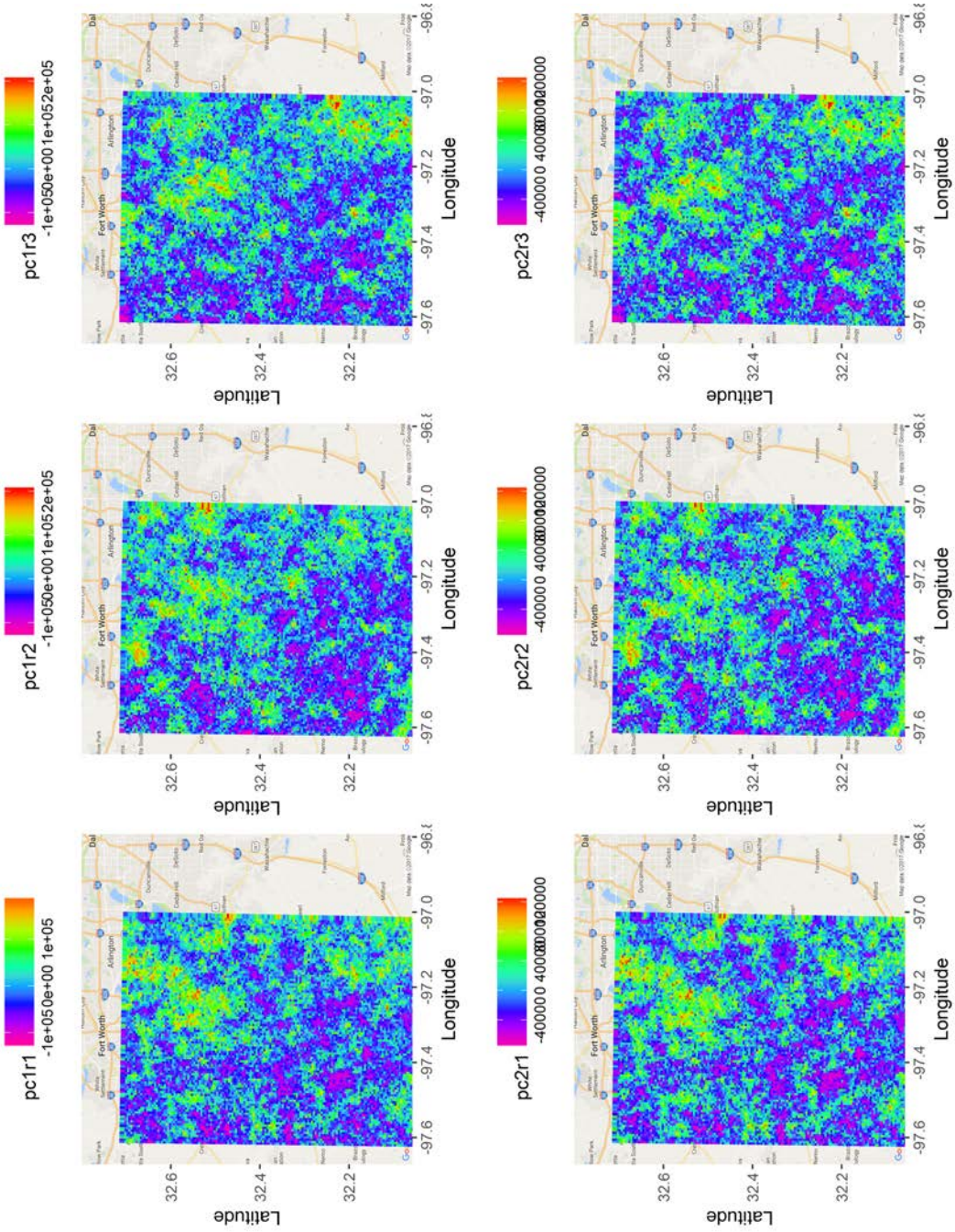


Figure 3.9: *A few realizations of fpc co-simulation on rotated fpc scores*

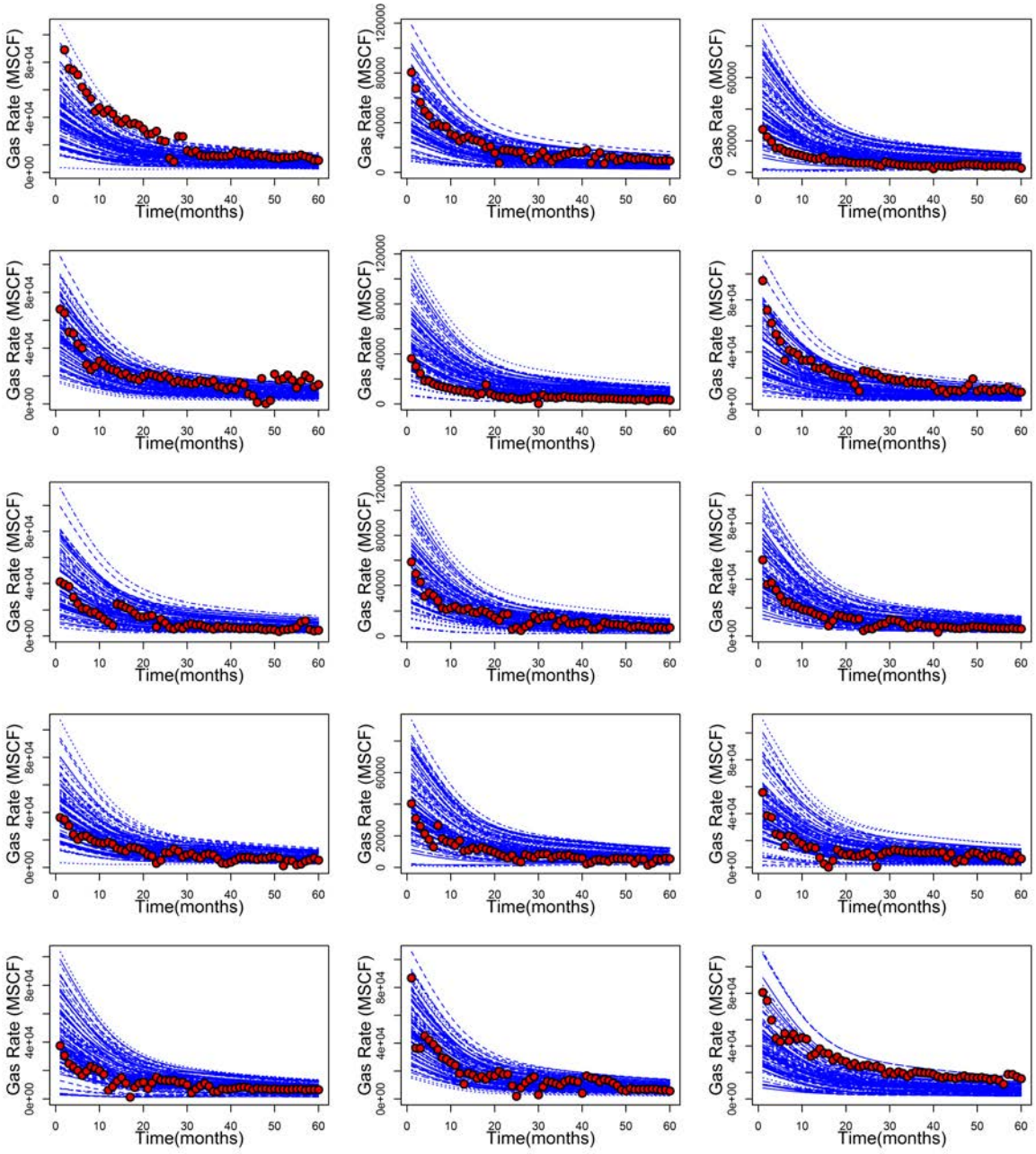


Figure 3.10: *A few randomly selected forecasts produced with co-simulation of rotated fpcs. Red dots represent real production data, blue lines are the forecasts produced with co-simulation of rotated fpc scores.*

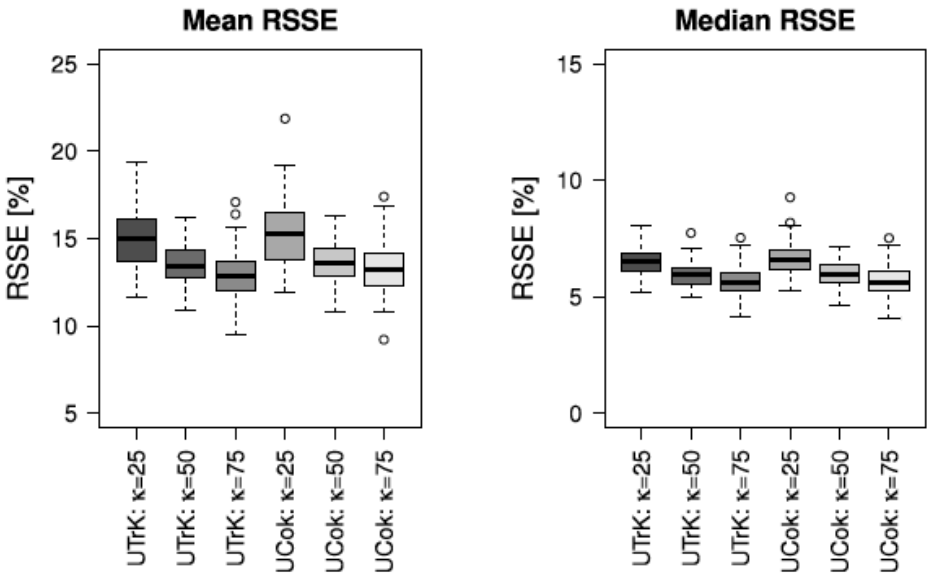


Figure 3.11: *The results of the Monte Carlo Analysis. UTrK = Universal Trace Kriging, UCoK = Cokriging of fpc scores. κ = percentage of the entire dataset used for training. The plot is adapted from Menafoglio et al. [2016b].*

Table 3.2: *The results of the Monte Carlo study*

	Training size	Median	Mean	Std
Mean RSSE(UCok)	$\kappa = 25$	0.153	0.153	0.019
	$\kappa = 50$	0.136	0.137	0.012
	$\kappa = 75$	0.132	0.133	0.014
Median SSE(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 SSE(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 SSE(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

3.5 Chapter Conclusion

In this chapter, we reviewed the principles of geostatistics for functional data. Two methods, universal trace kriging (UTrK) by Menafoglio et al. [2013] and ordinary co-kriging of basis coefficients by Nerini et al. [2010], were reviewed. We demonstrated that the latter method can be easily extended with universal co-kriging to accommodate for non-stationary functional data and serve as an alternative to UTrK methodology. The new extension and the UTrK methodologies were compared and evaluated on a real reservoir case study of the Barnett shale, an unconventional gas reservoir. While intrinsically different the two methods were found to have a similar performance over many test sets analyzed in our Monte Carlo study. UTrK was found to require a slightly lower modeling effort than the method based on universal co-kriging of basis coefficients (or fpc scores). However, in combination with the varimax rotated functional principal components and sequential Gaussian co-simulation the universal co-kriging of fpc scores was found to have a much greater interpretative and practical forecasting power.

Chapter 4

Forecasting of Spatially Correlated Functional Data in Presence of Non-Spatial Covariates

4.1 Introduction

In many fields of Earth sciences, it is quite common to observe spatially correlated functional data along with a certain number of explanatory variables or covariates. One example of such data are unconventional reservoir hydrocarbon production curves whose shape and magnitude depends on the location of the well within the reservoir (geology) and hydraulic fracturing parameters that are a consequence of human activity at that particular well. From a statistical perspective, the analysis and forecasting of such data is difficult since one needs to jointly model and analyze the influence of the location as well as the influence of explanatory variables[1]. The previously presented methods for interpolation of functions are still applicable (with slight modifications) to this problem. In this chapter, we will present detailed theoretical derivations of the modifications of the two methods and demonstrate and compare them on a real unconventional reservoir dataset provided by Anadarko Petroleum Corporation (APC). The dataset consists of 188 horizontal, hydraulically fractured horizontal wells that

[1]The Barnett shale case study we presented in the previous chapter was a special case where all wells had similar hydraulic fracturing parameters hence we modeled their dependence on spatial location.

produce oil, gas and flowback water (more on that later). Additionally, we will also present practical solutions to problems that arise when working with such real dataset.

4.2 Methodology

Here, we will consider a set of smooth functions $\{\mathcal{X}_i(t), t \in T\}_{i=1}^{N}$ (i.e. oil production curves) observed over a set of spatial locations $\boldsymbol{s}_i \in R^2$ along with a certain number of explanatory variables or covariates $\boldsymbol{z}_i \in R^n$ (i.e. hydraulic fracturing parameters). We will jointly refer to all spatial and non-spatial variables as $\boldsymbol{x}_i = \{\boldsymbol{s}_i, \boldsymbol{z}_i\}$. Further in our developments, we assume that all functions in the set are realizations of a non-stationary random process that can be decomposed into a deterministic functional drift and globally second order stationary residual

$$\mathcal{X}_i(t) = m_i(t) + r_i(t) \tag{4.1}$$

we also assume that the drift is a function of all parameters and that it can be modeled with a functional linear regression model $m_i(t) = f(\boldsymbol{x}_i) = \sum_{l=0}^{n+2} f_l(\boldsymbol{x}_i) a_l(t)$. Furthermore, we assume that the residuals are spatially correlated and that their covariances are a function of spatial distance $cov(r_i(t), r_j(t)) = C(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$.

Here, we refer to the overall mean (or the average) of the ensemble of curves as $\mu(t)$, ortho-normal set of empirical fpc's as $e_k : \{\phi_1(t), \phi_2(t), ..., \phi_k(t)\}$ and functional principal component scores as $\xi_i^k = \langle \mathcal{X}_i(t) - \mu(t), \phi_k(t) \rangle$.

4.2.1 Universal Trace Kriging-based Forecasting

As in the previous chapter, we are seeking predictions of an unobserved function (production from an undrilled well) at some location $\boldsymbol{s}_0$ with a user specified set of covariates $\boldsymbol{z}_0$ as the best linear unbiased combination of all observed functions (production from existing wells)

$$\mathcal{X}_0^*(t) = \sum_{i=1}^{N} \lambda_i \mathcal{X}_i(t) \tag{4.2}$$

As in chapter 3, we seek $\lambda_1, \lambda_2, ..., \lambda_N$ that minimize the following objective criterion:

$$\underset{\lambda_1,\lambda_2,...,\lambda_N}{argmin}\, \mathbb{E}\left[\|\mathcal{X}_0^*(t) - \mathcal{X}_0(t)\|^2\right] \quad s.t. \quad \mathbb{E}\left[\mathcal{X}_0^*(t) - \mathcal{X}_0(t)\right] = 0 \tag{4.3}$$

The unbiasedness constraint is developed from the second term in (4.3) as follows

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathcal{X}}_0(t) - \mathcal{X}_0(t)\right] &= \mathbb{E}\left[\sum_{i=1}^{N} \lambda_i \mathcal{X}_i(t)\right] + \mathbb{E}\left[\mathcal{X}_0(t)\right] \\
&= \sum_{i=1}^{N} \lambda_i \mathbb{E}\left[\mathcal{X}_i(t)\right] - \mathbb{E}\left[\mathcal{X}_0(t)\right] \\
&= \sum_{i=1}^{N} \lambda_i \sum_{l=0}^{L} f_l(\boldsymbol{x}_i) a_l(t) - \sum_{l=0}^{L} f_l(\boldsymbol{x}_0) a_l(t) \\
&= \sum_{l=0}^{L} a_l(t) \left(\sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{x}_i) - f_l(\boldsymbol{x}_0)\right)
\end{aligned}$$

Obviously, this functional will be equal to zero if and only if

$$\sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{x}_i) = f_l(\boldsymbol{x}_0) \tag{4.4}$$

The updated optimization problem is as follows

$$\underset{\lambda_1,\lambda_2,...,\lambda_N}{argmin}\, \mathrm{Var}_t\left[\mathcal{X}_0^*(t) - \mathcal{X}_0(t)\right] \quad s.t. \quad \sum_{i=1}^{N} \lambda_i f_l(\boldsymbol{x}_i) = f_l(\boldsymbol{x}_0) \tag{4.5}$$

This optimization problem is solved in the same manner as universal trace kriging we presented in chapter 3 by introducing $L + 1$ Lagrange multipliers and setting the partial derivatives with respect to weights to zero. Therefore, the weights are computed with the following system of equations that we express here in matrix

form

$$\begin{bmatrix} C(\boldsymbol{s}_1,\boldsymbol{s}_1) & C(\boldsymbol{s}_1,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_1,\boldsymbol{s}_N) & f_0(\boldsymbol{x}_1) & f_1(\boldsymbol{x}_1) & \cdots & f_L(\boldsymbol{x}_1) \\ C(\boldsymbol{s}_2,\boldsymbol{s}_1) & C(\boldsymbol{s}_2,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_2,\boldsymbol{s}_N) & f_0(\boldsymbol{x}_2) & f_1(\boldsymbol{x}_2) & \cdots & f_L(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\boldsymbol{s}_N,\boldsymbol{s}_1) & C(\boldsymbol{s}_N,\boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_N,\boldsymbol{s}_N) & f_0(\boldsymbol{x}_N) & f_1(\boldsymbol{x}_N) & \cdots & f_L(\boldsymbol{x}_N) \\ f_0(\boldsymbol{x}_1) & f_0(\boldsymbol{x}_2) & \cdots & f_0(\boldsymbol{x}_N) & 0 & 0 & \cdots & 0 \\ f_1(\boldsymbol{x}_1) & f_1(\boldsymbol{x}_2) & \cdots & f_1(\boldsymbol{x}_N) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_L(\boldsymbol{x}_1) & f_L(\boldsymbol{x}_2) & \cdots & f_L(\boldsymbol{x}_N) & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \eta_0 \\ \eta_1 \\ \vdots \\ \eta_L \end{bmatrix} = \begin{bmatrix} C(\boldsymbol{s}_1,\boldsymbol{s}_0) \\ C(\boldsymbol{s}_2,\boldsymbol{s}_0) \\ \vdots \\ C(\boldsymbol{s}_N,\boldsymbol{s}_0) \\ f_0(\boldsymbol{x}_0) \\ f_1(\boldsymbol{x}_0) \\ \vdots \\ f_L(\boldsymbol{x}_0) \end{bmatrix} \tag{4.6}$$

Note that the only difference with respect to the universal trace kriging (UTrK) system of equations from chapter 3 is in the transformation functions $f(.)$ that are now operating on the entire vector $\boldsymbol{x}$ instead on vector $\boldsymbol{s}$ only.

Parameter inference is analogous to the parameter inference in universal trace kriging

1. Fit a piece-wise functional OLS model on pre-smoothed functional data[2].
2. Compute trace variogram on the residuals and fit one of the admissible variogram structures (Gau, Mat, Exp, Sph,...).
3. Fit a piece-wise functional GLS model to pre-smoothed functional data with covariance matrix corresponding to the trace variogram fitted in the previous step.
4. Iterate steps 2 and 3 a few times.
5. Use the final model for interpretation and forecasting.

4.2.2 FPCA-based Forecasting

Similar to trace kriging-based approach presented above, we can modify the universal cokriging of basis coefficient approach to accommodate for covariates $\boldsymbol{z}_0$. As shown

[2]Recall that functional regression produces the same results as piece-wise OLS on pre-smoothed functional data.

in the previous chapter, the drift in functional data can be expressed in terms of the drift in the coefficients of basis expansion or functional principal component scores. Namely, for each fpc score equation (3.31) implies the following decomposition

$$\xi_i^k = m_i^k + r_i^k \tag{4.7}$$

where the drift component depends on all covariates $m_i^k = \sum_{l=0}^{L} \beta_l f_l(\boldsymbol{x}_i)$ and the residual r_i^k is second order stationary and spatially correlated $\text{Cov}(r_i^k, r_j^l) = C_{kl}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$.

We forecast a new function $\mathcal{X}_0(t)$ at some location $\boldsymbol{s}_0$, with its own set of covariates $\boldsymbol{z}_0$, by forecasting its fpc scores $\{\xi_0^1, \xi_0^2, ..., \xi_0^k\}$ with the best linear unbiased combination of the fpc scores of all observed curves

$$\xi_{0j}^* = \sum_{i=1}^{N} \lambda_{ij}\xi_{ij} + \sum_{\substack{k=1 \\ k\neq j}}^{K} \sum_{i=1}^{N} \lambda_{ik}\xi_{ik} \tag{4.8}$$

As before, the weights are sought with an objective to minimize the mean squared error in predictions under unbiasedness constraints

$$\underset{\boldsymbol{\lambda}}{\text{argmin}} \quad \mathbb{E}\left[\left(\xi_{0j}^* - \xi_{0j}\right)^2\right] \quad s.t. \quad \mathbb{E}\left[\xi_{0j}^* - \xi_{0j}\right] = 0 \tag{4.9}$$

The unbiasedness constraints are developed from the second term in (4.9) as follows:

$$\begin{aligned}
\mathbb{E}\left[\xi_{0j}^* - \xi_{0j}\right] &= \mathbb{E}\left[\sum_{i=1}^{N} \lambda_{ij}\xi_{ij} + \sum_{i=1}^{N}\sum_{k\neq j}^{K} \lambda_{ik}\xi_{ik}\right] - \mathbb{E}\left[\xi_{0j}\right] \\
&= \sum_{i=1}^{N} \lambda_{ij}\mathbb{E}\left[\xi_{ij}\right] + \sum_{i=1}^{N}\sum_{k\neq j}^{K} \lambda_{ik}\mathbb{E}\left[\xi_{ik}\right] - \mathbb{E}\left[\xi_{0j}\right] \\
&= \sum_{i=1}^{N} \lambda_{ij}\sum_{l=0}^{L} b_l^j f_l(\boldsymbol{x}_i) + \sum_{i=1}^{N}\sum_{k\neq j}^{K} \lambda_{ik}\sum_{l=0}^{L} b_l^k f_l(\boldsymbol{x}_i) - \sum_{l=0}^{L} b_l^j f_l(\boldsymbol{x}_0) \\
&= \sum_{l=0}^{L} b_l^j \left(\sum_{i=1}^{N} \lambda_{ij} f_l(\boldsymbol{x}_i) - f_l(\boldsymbol{x}_0)\right) + \sum_{k\neq j}^{K}\sum_{l=0}^{L} b_l^k \sum_{i=1}^{N} \lambda_{ik} f_l(\boldsymbol{x}_i)
\end{aligned} \tag{4.10}$$

Equation (4.10) will be equal to zero if and only if

$$\begin{aligned} &\sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{x}_i) = f_l(\boldsymbol{x}_0), \quad \forall l; \\ &\sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{x}_i) = 0, \quad \text{for} \quad j \neq k, \quad \forall l; \end{aligned} \tag{4.11}$$

Therefore, we arrive to an updated constrained optimization problem

$$\underset{\boldsymbol{\lambda}}{\text{argmin}} \quad \text{Var}\left[\xi_{0j}^* - \xi_{0j}\right] \quad s.t. \quad \begin{cases} \sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{x}_i) = f_l(\boldsymbol{x}_0), & \forall l; \\ \sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{x}_i) = 0, & \text{for} \quad j \neq k, \quad \forall l \end{cases} \tag{4.12}$$

The solution to this constrained optimization problem is found in the same manner as in chapter 3 by developing the variance term, introducing $K \times (L+1)$ Lagrangian multipliers (η_{lk}), and setting the partial derivatives with respect to weights to zero. The final form of the system of equations that solves the constrained optimization problem is given below in matrix form.

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1K} & \boldsymbol{F} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2K} & \boldsymbol{0} & \boldsymbol{F} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K1} & \boldsymbol{C}_{K2} & \cdots & \boldsymbol{C}_{KK} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{F} \\ \boldsymbol{F}^T & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{F}^T & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{F}^T & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda_1} \\ \boldsymbol{\lambda_2} \\ \vdots \\ \boldsymbol{\lambda_K} \\ \boldsymbol{\eta_1} \\ \boldsymbol{\eta_2} \\ \vdots \\ \boldsymbol{\eta_K} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{10} \\ \boldsymbol{c}_{20} \\ \vdots \\ \boldsymbol{c}_{K0} \\ \boldsymbol{1} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \end{bmatrix} \tag{4.13}$$

Where:

$$\boldsymbol{F} = \begin{bmatrix} f_0(\boldsymbol{x}_1) & f_1(\boldsymbol{x}_1) & \cdots & f_L(\boldsymbol{x}_1) \\ f_0(\boldsymbol{x}_2) & f_1(\boldsymbol{x}_2) & \cdots & f_L(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(\boldsymbol{x}_N) & f_1(\boldsymbol{x}_N) & \cdots & f_L(\boldsymbol{x}_N) \end{bmatrix}, \boldsymbol{\lambda}_k = \begin{bmatrix} \lambda_{1k} \\ \lambda_{2k} \\ \vdots \\ \lambda_{Nk} \end{bmatrix}, \boldsymbol{\eta_k} = \begin{bmatrix} \eta_{0k} \\ \eta_{1k} \\ \vdots \\ \eta_{Lk} \end{bmatrix}$$

and $[\boldsymbol{C}_{kl}]_{ij} = C_{kl}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \text{Cov}(\xi_i^k, \xi_j^l)$

Once again the only thing that changes with respect to the system of equations given by equation 3.39 are the transformation functions $f(.)$ that are now operating on the entire vector $\boldsymbol{x}$ instead on vector $\boldsymbol{s}$ only.

4.3 Unconventional Reservoir Case Study

In this section, we evaluate the presented methodologies on a real unconventional reservoir case study. The dataset considered in this case study comes from one of the most prolific unconventional reservoirs in the United States and it was provided to us by Anadarko Petroleum Corporation (APC)[3]. The dataset contains 188 horizontal wells with multiple hydraulic fractures (stages). The horizontal wells produce oil, gas and in early months flow-back water from hydraulic fracturing operations. One example of well production data is shown in figure 4.1.

Since the presented forecasting methodologies are capable of forecasting only one functional variable at a time, in this case study the focus will be on oil production curves only[4].

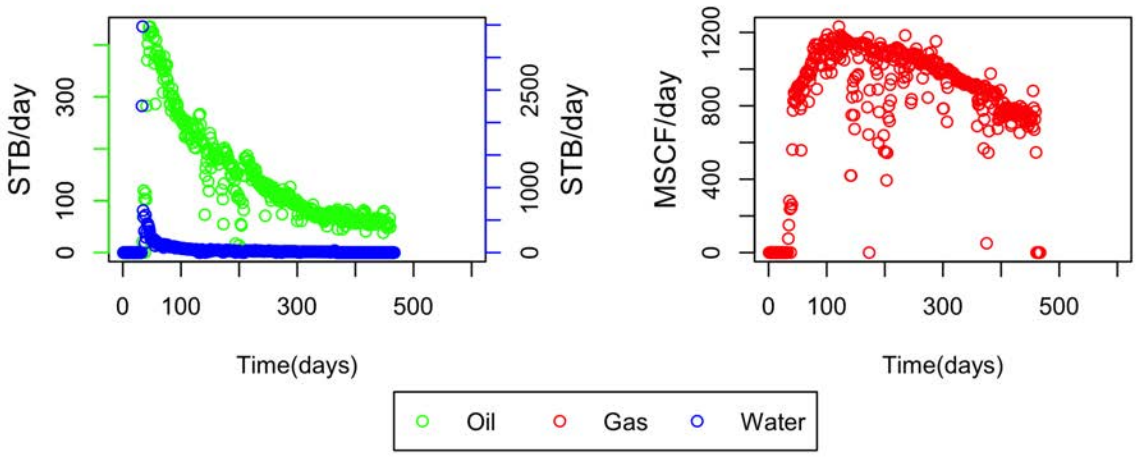


Figure 4.1: *An example of production data from one well in APC dataset. Time represents the number of days since the first day of production.*

[3]The exact name of the shale play is omitted due to the data confidentiality agreement with Anadarko Petroleum Corporation

[4]Multivariate forecasting of functions is left for later chapter

As a part of data pre-processing, for each well, we discarded oil production data that preceded the peak in oil rate since during those days wells mostly produced flow-back water from hydraulic fracturing operations[5]. Besides daily production rates, the dataset contained information on daily averages of well-head and down-hole pressures, daily well downtime and daily choke sizes for each well.

In addition to production data, a total of 26 parameters (hydraulic fracturing, petro-physical and geographical) were available for each well. A complete list of all available well parameters with their respective ranges is given in table 4.1.

4.3.1 Production Data Smoothing

The first step of functional data analysis is to convert the raw functional observations into smooth continuous curves. Unlike the Barnett shale case study (chapter 3), here we had a much richer dataset that enabled us to perform more advanced data smoothing. In particular, the availability of daily downtime information enabled us to handle the noise in production data more effectively.

Consider a plot of one production profile given in figure 4.2 left. Notice that the profile is "noisy" since many data points deviate from the overall trend in production decline. The problem here is that this "noise" is not noise per se[6]. Every single data entry in oil production curves has an explanation. In particular, consider the coloring of points in figure 4.2 left, where the color corresponds to daily well downtime. Notice that daily rates that deviate from the overall trend in production correspond to days during which the well was in production for less than 24 hours. The higher the downtime the larger the deviation from the overall trend. To eliminate this noise, it is first important to emphasize that reported daily production rates represent daily cumulative and not instantaneous rates. Secondly, reporting days represent calendar dates since the beginning of production and not the actual time in production. In such setting, a much more appropriate measure of well production performance are cumulative production curves plotted vs actual time in production. Such curves are often smoother than the cumulative production curves plotted vs reporting days (figure 4.2 right) since they effectively remove data entries with 24 hours of daily

[5]There is no conate water in the analyzed reservoir.

[6]i.e. normally distributed noise that comes as a consequence of imperfections in measurement device.

Table 4.1: *Well parameters*

Type	Parameter Name	**min**	**max**	Source	Unit
Geographical	GeolX_Rel*	865.0	89424.2	-	m
	GeolY_Rel*	0.0	-66243.3	-	m
Nearby Wells	ProdVert200*	0.00	0.29	-	scaled
	ProdVert300*	0.00	0.58	-	scaled
Petrophysical	Porosity	0.06	0.10	Well log	%
	Water Saturation	0.17	0.49	Well log	%
	Total organic content	0.04	0.06	Well log	-
	Vol. fract. of clay	0.06	0.21	Well log	%
	Vol. fract. of carbonates	0.48	0.75	Well log	%
	Vol. fract. of quartz	0.09	0.19	Well log	%
	Vol. fract. of pyrite	0.01	0.02	Well log	%
Fracturing	StagesPumped*	8	45	-	count
	Stimulated lateral length*	2194	9526	frac. log	ft
	Average fracture spacing*	99	476	frac. log	ft
	Number of screenouts	0	2	frac. log	count
	Total fluid pumped*	40507.74	259617.81	frac. log	bbl
	Amt. of slick water	0.00	199703.87	frac. log	bbl
	Amt. cross-link fluid	0.00	71978.74	frac. log	bbl
	Amt. of acid*	0.00	738.11	frac. log	bbl
	Amt. of linear fluid	0.00	35233.12	frac. log	bbl
	Clean fluid total	0.00	73578.78	frac. log	bbl
	Total proppant used*	1276781	8410537	frac. log	lbs
	Proppant per foot	323.53	1524.37	frac. log	lbs/ft
	100 Mesh sand total	0	92637	frac. log	lbs
	Resin coated sand total	0	360878	frac. log	lbs
PVT	Oil API gravity	42.8	53.6	-	API

downtime. They are also much easier to smooth with basis expansion, and given the fact that the expansion is performed with analytic basis functions[7] it is trivial to convert these curves into production rate vs time in production curves. One example of a smoothed cumulative oil vs time in production curve is given in figure 4.3 along with its first derivative (daily rate) plotted over reported daily production data vs reporting time.

[7]The derivative of the fit is available in this case which is convenient since the rate production curves are given by the first derivative of the cumulative production fits.

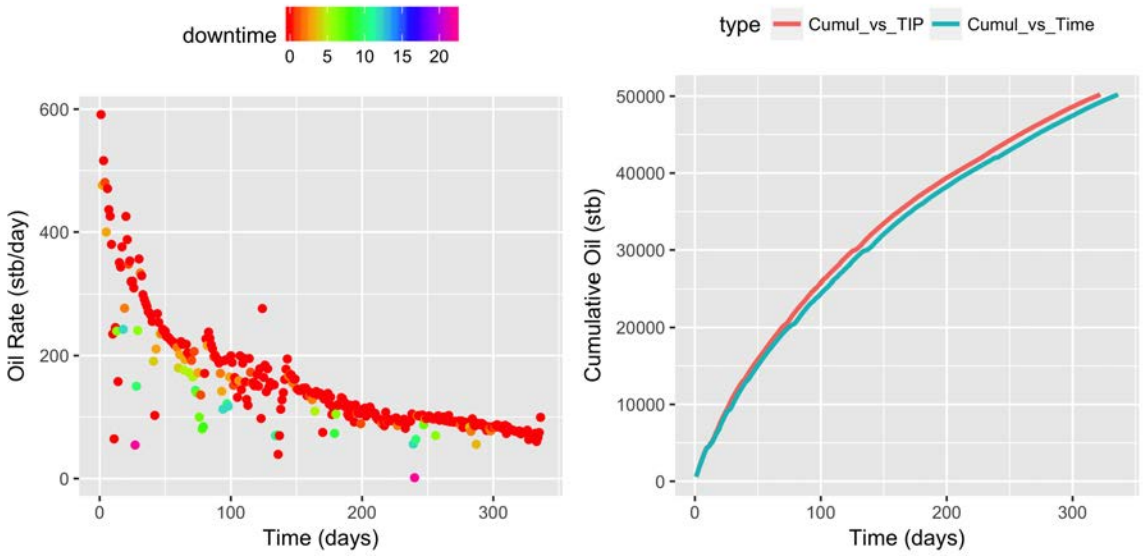


Figure 4.2: *Left - An example of noisy production data colored by daily downtime. Right - Cumulative production plotted vs time in production (TIP) and vs reporting time (Time).*

Due to the fact that wells did not start producing on the same date, at the time of dataset generation the length of production profiles varied between the wells. Given that all of the proposed forecasting methodologies assume the same length of functional data, data completion strategy outlined in chapter 2 had to be employed. We chose to work with a common time domain of 600 days in production since there were about 75 wells whose production profiles were of that or longer length. This number of wells was enough to accurately estimate the functional principal components and the mean as a common basis for curve completion (smoothing). After performing data smoothing and curve completion on the cumulative production, we computed the final ensemble of rate vs time curves by taking the first derivative of the final cumulative fits. The final ensemble of smoothed curves is given in figure 4.4 right.

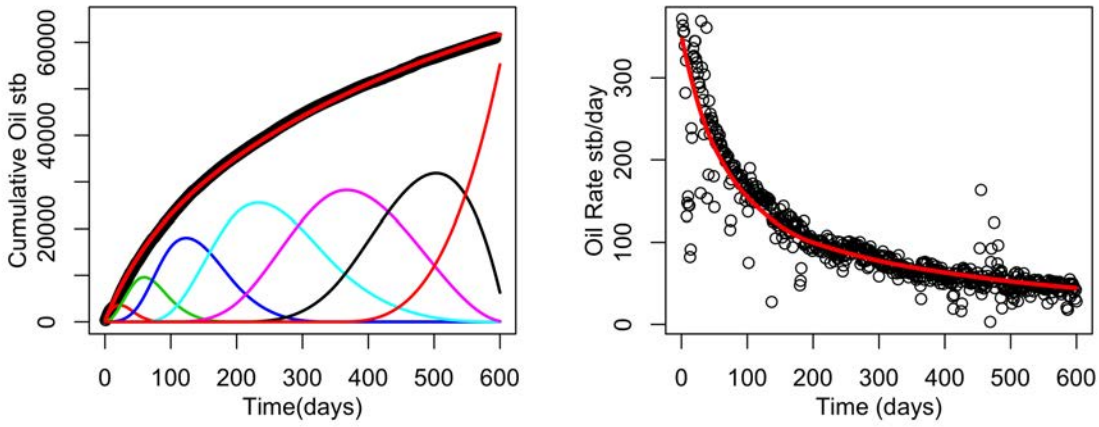


Figure 4.3: *An example of curve smoothing. Left - cumulative production vs. time in production with basis expansion and resulting fit. Right - production rate vs. time in days with the first derivative of the fit on the left (red curve).*

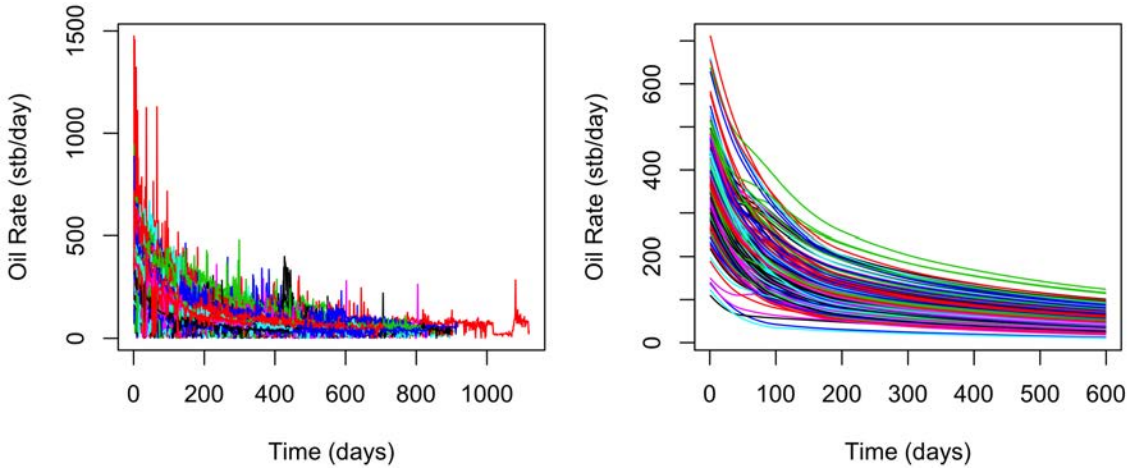


Figure 4.4: *Left - Raw rate vs time data. Right - the final smoothed ensemble of curves.*

4.3.2 Sensitivity Analysis

In this section, we outline the results of sensitivity analysis that helped us better understand the relationships between available well parameters and oil production. The results were produced with distance based generalized sensitivity analysis (DGSA,

Fenwick et al. [2014]). The DGSA method starts by clustering production data, and then computing cluster specific cumulative density functions on each of the available parameters. The cluster specific cumulative density functions are compared to the overall cumulative density function of the analyzed parameter, computed on the entire dataset (all clusters). Departures in cluster specific cumulative density functions from the overall cumulative density function suggest that the parameter is influential on the analyzed response, while the lack of departure suggests the converse. The actual quantification of sensitivity is performed by averaging L_1 norms computed between the cluster specific CDFs and the overall CDF. In this way computed sensitivities are then used for parameter ranking.

In our analysis, we first computed Euclidean distances between the smoothed oil production curves, followed by multidimensional scaling (MDS) that produced a two dimensional space for clustering (figure 4.5) in which we simply employed k-means method (Hastie et al. [2009]) to cluster the responses. The plots of cluster specific and the overall CDFs of each well parameter are shown in figure 4.7, while the low dimensional distance plots colored by each of the input parameters are shown in figure 4.8. Influential parameters show trends in low dimensional colored plots. For example, stimulated lateral length is an important parameter since its trend is quite apparent in the colored low-dimensional distance plot, while "PetroSwt" is a non-influential parameter since it shows complete absence of a trend in the colored low-dimensional distance plot.

The final ranking of the parameters is given in figure 4.6. We observe that hydraulic fracturing parameters appear as the most influential on oil response followed by location and PVT parameters. What is surprising is the low ranking of petrophysical parameters, in particular "PetroTOC" (total organic content). This low ranking is due to the lack of variability in input parameters that is also apparent from the CDF plots. "PetroTOC" took only three values (0.04, 0.05, 0.06) that were also inconsistent with the total organic content values extracted from vertical well logs drilled in the study area[8].

[8]Besides horizontal well data the APC dataset contained 2500 vertical well logs from the same study area.

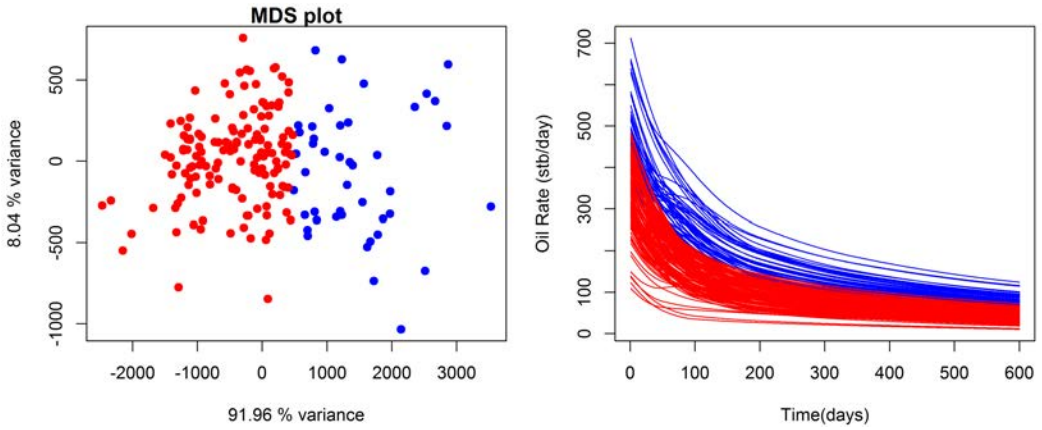


Figure 4.5: *Left - MDS plot of production produced with Euclidean distance and clusters produced with k-means clustering. Right - k-means clustering viewed on original production*

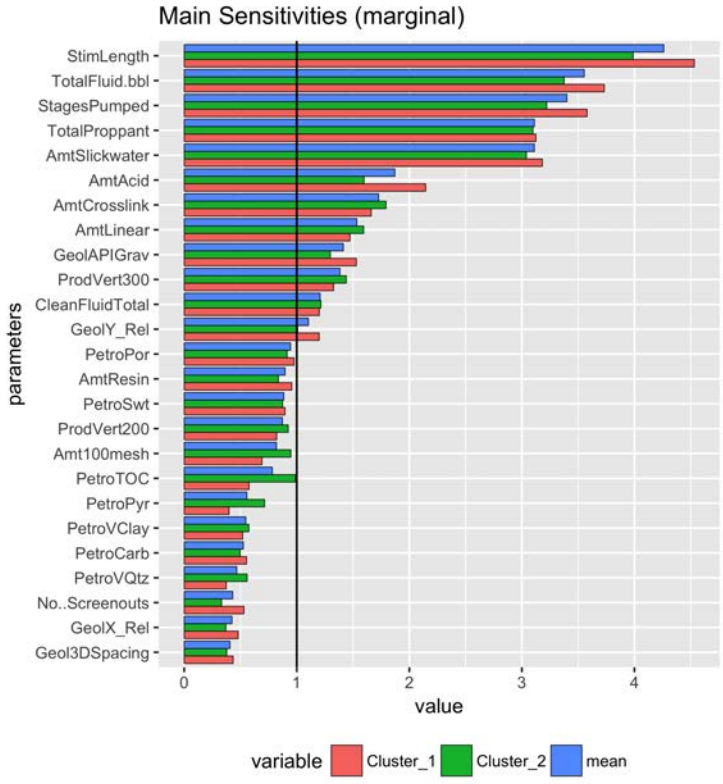


Figure 4.6: *DGSA - Pareto plot*

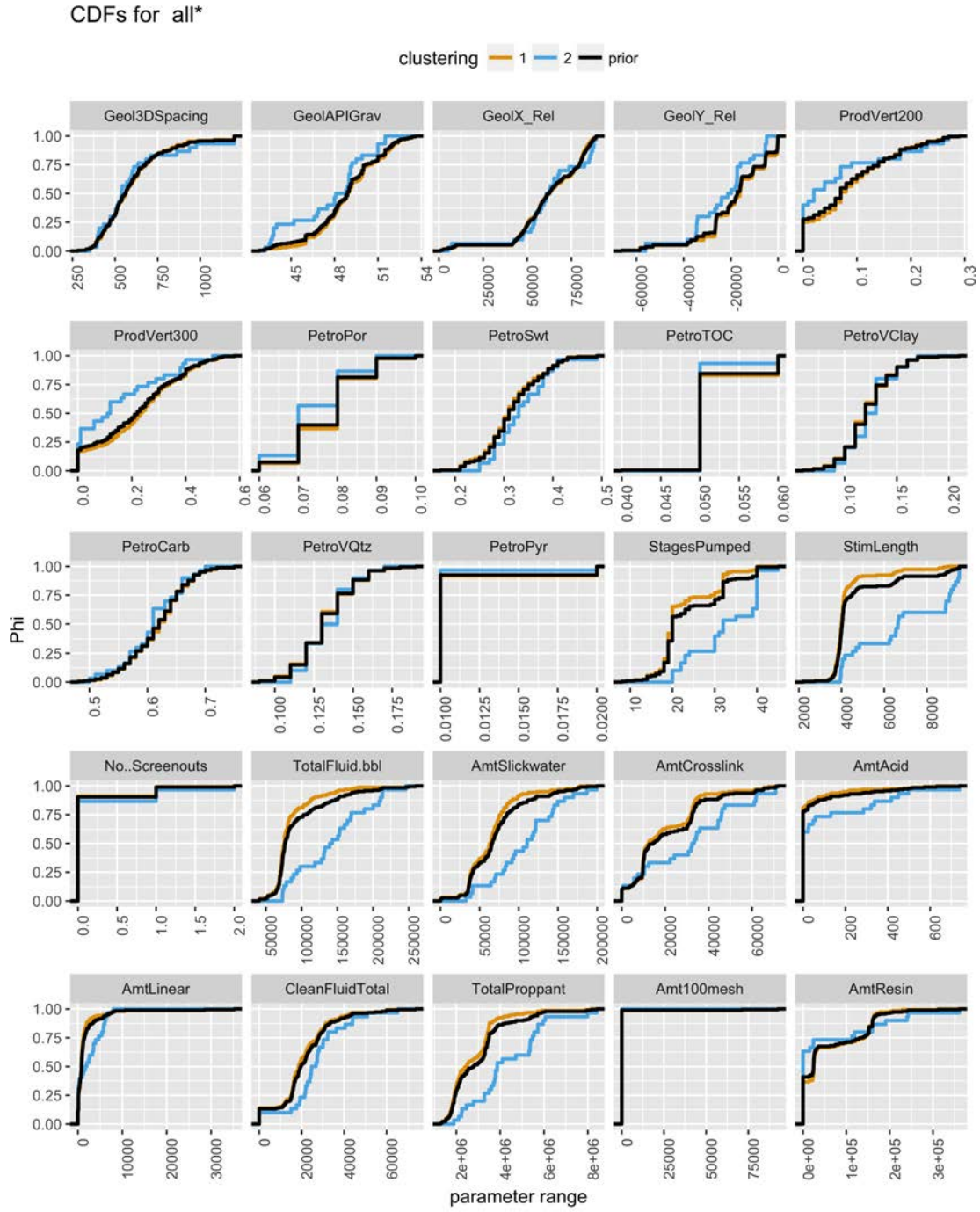


Figure 4.7: *DGSA - CDF analysis*

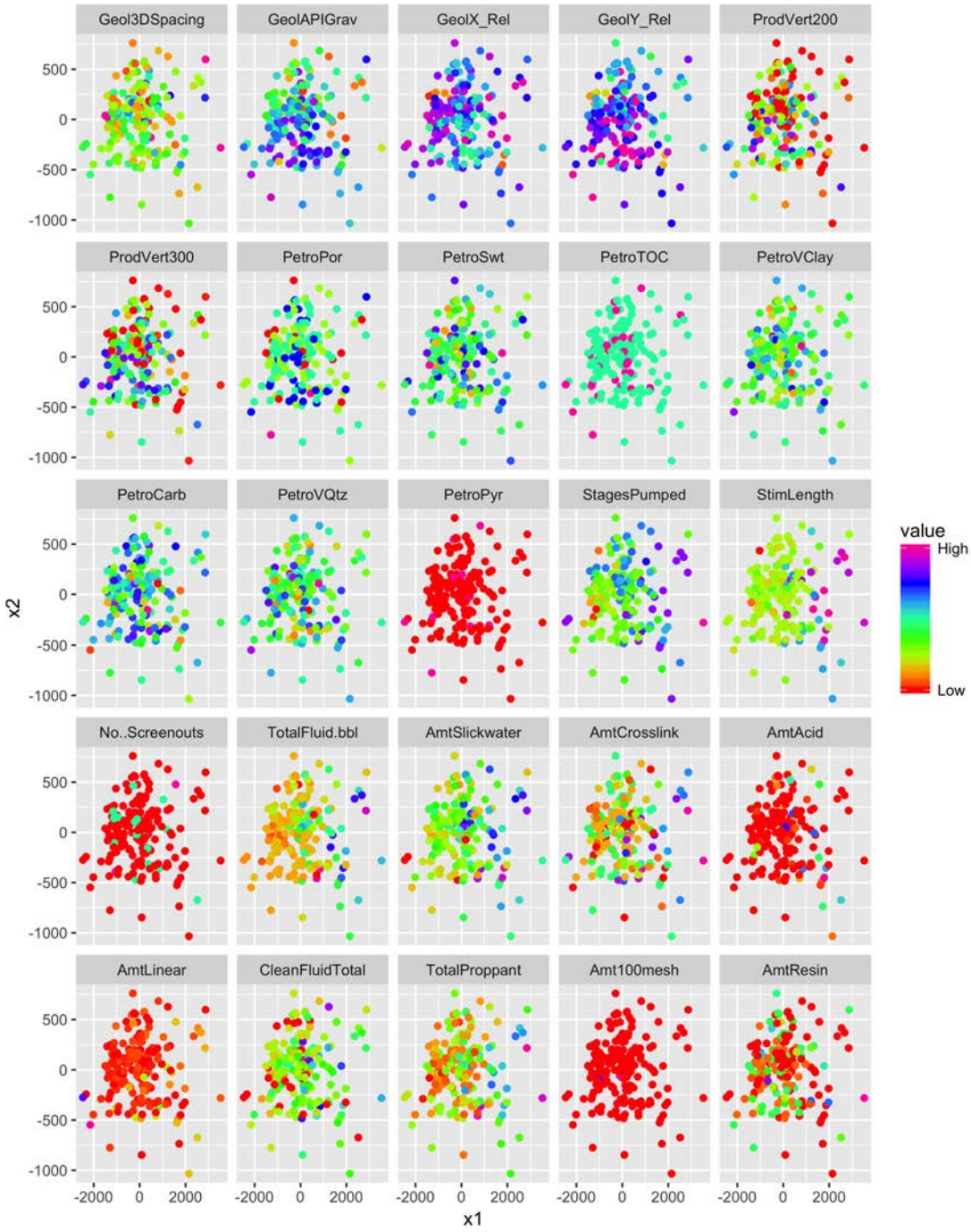


Figure 4.8: *DGSA - Scatter Plot Analysis*

4.3.3 Functional PCA

The next step in our data mining workflow is to examine the functional principal components. A plot of the first two fpcs as perturbations around the mean function is given at the top of figure 4.9. This situation is fairly similar to the fpcs computed on the Barnett shale dataset. The first fpc acts as a scalar that shifts curves upwards or downwards, while the second fpc describes the shift in time, potentially describing the onset of bilinear flow. In general, interpretation of fpcs is difficult since they both describe variations in all parts of the analyzed time domain. To improve interpretability, as before, we apply varimax rotation to this set of fpcs and arrive to a set of rotated fpcs shown at the bottom of figure 4.9. At this point things become much clearer. The first rotated fpc describes the variation in the tail of the production while the second rotated fpc describes the variation in early oil production. This result is also very similar to the rotated fpcs on the Barnett shale dataset.

Given that all wells have different completions, one interpretation of this result is as follows. The first rotated fpc most likely depends on the well locations and total organic content, or in other words reserves. The second fpc is most likely correlated with hydraulic fracturing parameters, since in early days wells are draining the network of artificial fracture networks.

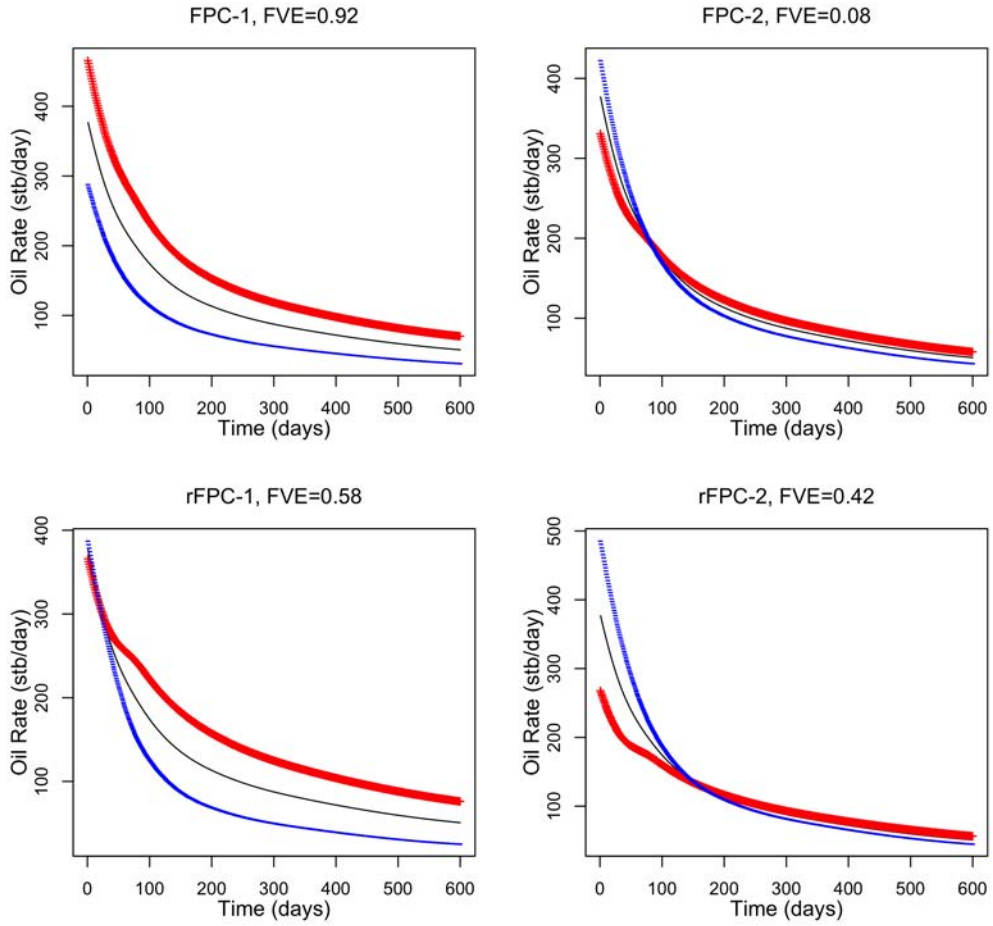


Figure 4.9: *Top - The first two fpcs as perturbations around the mean function. Bottom - The two rotated fpcs as perturbations around the mean.*

4.3.4 Geostatistical Analysis

Next in our analysis, we will examine the spatial correlations between the production curves. Here, we will apply the two methodologies outlined in the methodology section of this chapter, on the entire dataset (188 curves). We approached the problem from a forecasting perspective hence we chose to work with the parameters that are available to modelers before a well is drilled, namely the high ranking hydraulic fracturing parameters from the pareto-plot in figure 4.6 and well locations. Well parameters that were used in this study are marked with a star in table 4.1. Prior to this analysis, all non-spatial input parameters were standardized, while spatial parameters were scaled

to [0,1] scale. This was necessary due to numerical issues with drift modeling. First, we will examine the omni-directional trace variogram given in figure 4.10. The model fitted to this variogram is of Matérn type with a range of 0.17 scaled units (15km in original units). This spatial correlation represents the continuity in the overall reservoir quality and not the continuity of some reservoir parameter (i.e. porosity). What this variogram informs is that one can expect a similar production profiles of two wells completed in the same way, and separated by at most 15km.

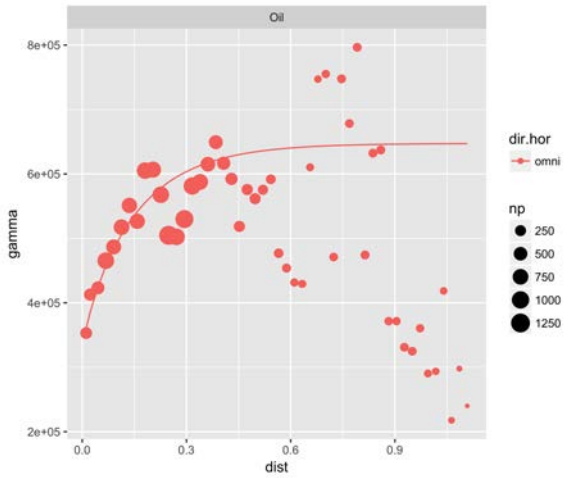


Figure 4.10: *Trace variogram on oil rates.*

We fitted a universal co-kriging with covariates model to the rotated functional principal component scores. The variograms computed on the residuals are given in figure 4.11. The model fitted to these empirical variograms is Matérn with a range of 0.17 (15km). These variograms are very similar to the trace variogram analyzed previously. We observe that the residuals have a negative spatial cross-correlation in this case.

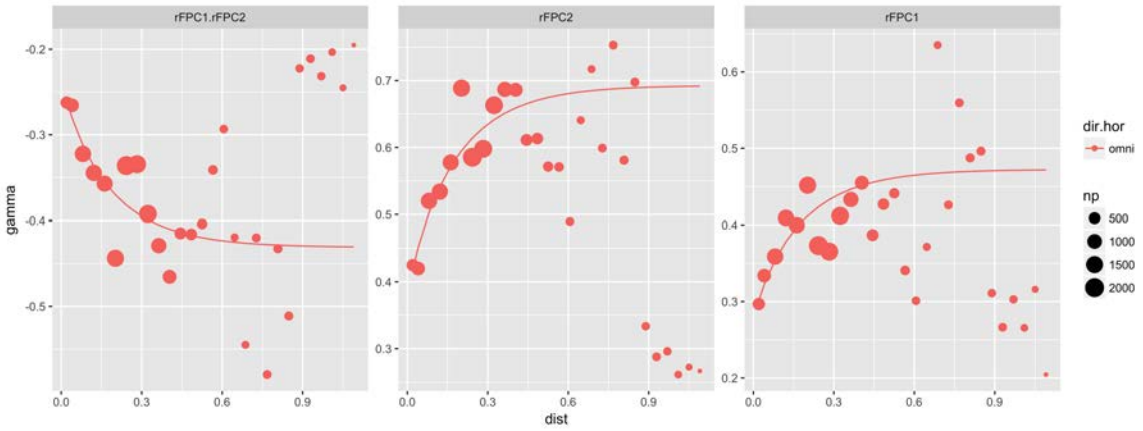


Figure 4.11: *Variograms of the residuals of the rotated functional principal component scores*

4.3.5 Forecasting Study

Next, we evalute the forecasting capabilities of the presented methodologies. We split the dataset into 100 randomly selected wells that were used for training, and 88 wells that were used for testing. We recomputed the rotated functional principal components on the training set and fitted trace-based and projection-based forecasting models. We then used the models to predict the wells from the test set. A few forecasts of the test wells are shown in figure 4.12. Notice that in this case the two modeling approaches produced similar forecasts since the variogram ranges were the same. SSE errors are summarized in table 4.2.

Table 4.2: *SSE Error table*

Method	min	mean	median	sd	max
Universal Trace Kriging	0.002	0.451	0.214	0.63	3.762
U. Cokriging of rot. fpc scores	0.004	0.446	0.211	0.615	3.606

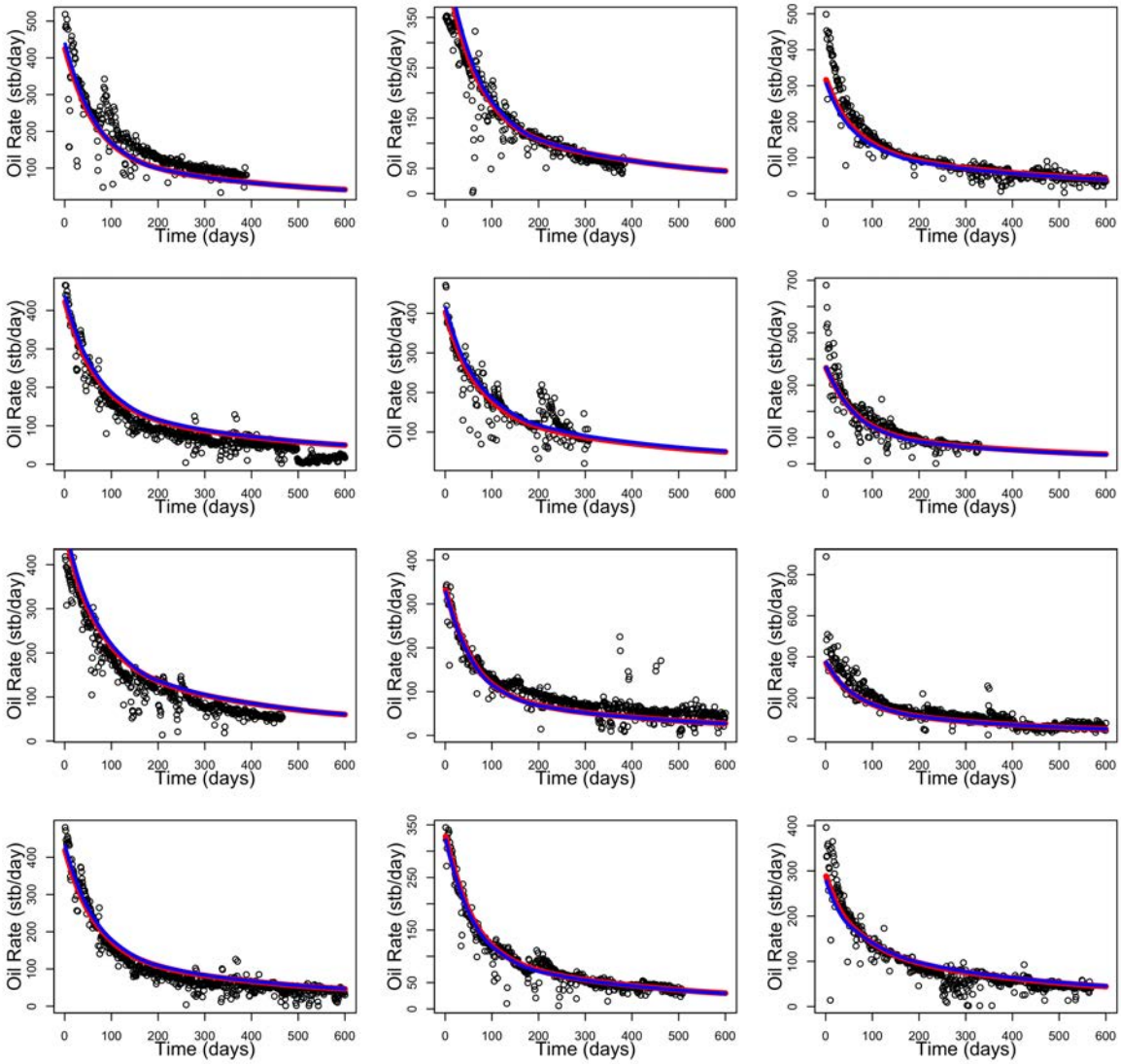


Figure 4.12: *A few forecasts. Black dots represent true data, red curves are forecasts produced with universal co-kriging on rotated fpcs and blue curves are universal trace kriging forecasts.*

Next we applied sequential Gaussian co-simulation to produce confidence bands around the forecasts. In figure 4.13 we are plotting three realizations of the rotated functional principal components with added spatial components of the drift. The overall trend in the map of the first rotated fpc corresponds to the trend in API gravity (figure 4.14) of oil.

A total of 100 co-sgsim realizations of the residuals were produced and then combined with the drift term computed for each test well to produce 88 forecasting ensembles of curves (one ensemble for each test well). A few randomly selected ensembles are plotted in figure 4.15 along with the actual production data. Notice that, in general, the ensembles of the forecasts fully enclose the true data.

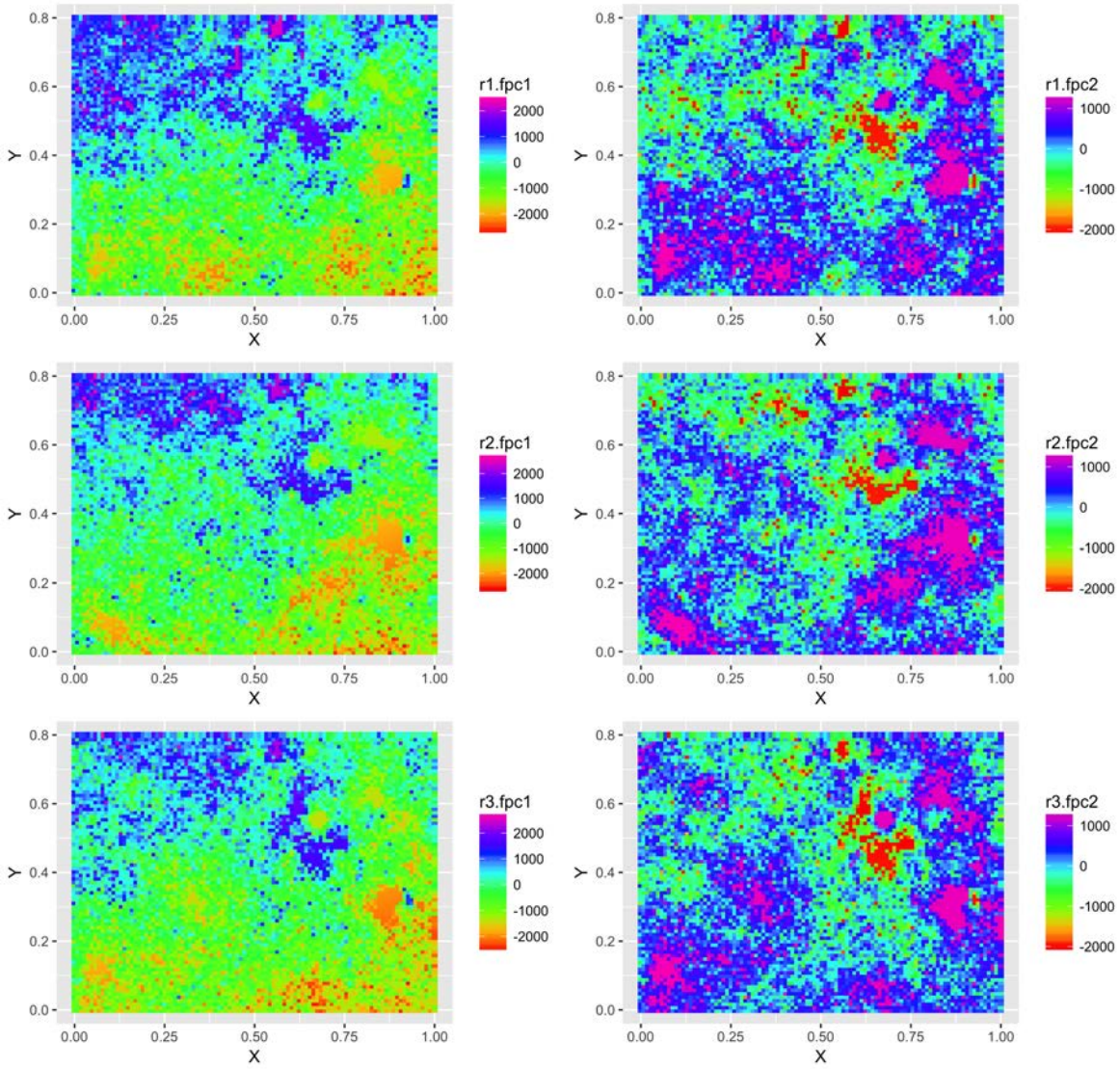


Figure 4.13: *Co-sgsim realizations of rotated fpc scores.*

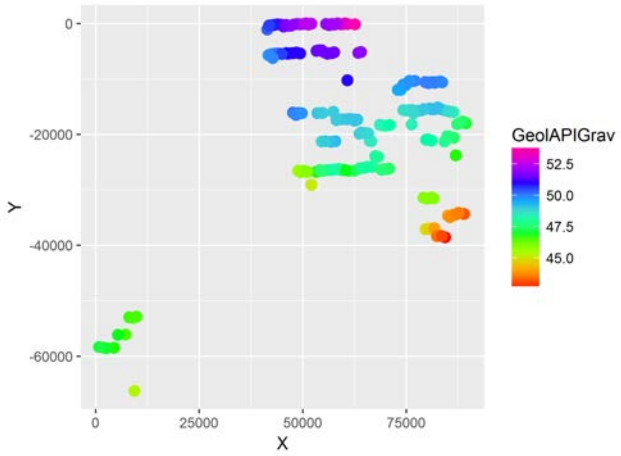


Figure 4.14: *Well locations colored by API gravity.*

4.3.6 Monte Carlo Study

Next we evaluated the performance under variable training set sizes. We varied the size of the training set from 30 to 150 wells and on each iteration we predicted a non-overlapping test set of size 38. For each training set size we produced 100 random train/test splits of the available data and fitted universal trace kriging and universal co-kriging of rotated fpc scores models. At every iteration we computed the mean and the median of the SSE on the test set. Distributions of the mean and the median of the SSE are summarized in figure 4.16. We observe that the distribution of the median was similar for the two methods, however the mean was slightly higher for universal co-kriging of coefficients. This discrepancy is a consequence of two things: fpc truncation error and difficulties with LMC fitting with low sample sizes. In our study we also observed a lot of outliers in the mean of the SSE. These outliers are a consequence of extrapolation. Kriging methods are great interpolators but very poor extrapolators which is one limitation of the presented methods. From figure 4.16 we also observe that the median stabilizes for training set sizes of 60-70 wells suggesting that is the minimum number of wells that is necessary to produce reliable forecasts with the proposed methodologies.

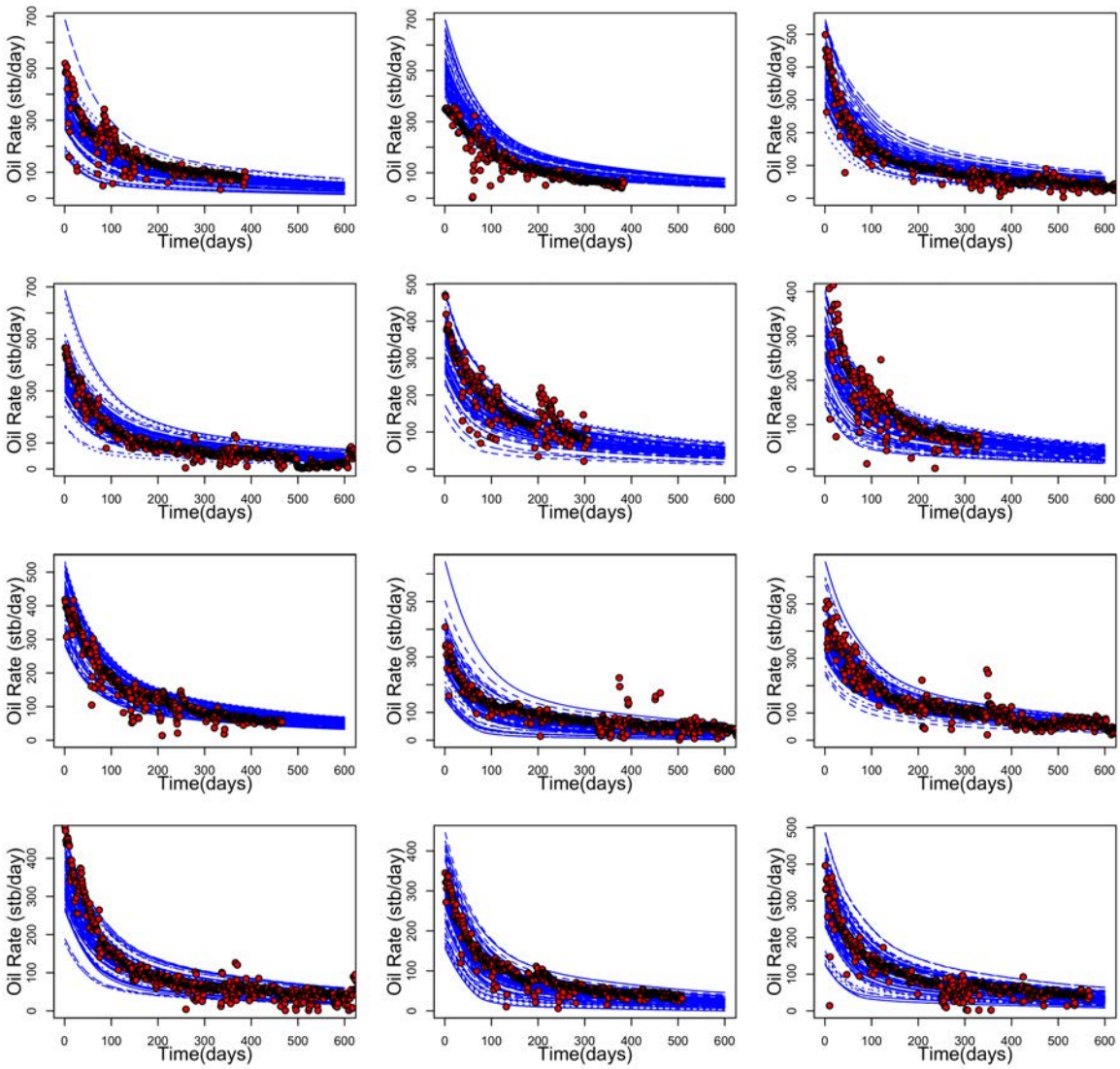


Figure 4.15: *Blue curves - forecasts produced with co-simulation of rotated fpc scores. Red dots - true data.*

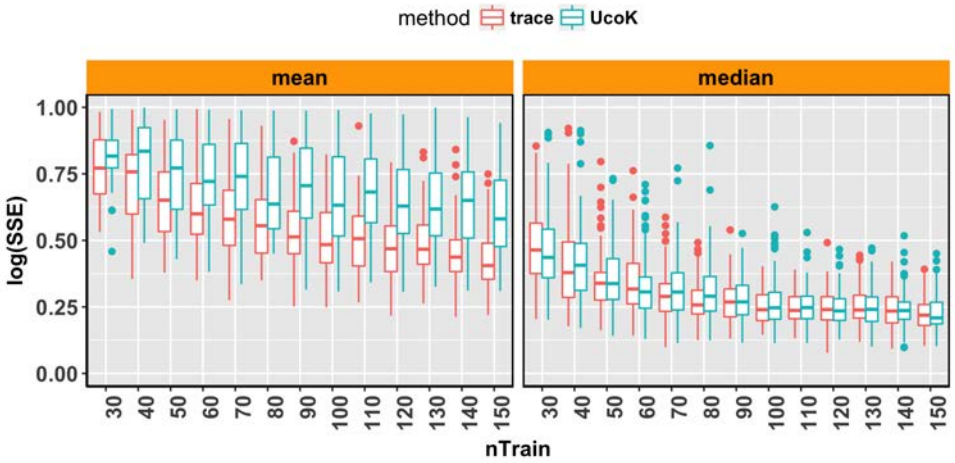


Figure 4.16: *Monte Carlo analysis - The influence of the training set size on forecasting capabilities.* ***trace*** *= Universal Trace Kriging,* ***UcoK*** *Universal co-kriging of coefficients*

4.4 Chapter Conclusion

In this chapter, we showed that the methods from chapter 3 can be used with slight modification for forecasting of shale hydrocarbon production curves with variable hydraulic fracturing parameters. The methodologies were found to produce similar results in both data mining and forecasting studies on the Anadarko data set with 188 horizontal wells with multiple frac-jobs. However, the approach with rotated functional principal components in combination with sequential Gaussian co-simulation was found to be the most adequate for the types of analyses and forecasting studies in unconventional reservoir engineering. The first rotated fpc that describes the variation in late oil production was found to be highly correlated with API gravity of oil, while prediction bands generated with sequential Gaussian simulation fully enclosed the true data. Currently, the methodologies rely on distance-based generalized sensitivity analysis for evaluation of parameter importances and selection since parameter selection for functional regression is an ongoing research area. This practical issue leaves space for improvement in future research work.

Chapter 5

Interpretation and Forecasting of Multivariate Functional Data with Regression Trees

While trace variograms and variograms evaluated on the residuals of basis coefficients are powerful interpretation tools, drift modeling techniques for functional data are not quite interpretative. The coefficients in functional regression are functions that are in many cases almost impossible to interpret. On the other hand the nature of oil and gas data is such that explanatory variables are always correlated[1] that ultimately makes the t-tests in principal component regression highly misleading[2]. Moreover, the techniques outlined in the previous chapter are capable of working with single variate functional outputs. This is a significant limitation given that hydrocarbon wells often produce multiple fluids (i.e. oil and gas) giving rise to multivariable functional forecasting problem.

In this chapter, we develop a regression tree-based methodology that is capable of producing highly interpretable drift models. In the original developments by Breiman et al. [1984], regression trees were designed to work with scalar outputs. Later work by Segal [1992] expanded the method to accommodate for multivariate outputs. Here, we propose an expansion of the original idea that enables us to grow regression trees with multivariate outputs of any type, including multivariate functional outputs such

[1] i.e. the number of fracturing stages is often correlated with the amount of injected fluid/proppant
[2] Multicollinearity problem in ordinary least-squares (Hastie et al. [2009])

as oil and gas production curves.

This chapter is organized as follows. In the methodology section we outline the basics of regression trees followed by a review of the method by Segal [1992] for growing regression trees with multivariate outputs (vectors). We then proceed to develop an approach for growing regression trees with multivariate functional outputs. The method is demonstrated on the Anadarko dataset introduced in chapter 4; only, in this case, we are analyzing and forecasting all functional outputs, oil and gas curves, together at the same time.

5.1 Methodology

Here, we consider a set of functions $\{\mathcal{X}_i(t), t \in T\}_{i=1}^N$ (i.e. oil production curves) observed over a set of spatial locations $\boldsymbol{s}_i \in D \subset R^2$ along with a set of explanatory variables or covariates $\boldsymbol{z}_i \in R^n$. We will refer to all spatial and non-spatial covariates as $\boldsymbol{x}_i = \{\boldsymbol{z}_i, \boldsymbol{s}_i\}$. As in the previous chapters, we assume that the functions are non-stationary and that they can be decomposed into a deterministic mean and globally second order stationary functional residual

$$\mathcal{X}_i(t) = m_i(t) + \delta_i(t) \tag{5.1}$$

the drift term $m_i(t)$ is assumed to depend on all spatial and non-spatial covariates $\boldsymbol{x}$, and in this chapter it will be modeled with functional regression trees that we introduce next. The residual will be assumed to depend only on spatial locations and, as such, it will be modeled with the functional interpolation methods outlined in chapter 3 (OTrK, UCoK).

5.1.1 Regression Trees

Consider a training set $\mathcal{T} : \{(\boldsymbol{x_i}, y_i)\}_{i=1}^N$ where $\boldsymbol{x}_i$ is a vector in R^n and y_i is a scalar output ($y_i \in R$), and let $f : \boldsymbol{x} \to y$ be a function that maps $\boldsymbol{x}_i$'s to y_i's. The idea of regression trees is to partition the p-dimensional input space spanned by $\boldsymbol{x}_i$'s into M disjoint sub-regions R_m's, and then in every sub region approximate the true function f with some local function f_m. This local function can be a constant (i.e. the local mean), a local regression or even a Gaussian process. Input space partitioning can

be performed in many ways, however, one of the most widely adopted partitioning schemes is the binary recursive splitting method proposed by Breiman et al. [1984]. Their method is binary because it considers binary splits of R^n along one predictor (covariate) at a time. It is recursive because regions are recursively split into sub-regions until some stopping criteria is met, or, no more training data is left to split. The procedure is also greedy since, in determining the best split of a region it does not consider the quality of later splits. To determine the best split of one region, the method relies on user specified cost function G. For example, when considering some split s along predictor p of region R_m into two sub-regions R_{ml} and R_{mr} one would compute the quality of the split as follows:

$$Q^m_{s,p} = G(R_m) - [G(R_{ml}) + G(R_{mr})] \tag{5.2}$$

The split that has the highest quality is accepted and the procedure continues to further refine the newly formed regions. For trees with scalar outputs that approximate the true function with the local mean, the most appropriate cost function is the sum of squared residuals $G_{sse}(R_m) = \sum_{y_i \in R_m} (y_i - \mu_m)^2$.

Regression trees have very high interpretative capabilities. Every step of the recursive partitioning procedure can be recorded in a form of a decision tree that visually puts the whole splitting process into perspective. An example of a recursive regression tree produced on an input space spanned by two parameters (X_1 and X_2) is given in figure 5.1.

Recursive splitting can be performed on pretty much any type of input parameters. Splitting on continuous input parameters is trivial, the data is simply ordered and the split point is moved from the lowest to the highest point of the parameters range. Categorical predictors have two cases, orderable and unorderable. For orderable categorical predictors the splitting procedure is the same as for continuous, while for unorderable one has to consider different combinations of categories. Obviously this becomes tedious and numerically difficult for a large number of categories[3].

Another type of predictors that commonly occurs in Earth sciences are the functional predictors (i.e. relative permeability curves, porosity distributions etc.). This is a special kind of input parameter that cannot be directly treated with any of the previously considered splitting approaches. Instead, we propose a simple two step

[3]In our experience more than six categories are already intractable.

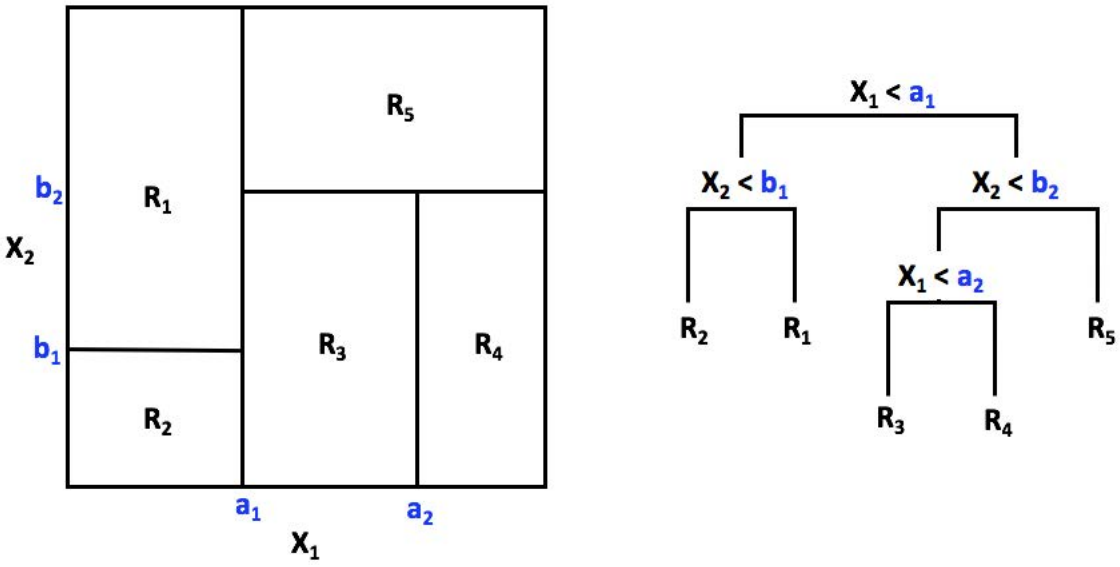


Figure 5.1: *An example of recursive splitting. Left - Recursively partitioned input space; Right - The corresponding regression tree*

ordering procedure for this job.

- First, the realizations of a complex input parameter (i.e. rel permeability curves) are clustered with hierarchical clustering (Eisen et al. [1998], Hastie et al. [2009]) with a distance metric appropriate for the type of the analyzed functional predictor.
- A leaf reordering algorithm (Bar-Joseph et al. [2001]) is run as a post-processor on the hierarchical tree, from the previous step, to create an ordered sequence of complex parameter realizations.
- Once an ordered sequence is established we employ the same splitting technique as in the case of orderable continuous or categorical predictors.

An example of ordering of a complex predictor is given in figure 5.2.

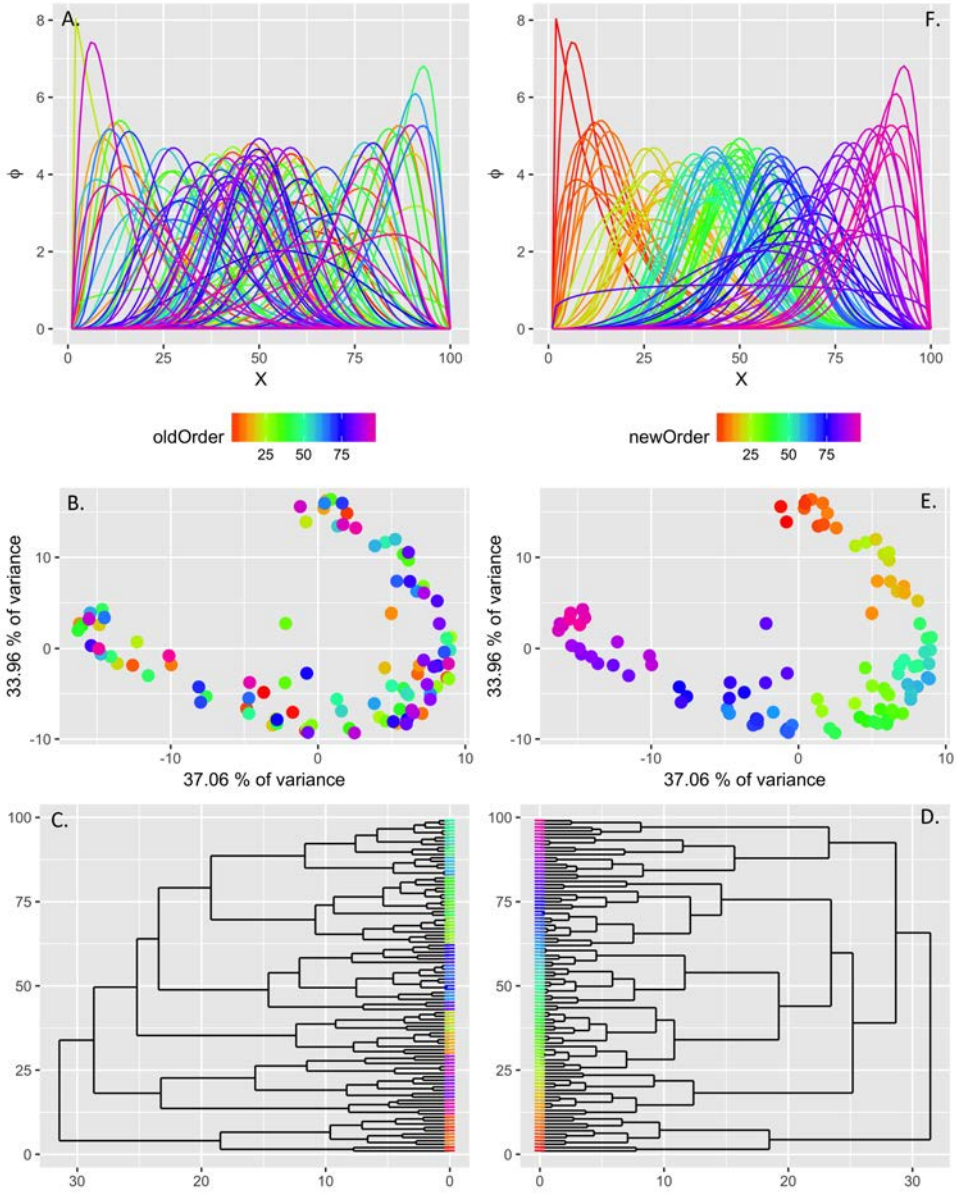


Figure 5.2: *An example of ordering of complex predictors. A - Raw unordered functional predictor data. B - Low dimensional (MDS) representation of the data colored by original (old) ordering. C - Hierarchical clustering performed on the raw data. D - Leaf reordered hierarchical clustering dendrogram. E - A low dimensional representation (MDS) colored by the new ordering of the data. F - A plot of the original data colored by the new ordering.*

Bootstrapped Regression Trees

While having great interpretative properties, regression trees forecasting capabilities are not so impressive since they do not produce any type of prediction bands and in addition, they also tend to over-fit the training data. As a remedy for these problems, Breiman [1994] proposed bootstrapping. The idea is simple, the training data is sampled many times with replacement and a regression tree is fitted to each sample. In this way, an ensemble of regression trees is produced. This ensemble is then used on a new set of predictors to produce an ensemble of forecasts. Finally, the ensemble of forecasts is used to construct prediction bands and the mean prediction. This procedure of generating ensembles of bootstrapped trees is commonly referred to as "bagging".

In later research, it was observed that predictors that are highly correlated with the output always placed high in the dendrogram of a regression tree thereby never giving a chance to less but still significantly important predictors. To give equal chance to all predictors in the training data and avoid over-fitting, Breiman [1999] proposed a double bootstrapping procedure or better known as "random forest". This idea also starts by bootstrapping the observations as bagging, however when growing regression trees, on each split, it only considers a randomly selected (without replacement) subset of predictors rather than all of the available predictors. This procedure is numerically faster than bagging, and it was also found to produce better forecasts in some cases.

Variable Importance

The decision tree topology is indicative of variable importance. Since the tree building procedure is greedy, input parameters that are the most correlated with the output are usually used in earlier splits, while the less correlated parameters either appear on later splits or are not used in splitting at all. To quantify input parameter importance, Breiman et al. [1984] proposed the following variable importance (sensitivity) index

$$S_p = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N_m} max\{Q_{s,p}^m, \forall s\} \tag{5.3}$$

Where: $Q_{s,p}^{m}$ is given by equation (5.2), M is the number of splits in a tree and N_m is the number of training points within region m.
The sensitivity index S_p quantifies the overall reduction in cost function caused by splitting on parameter p. This index can also be computed for bagged trees by simply averaging over all trees in the ensemble. The value of the sensitivity index is not influenced by the correlations between the input parameters (unlike t-test in linear regression). For example, sensitivity indices S_p of perfectly correlated input parameters would have the same value[4].

5.1.2 Functional and Multivariate Regression Trees

Previously reviewed regression trees considered a simple situation with scalar outputs ($y_i \in R$). In this sub-section we will review and develop strategies for building regression trees when the output data are vectors ($\boldsymbol{y} \in R^n$), functions ($\mathcal{X}(t), t \in \tau$), and multiple functions ($\boldsymbol{\mathcal{X}} = (\mathcal{X}^1(t), \mathcal{X}^2(t), ..., \mathcal{X}^k(t)), t \in \tau$). To grow regression trees with such complex outputs, modifications at the level of cost functions ($G(R_m)$) are needed. For instance, when the outputs are vectors, Segal [1992] proposed to grow regression trees with the following cost function

$$G(R_m) = \sum_{\boldsymbol{y}_i \in R_m} (\boldsymbol{y}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_{\boldsymbol{m}}^{-\mathbf{1}} (\boldsymbol{y}_i - \boldsymbol{\mu}_m) \tag{5.4}$$

This cost function is essentially a sum of Mahalanobis distances between the output vectors $\boldsymbol{y}_i$ contained in m-th region, and the regions mean vector $\boldsymbol{\mu}_m$. One practical difficulty with this approach lies in the fact that the covariance matrix ($\boldsymbol{\Sigma}_m$) needs to be estimated in every region (for every split) and that such matrix must be positive semi-definite. A simple practical workaround is to use a common covariance matrix for all splits in a tree.
Regression trees with functional outputs. When the outputs are functions, there are several ways in which one could proceed. The simplest approach is to expand the functional data onto a set of K basis functions[5]

[4]For a detailed discussion on the topic of variable importance in regression trees please consult Louppe et al. [2014]
[5]for example, a B-Spline basis system or a set of functional principal components

$$\mathcal{X}_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

and then work with the expansion coefficients, c_{ik}'s. In this way, functional data is transformed into multivariate data and regression trees can be grown with the cost function given with equation (5.4) (Yu and Lambert [1999]). As an alternative, for smooth functional data, such as hydrocarbon decline curves considered in this dissertation, one could grow regression trees with a cost function that is widely used in functional data analysis (Ramsay and Silverman [2005])

$$G(R_m) = \sum_{\mathcal{X}_i(t)\in R_m} \int_T (\mathcal{X}_i(t) - \mu_m(t))^2 dt \tag{5.5}$$

Where $\mu_m(t)$ is the mean function of m-th region.

This cost function produces the same results as the following cost function:

$$G(R_m) = \sum_{\mathcal{X}_i(t)\in R_m} \|(\mathcal{X}_i(t) - \mu_m(t))\|^2 \tag{5.6}$$

which is a special case of the cost function (5.4) for $\mathbf{\Sigma}_m = \boldsymbol{I}$, that computes the sum of Euclidean distances from the regions mean vector.

Distance-based tree growing strategy. A universal and much more general strategy for growing regression trees is by means of similarity distances. Let $\hat{D}^m_{K\times K}$ be a similarity distance matrix (i.e. Euclidean) between the outputs in region m. To split the region based on distances, we formulate the following cost function

$$G(R_m) = min\{\sum_{i=1}^{K} d^m_{ij}; 1 < j < K\} = \sum_{i=1}^{K} d^m_{i,medoid} \tag{5.7}$$

where d^m_{ij} is ij-th element of $\hat{D}^m_{K\times K}$.
Unlike the previously introduced cost functions that compute either the sum of Mahalanobis or Euclidean distances from the regions mean, this cost function computes the sum of distances from the most central data point of the region (in terms of

outputs), the medoid[6]. This concept is very robust and applicable to a variety of situations commonly occurring in Earth sciences. For example, it is not unusual to have multiple outputs of different type such as functions and vectors, or functions and images, or as it is the case in this dissertation oil and gas production curves. To grow trees with such complex outputs, one would compute appropriate distance matrices on each output type and then simply compute a joint distance matrix as follows

$$D_{mv} = \sum_{k=1}^{K} w_k D_k^* \tag{5.8}$$

Where:

D_k^* is a scaled distance matrix computed on k-th output type

w_k is an optional weight given to k-th output type[7]

This matrix of joint distances is then used to build a regression tree with the cost function given by equation (5.7). Computation of parameter importances on such regression tree is performed in the same way as before, only in this case parameter importance indices reflect parameters influence on all considered output types together. Forecasts produced with this type of tree are the local means of each output type.

5.1.3 Method Summary

Several modeling workflows can be envisioned with functional regression trees and spatial interpolation methods outlined in chapter 3. Table 5.1 outlines all possible modeling workflows.

[6]Working with medoids is computationally faster since distances between outputs are computed only once prior to running the partitioning algorithm

[7]This parameter is determined through cross validation or its simply user specified.

Table 5.1: *Possible modeling workflows*

Funct. Output Type	Drift Model	Residual Model
Single-Variate	1. Reg. tree with cost (5.6) 2. Basis expansion and Reg. tree w/ cost (5.4) 3. Distance modeling and Reg. tree with cost (5.7)	1. Ordinary Trace Kriging 2. Ordinary Co-kriging of basis coefficients (or fpcs)
Multi-Variate	Distance modeling of each output type. Combine distances with eq(5.8), then Reg. tree w/ cost (5.7)	1. Ordinary Trace Kriging of each output type 2. Ordinary Co-kriging of basis coefficients of each output type 3. Ordinary Co-kriging of basis coefficients of all output types[8]

5.2 Case Study

In this section, we demonstrate the previously introduced modeling methodology on oil and gas responses from 188 horizontal wells in the Anadarko dataset introduced in the previous chapter. To smooth gas production curves we applied the same strategy as in chapter 4. We smoothed cumulative gas rates vs. time in production (in hr) and then we took the first derivative of each fit to arrive to a set of gas rate vs time in production curves. The final ensemble of oil and gas production curves is given in figure 5.3. Prior to smoothing, we conducted the same data pre-processing as in chapter 4. Production data that preceded the peak in oil rate was discarded, since it did not represent the actual reservoir response due to flow back water from hydraulic fracturing. It is interesting to notice in figure 5.3 right that a large number of gas production curves did not decline concurrently with the oil production curves. Instead, many of the gas production curves increase until some peak in gas rate is achieved before they start declining. This behavior is most likely a consequence of sorbed gas being released from solution with the drop in reservoir pressure. This particular case is a perfect example of the power of non-parametric curve smoothing.

[8] We outline this procedure in chapter 6.

Gas production behavior described previously cannot be adequately parameterized with any of the widely used decline curve models.

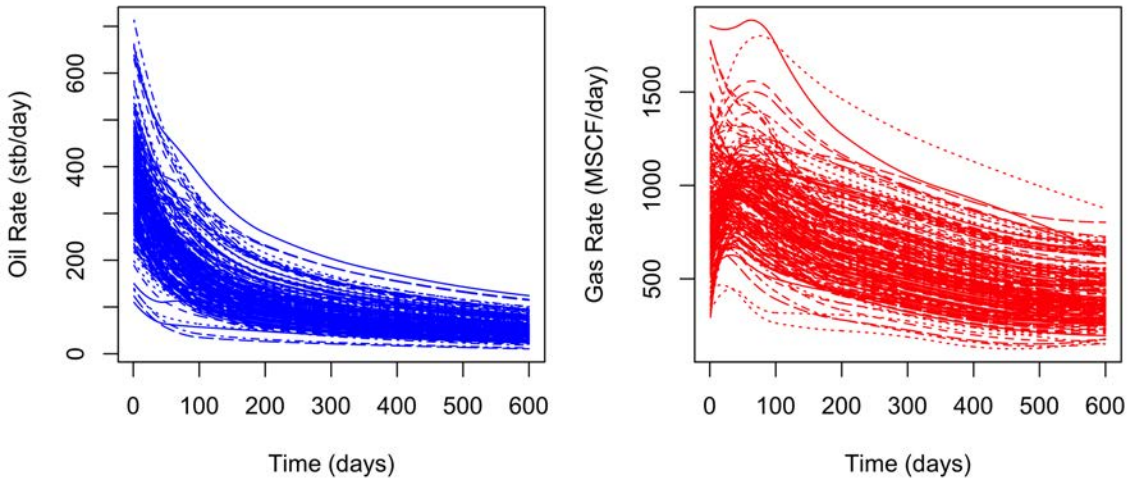


Figure 5.3: *Left - Smoothed oil production curves. Right - Smoothed gas production curves*

5.2.1 Data Analysis

First, we will use the tree based methodology to perform data mining. We computed Euclidean distances between all 188 oil curves (D_{oil}) and Euclidean distances between all 188 gas curves (D_{gas}). Both distance matrices were scaled to [0,1], and later combined into a joint distance matrix with equation (5.8)

$$D_{joint} = D^*_{oil} + D^*_{gas}$$

The joint distance matrix was then used to build a multivariate regression tree with the distance-based cost function (5.7). The fitted multivariate tree is given in figure 5.4, while the variable importance plot corresponding to this tree is given in figure 5.5 right. On the same joint distance matrix, we performed distance-based generalized sensitivity analysis (Fenwick et al. [2014]) in order to compare its results with the results of the tree-based variable importance analysis. Distance-based sensitivity analysis requires the output data to be pre-clustered. This was done with k-means

method (Hastie et al. [2009]) on the joint distance matrix. The final parameter ranking based on DGSA is given in figure 5.5 left while the low dimensional[9] scatter plots (of the joint distance matrix) colored by each of the analyzed input parameters is shown in figure 5.7. We observe that the tree based method and DGSA produced quite similar results. In both cases, hydraulic fracturing parameters ranked very high, and were almost immediately followed by X and Y well locations suggesting a strong spatial dependence. Analysis of the regression tree in figure 5.4 also suggests strong parameter interactions. In different parts of the reservoir, different hydraulic fracturing parameters become important. Proper quantification of parameter interactions is left for future work.

Next in our data analysis, we performed trace variography. To model the drift, we used bagging procedure with the same multivariate setup as outlined previously. The fitted bagged trees were used to predict the entire training set[10] and compute the functional residuals. Trace variograms shown in figure 5.6 were computed on the residuals of oil and gas rates individually. In this case, we do not observe any meaningful spatial structure. Both variograms appear to be pure nuggets.

5.2.2 Forecasting Study

In this subsection, we evaluate the forecasting capabilities of the proposed functional regression tree based methodology. As in previous chapter, we randomly split the dataset into 100 training wells and 88 testing wells. We computed Euclidean distances between the oil rates and the gas rates separately and combined them into one joint distance matrix that was then used to build a distance based (eq. (5.7)) random forest. There are two reasons why we used random forest for forecasting. Firstly, boostrapped trees are capable of producing confidence bands around forecasts, which is always required in oil and gas uncertainty quantification studies. Secondly, the prediction accuracy of random forests is expected to be the same or better than of bagged trees. A few forecasts produced with the random forest are shown in figures 5.8 and 5.9. Since there was no apparent spatial structure on the residuals, geostatistical modeling was not performed on this dataset. SSE errors summarized in table 5.2

[9] low dimensional representation of the distance matrix was computed with multidimensional scaling - MDS (Borg and Groenen [2005])

[10] Bagged trees are an ensemble procedure that makes many predictions for one set of inputs.

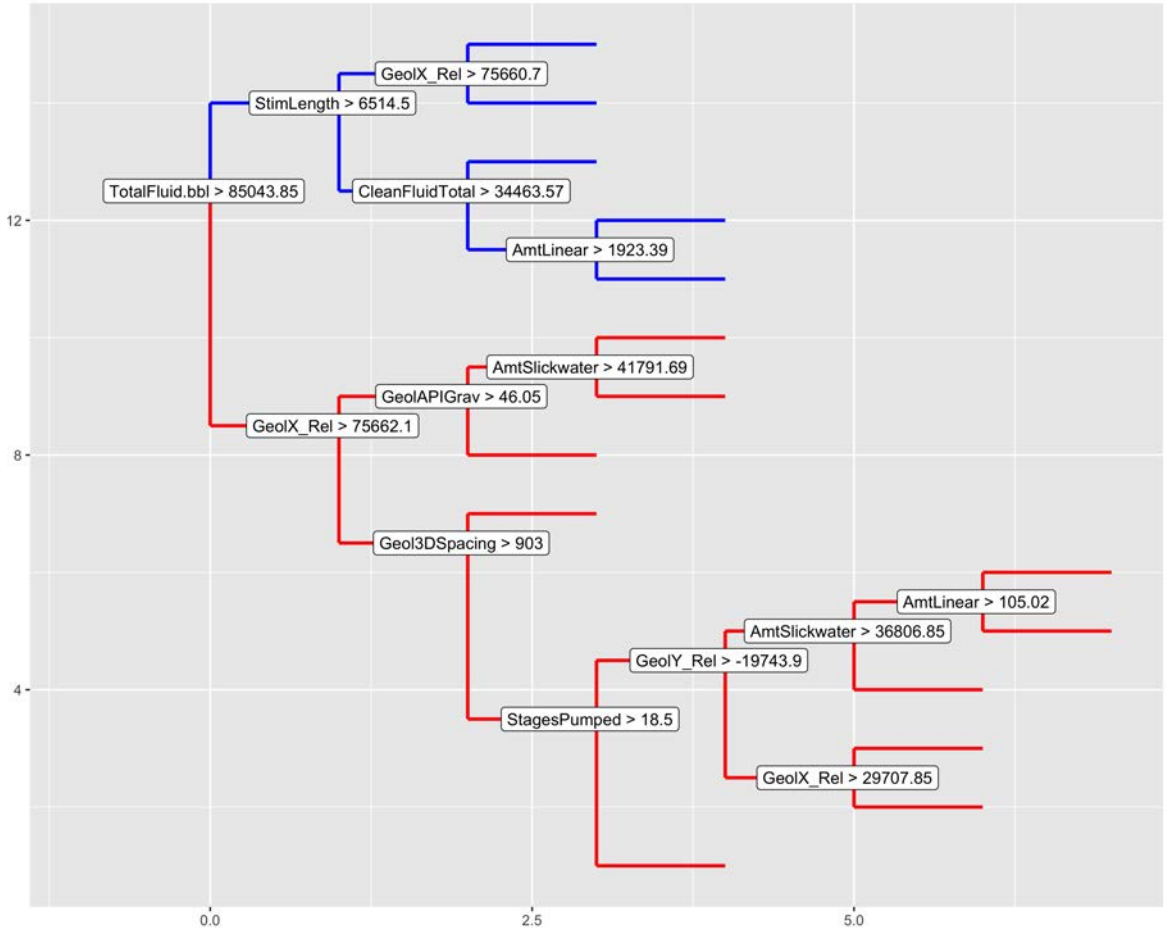


Figure 5.4: *Multivariate tree fitted on the entire dataset with cost function* (5.7) *and joint distance matrix computed on oil and gas responses*

were computed between the mean of the forecasts produced with the random forest and the smoothed versions of the raw data. As in the previous chapters, all errors were normalized with the trace variance of the entire dataset. The median of the error was around 22% in both cases, however the containment of the true data within prediction bands was 94% for oil and 89% for gas[11].

[11]To assess the containment within prediction bands we used the bagplot procedure by Rousseeuw et al. [1999]. The true data was considered to be outside the prediction bands if it was deemed an outlier in the bag plot analysis.

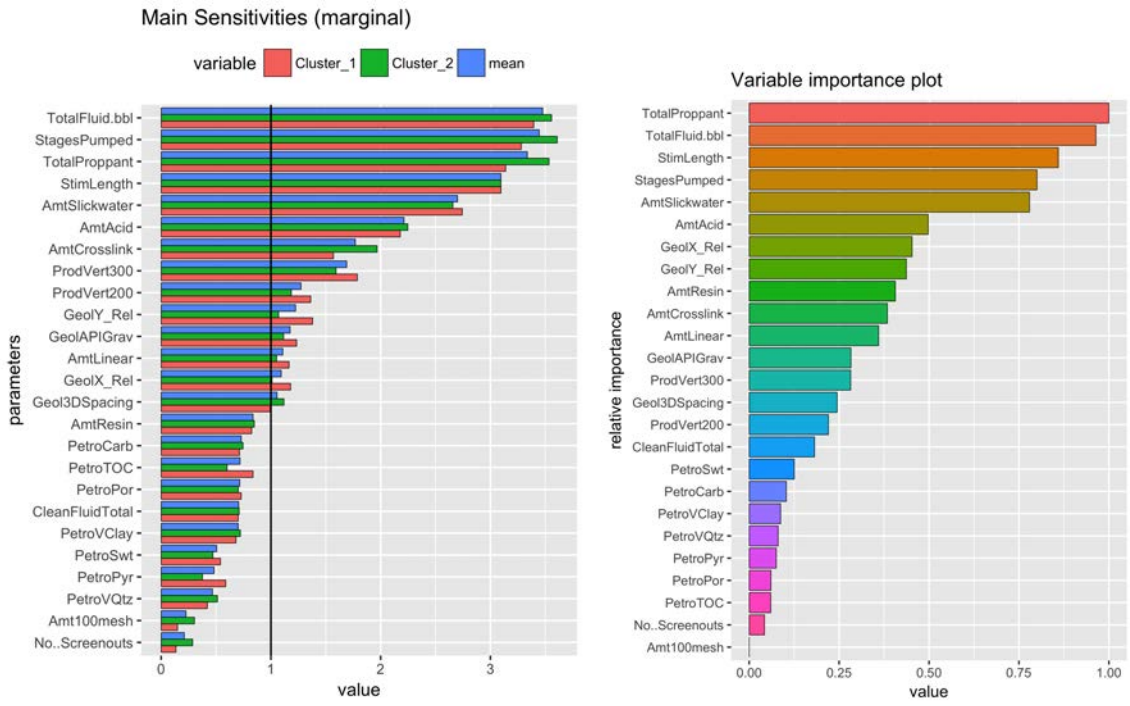


Figure 5.5: *Left - DGSA sensitivity on joint data; Right - multivariate tree variable importance.*

Monte Carlo Study

To better asses the error and its dependence on the size of the training set, we set up a Monte Carlo study in which we varied the size of the training set from 30 to 150 wells. For each training set size, we randomly sampled 100 training/testing sets on which we fitted a multivariate distance-based random forest, made predictions and computed the mean and the median of the test set SSE. In every case, the test sets had a size of 38 and were non-overlapping with the corresponding training data. For comparison purposes, we also assessed the predictive performance of universal trace kriging (UTrK) models fitted on oil and gas responses separately and universal co-kriging (UCoK) models fitted on the coefficients of oil and gas responses jointly[12]. The variation in the mean and the median of the SSE test error as a function of the training set size is shown in figure 5.10 for each modeling approach.

[12]Formulating a universal co-kriging system on the coefficients of multivariate functional data is straight forward. This approach is explained in great detail in the following chapter.

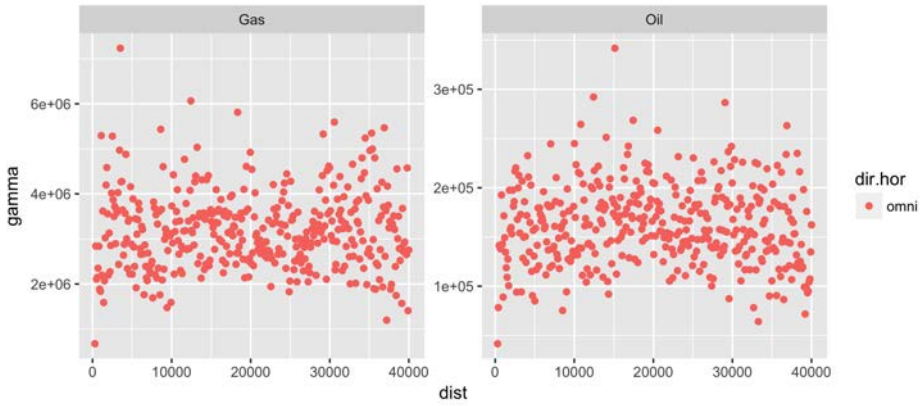


Figure 5.6: *Trace variograms computed on the residuals of oil and gas rates.*

Table 5.2: *Random Forest - SSE Error table (PB = Prediction Bands)*

Method	Type	min	mean	median	sd	max	# in PB	% in PB
Functional	Oil	0.006	0.43	0.237	0.514	2.393	83	94.3
Random Forest	Gas	0.002	0.50	0.235	0.660	3.276	79	89.7
Universal	Oil	0.006	0.477	0.264	0.672	4.135	-	-
Trace Kriging	Gas	0.002	0.513	0.204	0.832	4.602	-	-
U. Cokriging	Oil	0.001	0.619	0.215	0.971	6.149	-	-
of coefficients	Gas	0.003	0.639	0.248	1.029	5.141	-	-

From figure 5.10 we observe that in terms of the mean and the median of the SSE multivariate random forest outperformed the other methods for low training set sizes. For large training set sizes, universal trace kriging (UTrK) and multivariate random forest appear to produce similar results. Note the outliers in the plot of the mean of the normalized SSE, these are only present in the case of kriging based methods and are a consequence of erroneous solutions caused by extrapolation[13]. Note that the random forest error starts to stabilize around 50-60 wells in both the mean and the median, on both oil and gas responses. This suggests that the method starts to become reliable when the number of produced wells is around 50.

[13]It is well known that kriging is a good interpolator but a poor extrapolator

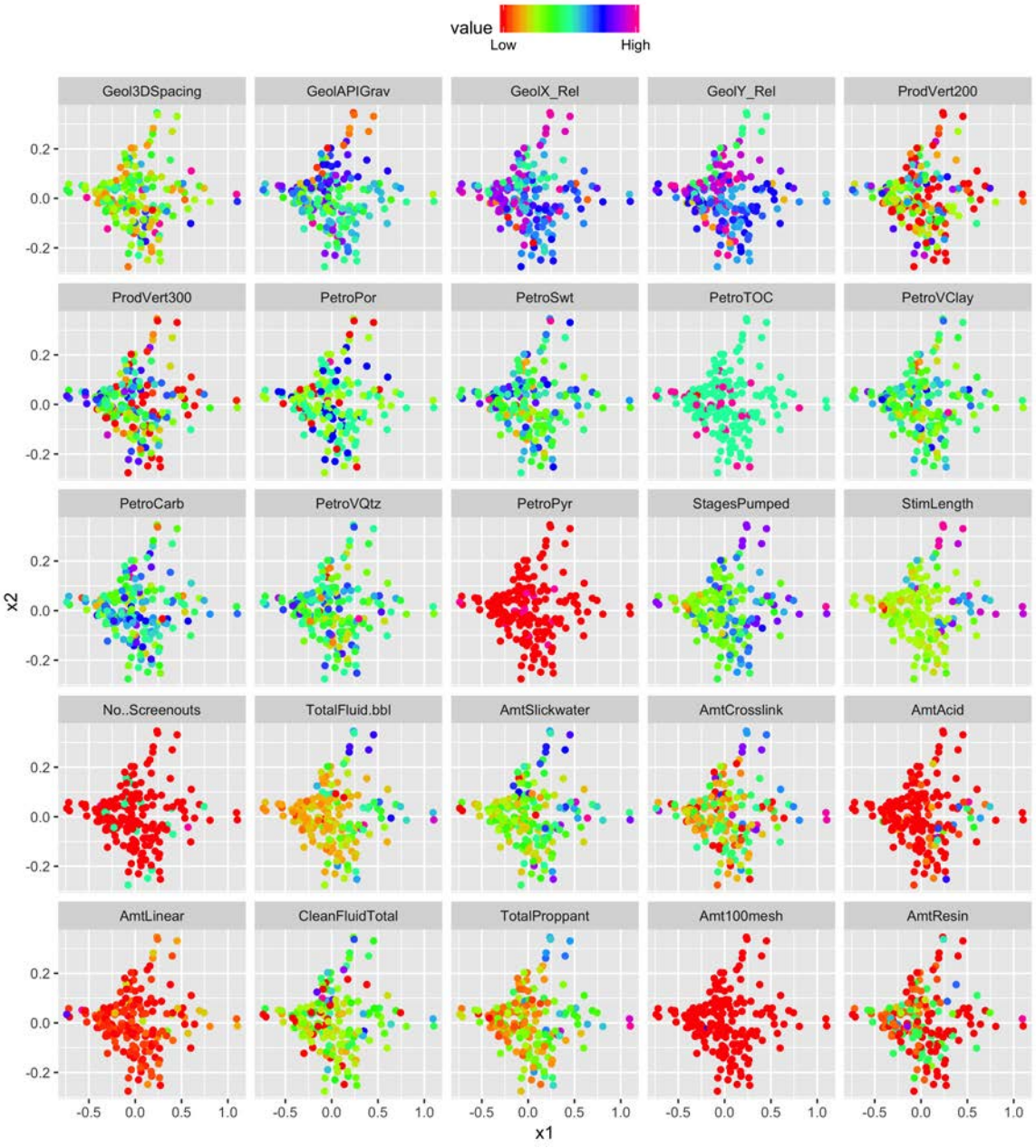


Figure 5.7: *Low dimensional scatter plots (MDS) based on the joint distance matrix and colored by each input parameter*

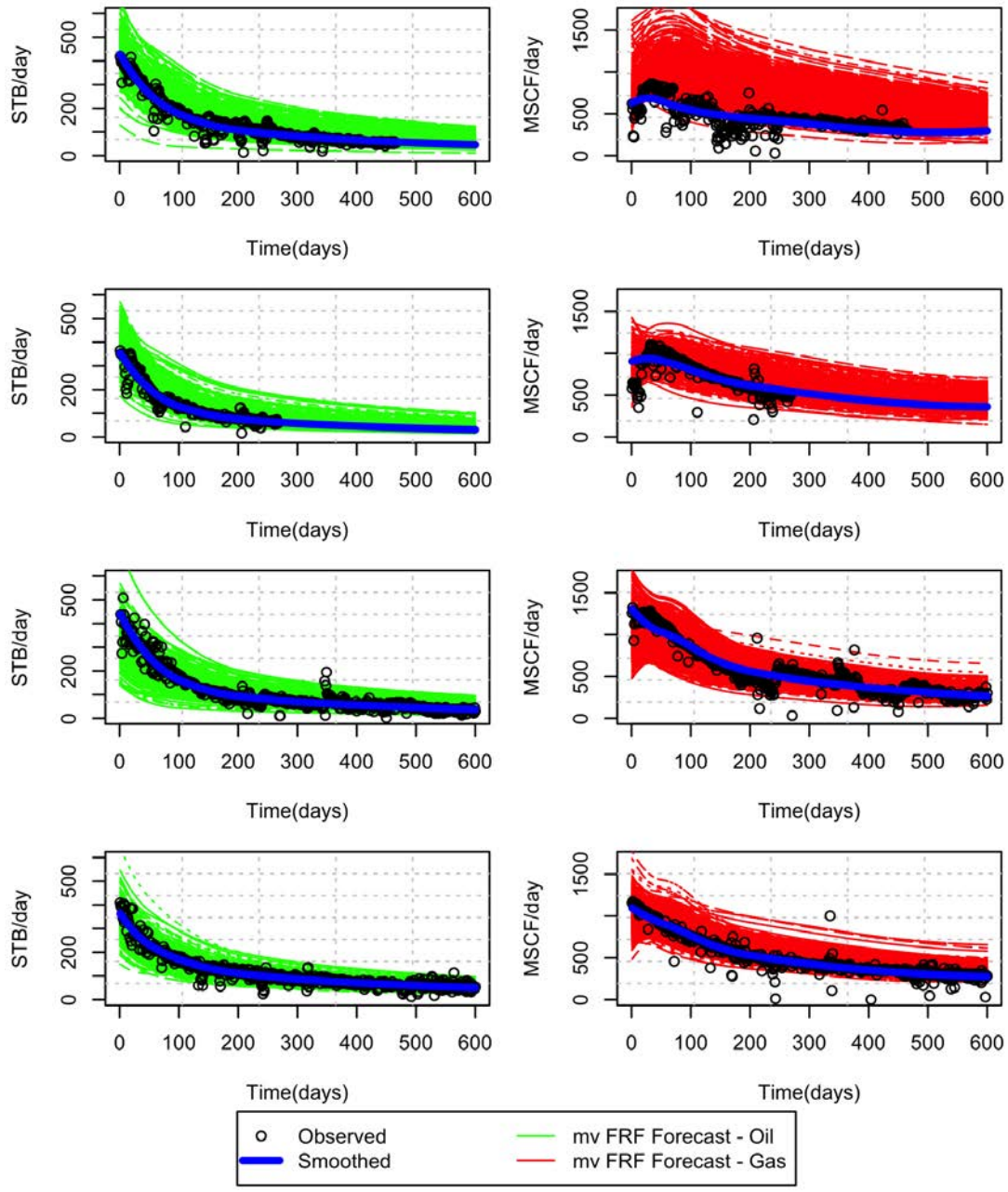


Figure 5.8: *A few forecasts produced with random forest. Left column are oil rates, right column are corresponding gas rates. "mv FRF" = multi-variate functional random forest*

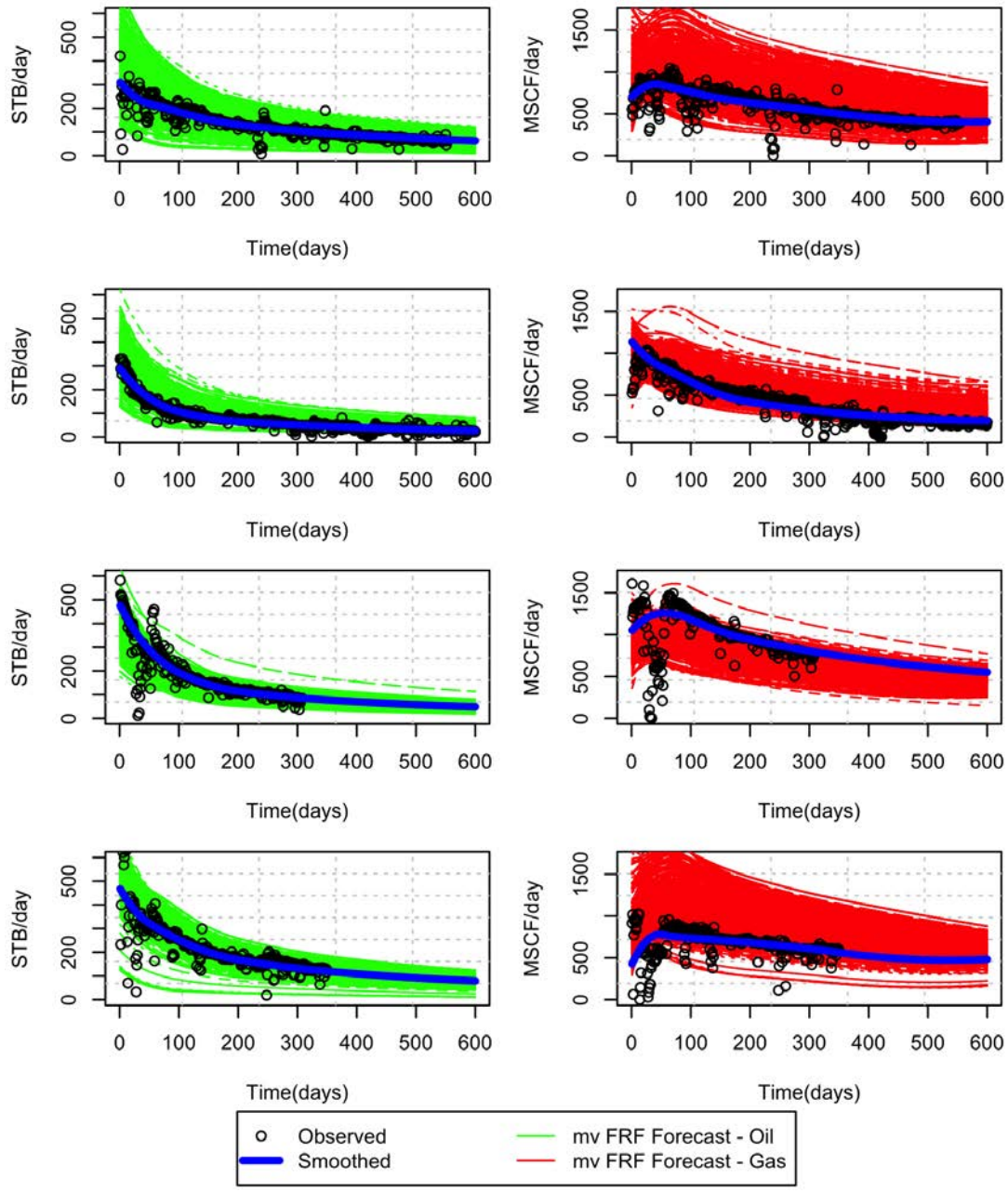


Figure 5.9: *A few forecasts produced with random forest. Left column are oil rates, right column are corresponding gas rates. "mv FRF" = multi-variate functional random forest*

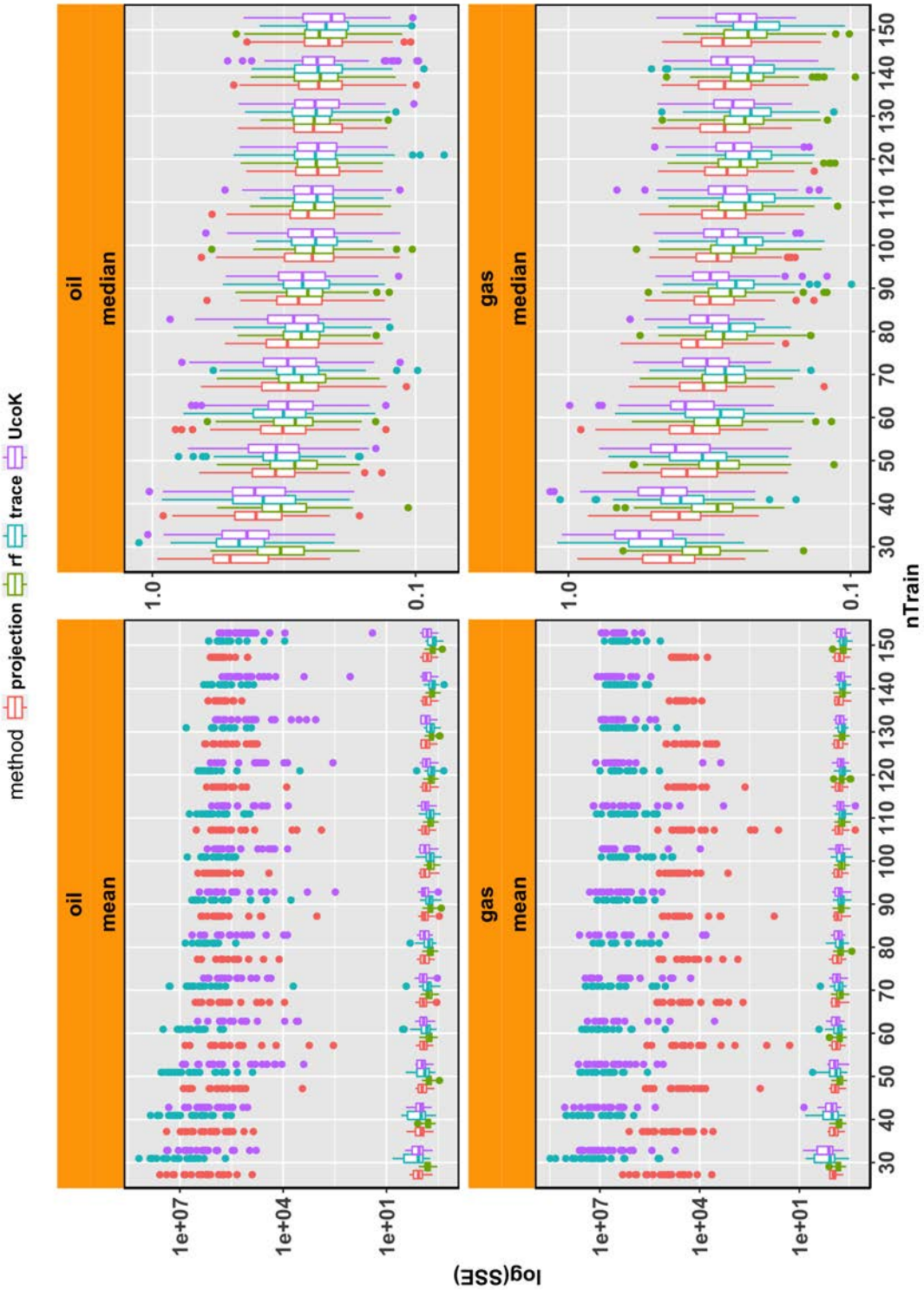


Figure 5.10: *The results of the Monte Carlo study. Abbreviations: "trace" = Universal Trace Kriging, "rf" - Multivariate random forest, "UcoK" - Cokriging of fpc scores, "projection" - joint UCoK of fpc scores of oil and gas responses.*

5.3 Chapter Conclusion

In this chapter, we presented a tree-based methodology for the analysis and forecasting of unconventional reservoir production curves. We proposed and demonstrated a robust approach for growing regression trees with similarity distances, that can accommodate any number and type of outputs. From a data analysis point of view, the method was found to produce similar variable importance results as distance based generalized sensitivity analysis (Fenwick et al. [2014]). Visual representations of regression trees also provide a deeper insight into parameter interactions, however, at this stage of research we did not make any attempts to quantify the importance of parameter interactions. This problem is left to be addressed in our future work. In addition, we also proposed a novel approach for growing regression trees with functional inputs. This proposition opens completely new avenues of research in sensitivity analysis of numerical reservoir models that commonly take probability density functions as inputs[14]. This topic will also be explored in our future work.

From a production data forecasting point of view, the method was found to produce reliable forecasts of oil and gas production curves simultaneously. In our forecasting study, we found that the prediction bands contained the true data in 94% of oil forecasts and 89% of gas forecasts. This result is very impressive given that in oil and gas industry forecasting is always conducted for the purpose of uncertainty quantification and decision making, hence, reliable prediction bands are of paramount importance.

The method also achieved a much smaller SSE error than the methods presented in the previous chapters. We observed a much smaller modeling effort compared to the previously proposed approaches. The tree-based approaches do not have numerical issues, they do not require standardization (rescaling) of input parameters and, with the developments presented in this chapter, they can work with pretty much any type of input parameters.

In our case study, we failed to obtain a meaningful variogram structure on the residuals of bagged functional regression trees. While this is unfortunate from a demonstration point of view, we still believe that the method is solid and that it may produce a better variogram on some other dataset. Identification of the exact reasons

[14] i.e. porosity distribution curves that are provided as inputs to sequential Gaussian simulation algorithm.

for nugget variograms will be in the focus of our future research work.

Finally, another source of uncertainty that surrounds all of the presented methods is the uncertainty in data smoothing. Data smoothing approaches presented in chapter 2 also provide us with covariance matrices of basis coefficients that define uncertainty of the fits. The regression tree bootstrapping paradigm allows us to easily incorporate such uncertainties into the forecasting framework presented in this chapter. When performing bootstrapping of the training data one would first sample observations and then also sample potential fits of one observation based on the covariance matrix of its basis expansion coefficients. Rigorous evaluation of this approach is also left for future work.

Chapter 6

Forecasting of Spatially Correlated Functional Data in Presence of Secondary Data

Numerical simulation and modeling is an irreplaceable component of all modern uncertainty quantification studies. Numerical models used in these studies are featured with high dimensional inputs and often produce multidimensional outputs of various types (scalars, functions, images, etc). Proper uncertainty quantification with such complex models entails exhaustive exploration of high-dimensional input spaces that is rarely achievable in practice due to extremely large computational requirements. For this reason, modelers often use computationally cheaper solutions by either ignoring certain physical aspects (lower fidelity) in their numerical models or by using statistical learning to build computer code emulators[1].

One of the most popular methods for emulation of computer experiments with scalar outputs is kriging for computer experiments (Sacks et al. [1989],Rasmussen and Williams [2006]). The method generalizes the concept of kriging to high-dimensional input spaces and it exactly reproduces the training data which is a desirable feature in this application. Another interesting generalization of geostatistical concepts comes from Kennedy and O'Hagan [2000] who applied co-kriging to aggregate information from numerical models of different levels of fidelity (multi-fidelity).

[1]This chapter as a whole was submitted and accepted for publication in the journal of Stochastic Environmental Research and Risk Assessment. Reference: Grujic et al. [2017]

Emulation of computer experiments that produce functional outputs is currently an active area of research. Josset et al. [2015] used functional regression to model functional errors between computer models of different levels of fidelity. Bottazzi and Della Rossa [2017] used functional interpolation by Nerini et al. [2010] to interpolate functional data over multidimensional input spaces. Thenon et al. [2016] also uses the functional interpolation approach by Nerini et al. [2010] in the context of functional multifidelity reservoir modeling. Trehan et al. [2017] constructs a functional error model with piece-wise regression.

Inspired by the work by Kennedy and O'Hagan [2000], in this chapter, we develop universal trace co-kriging, a novel method for interpolation of multivariate functional data that is applicable to emulation of computer codes of multiple levels of fidelity (multi-fidelity). The method is an extension of the universal trace kriging methodology (introduced in chapter 3) into multivariate context. In addition to these developments, we present a projection-based approach for interpolation of multivariate functional data which is an extension of the universal co-kriging of basis coefficients approach we developed in chapter 3. Besides the theoretical developments, we present detailed practical and methodological comparisons with the methods presented in chapter 3 on synthetic (oil reservoir) and real (Uranium contamination) numerical reservoir simulation case studies.

The chapter is organized as follows. In section 2 we present the theoretical developments of the universal trace co-kriging methodology followed by an outline of projection based approach. In section 3, we present the results of a synthetic reservoir case study, while in section 4, we present the results of uranium contamination case study. The chapter ends with conclusions based on extensive Monte Carlo analyses and ideas for future research.

6.1 A Trace-Cokriging Predictor for Multivariate Functional Data

Here, we consider a multivariate random process $\{\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}, \boldsymbol{s} \in D \subset R^n\}$ where each element $\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}$ is a vector of K random functional elements $\mathcal{X}^{(1)}_{\boldsymbol{s}_1}(t), ..., \mathcal{X}^{(K)}_{\boldsymbol{s}_n}(t)$ defined

on the same temporal domain ($t \in T$)[2]:

$$\boldsymbol{\mathcal{X}}_{\boldsymbol{s}} = \left(\mathcal{X}_{\boldsymbol{s}}^{(1)}(t), ..., \mathcal{X}_{\boldsymbol{s}}^{(K)}(t)\right)^T.$$

We call $\boldsymbol{m}_{\boldsymbol{s}}$ the spatial drift of the process at $\boldsymbol{s}$ in D, that is

$$\boldsymbol{m}_{\boldsymbol{s}} = \mathbb{E}[\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}] = \left(m_{\boldsymbol{s}}^{(1)}(t), ..., m_{\boldsymbol{s}}^{(K)}(t)\right)^T, \quad m_{\boldsymbol{s}}^{(k)}(t) = \mathbb{E}[\mathcal{X}_{\boldsymbol{s}}^{(k)}(t)].$$

To define a measure of multivariate spatial dependence, we generalize to the multivariate setting the concept of trace-covariogram previously presented in chapter 3.

$$Cov_t(\mathcal{X}_{\boldsymbol{s}}^{(k)}(t), \mathcal{X}_{\boldsymbol{u}}^{(l)}(t)) \quad = \quad \mathbb{E}\left[\int_T (\mathcal{X}_{\boldsymbol{s}}^{(k)}(t) - m_{\boldsymbol{s}}^{(k)}(t))(\mathcal{X}_{\boldsymbol{u}}^{(l)}(t) - m_{\boldsymbol{u}}^{(l)}(t))dt\right].$$

Note that this quantity cannot be defined in cases where components of multivariate functional process are defined on different temporal domains.

In this work, we assume that every element $\mathcal{X}_{\boldsymbol{s}}^{(k)}$ of the multivariate process $\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}$ is non-stationary, and that it can be represented by a sum of deterministic drift and zero-mean globally second-order stationary residual:

$$\mathcal{X}_{\boldsymbol{s}}^{(k)}(t) = m_{\boldsymbol{s}}^{(k)}(t) + \delta_{\boldsymbol{s}}^{(k)}(t) \tag{6.1}$$

here, the drift is assumed to be non-constant in space D and, analogously to universal trace kriging, modeled with a functional linear model:

$$m_{\boldsymbol{s}}^{(k)}(t) = \sum_{l=0}^{L} a_l^{(k)}(t) f_l(\boldsymbol{s}) \tag{6.2}$$

where $a_l^{(k)}(t)$ are functional coefficients in, and $f_l(.)$ are scalar regressors known over the entire domain D. Further, the residual is assumed to be globally second-order stationary in the sense of Menafoglio et al. [2013]. That is, we assume that the multivariate trace-covariogram structure depends only on the increment between locations, i.e. $C_{kl}(\|\boldsymbol{s} - \boldsymbol{u}\|) = Cov_t(\mathcal{X}_{\boldsymbol{s}}^{(k)}(t), \mathcal{X}_{\boldsymbol{u}}^{(l)}(t))$, for all $\boldsymbol{s}, \boldsymbol{u} \in D$.

[2]An example of such data are field oil production curves computed with K reservoir simulators, of different levels of fidelity, with the same set of input parameters.

We call $\boldsymbol{s}_1, ..., \boldsymbol{s}_{N_j}$ $(j = 1, ..., K)$ the measurement locations (or *design of experiment*), and $\mathcal{X}^{(j)}_{\boldsymbol{s}_1}(t), ..., \mathcal{X}^{(j)}_{\boldsymbol{s}_{N_j}}(t)$ the partial observation of the j-th element of the multivariate process at these locations. Within the former assumptions, we aim to predict the k-th element $\mathcal{X}^{(k)}_{\boldsymbol{s}_0}(t)$ of $\boldsymbol{\mathcal{X}}_{\boldsymbol{s}_0}$ at a target location $\boldsymbol{s}_0$ in D. To this end, we consider the trace-cokriging predictor, that is the best linear unbiased predictor within the class of linear predictors

$$\mathcal{X}^{(k)\lambda}_{\boldsymbol{s}_0}(t) = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \lambda_{ji} \mathcal{X}^{(j)}_{\boldsymbol{s}_i}(t) \tag{6.3}$$

To find the optimal weights, λ^*_{ji}, $j = 1, ..., K$, $i = 1, ..., N_j$, we minimize the mean squared error of prediction under the unbiasedness constraint, that is

$$\begin{aligned} \min_{\substack{\lambda_{ji} \in \mathbb{R}, \\ j=1,...,K, i=1,...,N_j}} \quad & \mathbb{E}\left[\|\mathcal{X}^{(k)\lambda}_{\boldsymbol{s_0}}(t) - \mathcal{X}^{(k)}_{\boldsymbol{s_0}}(t)\|^2\right] \\ \text{subject to} \quad & \mathbb{E}[\mathcal{X}^{(k)\lambda}_{\boldsymbol{s_0}}(t)] = m^{(k)}_{\boldsymbol{s_0}}(t). \end{aligned} \tag{6.4}$$

It is straightforward to see that the unbiasedness constraint reads as

$$\begin{aligned} &\sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\ &\sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{s}_i) = 0, \quad \text{for} \quad j \neq k, \quad \forall l; \end{aligned} \tag{6.5}$$

Therefore,

$$\mathbb{E}[\mathcal{X}^{(k)\lambda}_{\boldsymbol{s_0}}(t)] = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \lambda_{ji} m^{(j)}_{\boldsymbol{s_i}}(t)$$

and the latter quantity is equal to $m^{(j)}_{\boldsymbol{s_0}}(t)$ if and only if the condition 6.5 is fulfilled.

Developing the first line of Eq. 6.4 yields:

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{X}_{\boldsymbol{s}}^{(k)\lambda}(t)-\mathcal{X}_{\boldsymbol{s}}^{(k)}(t)\|^2\right] = C_{kk}(\mathbf{0})+\\ \sum_{j=1}^{K}\sum_{i=1}^{N_j}\sum_{j'=1}^{K}\sum_{i'=1}^{N_{j'}}\lambda_{ji}\lambda_{j'i'}C_{jj'}(\boldsymbol{s}_i-\boldsymbol{s}_{i'})\\ -2\sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jk}(\boldsymbol{s}_i-\boldsymbol{s}_0)\end{aligned} \tag{6.6}$$

Introducing $K\times(L+1)$ Lagrange multipliers to account for the unbiasedness constraints in Eq. 6.5 leads to the following objective functional

$$\begin{aligned}\Phi(\lambda) = C_{kk}(\mathbf{0})+\sum_{j=1}^{K}\sum_{i=1}^{N_j}\sum_{j'=1}^{K}\sum_{i'=1}^{N_{j'}}\lambda_{ji}\lambda_{j'i'}C_{jj'}(\boldsymbol{s}_i-\boldsymbol{s}_{i'})-\\ 2\sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jk}(\boldsymbol{s}_i-\boldsymbol{s}_0)+\\ 2\sum_{l=0}^{L}\eta_{kl}\left(\sum_{i=1}^{N_k}\lambda_{ki}f_l(\boldsymbol{s}_i)-f_l(\boldsymbol{s}_0)\right)+\\ 2\sum_{l=0}^{L}\sum_{\substack{j=1\\ j\neq k}}^{K}\eta_{jl}\left(\sum_{i=1}^{N_j}\lambda_{ji}f_l(\boldsymbol{s}_i)\right)\end{aligned} \tag{6.7}$$

After taking partial derivatives of equation 6.7 with respect to λ's and η's we arrive at the following system of linear equations:

$$\begin{aligned}\sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jj'}(\boldsymbol{s}_i-\boldsymbol{s}_{i'})+\sum_{l=0}^{L}\eta_{j'l}f_l(\boldsymbol{s}_i) = C_{j'k}(\boldsymbol{s}_{i'}-\boldsymbol{s}_0),\\ (j'=1,...,K; i'=1,...,N_{j'};);\\ \sum_{i=1}^{N_k}\lambda_{ki}f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0),\quad \forall l;\\ \sum_{i=1}^{N_j}\lambda_{ji}f_l(\boldsymbol{s}_i) = 0,\quad j\neq k,\quad \forall l;\end{aligned} \tag{6.8}$$

The trace-variance associated with predictor $\mathcal{X}_{\boldsymbol{s}_0}^{(k)*} = \sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}^{*}\mathcal{X}_{\boldsymbol{s}_i}^{(j)}$ is given by

$$\sigma_k^2(\boldsymbol{s}_0) = C_{kk}(\mathbf{0}) - \sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jk}(\boldsymbol{s}_i - \boldsymbol{s}_0) + \sum_{l=0}^{L}\eta_{kl}f_l(\boldsymbol{s}_0).$$

System (6.8) can be expressed in a matrix form as follows (for $k = 1$):

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1K} & \boldsymbol{F}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2K} & \mathbf{0} & \boldsymbol{F}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K1} & \boldsymbol{C}_{K2} & \cdots & \boldsymbol{C}_{KK} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F}_K \\ \boldsymbol{F}_1^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{F}_2^T & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F}_K^T & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_K \\ \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \\ \vdots \\ \boldsymbol{\eta}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{10} \\ \boldsymbol{c}_{20} \\ \vdots \\ \boldsymbol{c}_{K0} \\ \boldsymbol{f}_{01} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{6.9}$$

where:

$$[\boldsymbol{C}_{mn}]_{ij} = Cov_t\left(\mathcal{X}_{\boldsymbol{s}_i}^{(m)}, \mathcal{X}_{\boldsymbol{s}_j}^{(n)}\right) = C_{mn}(\boldsymbol{s}_i - \boldsymbol{s}_j)$$

$$\boldsymbol{c}_{j0} = \begin{bmatrix} C_{jk}(\|\boldsymbol{s}_1 - \boldsymbol{s}_0\|) \\ C_{jk}(\|\boldsymbol{s}_2 - \boldsymbol{s}_0\|) \\ \vdots \\ C_{jk}(\|\boldsymbol{s}_{N_j} - \boldsymbol{s}_0\|) \end{bmatrix}, \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \vdots \\ \lambda_{jN_j} \end{bmatrix}, \boldsymbol{\eta}_j = \begin{bmatrix} \eta_{j0} \\ \eta_{j1} \\ \vdots \\ \eta_{jd} \end{bmatrix},$$

$$\boldsymbol{F}_j = \begin{bmatrix} f_0(\boldsymbol{s}_1) & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ f_0(\boldsymbol{s}_2) & f_1(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_0(\boldsymbol{s}_{N_j}) & f_1(\boldsymbol{s}_{N_j}) & \cdots & f_L(\boldsymbol{s}_{N_j}) \end{bmatrix}, \boldsymbol{f}_{0j} = \begin{bmatrix} f_0(\boldsymbol{s}_0) \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{bmatrix}.$$

The system given in equation (6.9) is analogous to the system of universal co-kriging equations outlined in chapter 3 and in Chiles and Delfiner [1999].

Parameter inference is analogous to the parameter inference in conventional co-kriging. First, functional regression (Ramsay and Silverman [2005]) is used to compute the functional drift of each of the elements of multivariate functional data,

then the estimates of the trace-auto and trace-cross covariances are computed on the functional residuals and admissible covariance structures are fitted with the linear model of coregionalization (LMC, Goovaerts [1997]).

Auto-covariance estimation is performed simply by means of trace-variography Giraldo [2009], Menafoglio et al. [2013]. Recall the trace variogram estimator presented in chapter 3

$$\gamma_{k,k}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \int_T \left(\mathcal{X}^{(k)}_{\boldsymbol{s}_i}(t) - \mathcal{X}^{(k)}_{\boldsymbol{s}_j}(t) \right)^2 dt \tag{6.10}$$

where N(h) denotes the set of pairs (i,j) such that $h - \Delta h \leq \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \leq h + \Delta h$. To find the cross-covariance estimators we proceed by analogy with multivariate geostatistics by generalizing the well known cross-variogram (Goovaerts [1997]) and pseudo cross-variogram (Clark et al. [1987]) estimators:

1. The trace cross-variogram estimator:

$$\gamma_{k,l}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \int_T \left(\mathcal{X}^{(k)}_{\boldsymbol{s}_i}(t) - \mathcal{X}^{(k)}_{\boldsymbol{s}_j}(t) \right) \left(\mathcal{X}^{(l)}_{\boldsymbol{s}_i}(t) - \mathcal{X}^{(l)}_{\boldsymbol{s}_j}(t) \right) dt \tag{6.11}$$

2. The pseudo trace-cross-variogram estimator:

$$\gamma_{k,l}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \int_T \left(\mathcal{X}^{(k)}_{\boldsymbol{s}_i}(t) - \mathcal{X}^{(l)}_{\boldsymbol{s}_j}(t) \right)^2 dt \tag{6.12}$$

The properties of the trace cross-variograms are the same as their scalar counterparts. The pseudo trace cross-variogram is always positive and applicable to both isotopic and heterotopic data sampling, while the trace cross-variogram is only applicable in the case of isotopic data sampling (Wackernagel [2010]). In practice, inference and fitting of trace variograms over high dimensional input spaces is limited to omni-directional variograms. This is mainly due to difficulties with unidirectional (marginal) variogram estimation in high dimension that is necessary for product and sum covariance structures (De Cesare et al. [2001]).

Currently, the method of moments is the only possible parameter inference procedure in trace-cokriging. This is because the concept of density for functional data is

not mathematically defined (Delaigle and Hall [2010]) making it difficult to formulate any type of automated maximum likelihood-based parameter inference procedure.

The range of applicability. As mentioned previously, the temporal domain over which the elements of the vector $\boldsymbol{\mathcal{X}}_s$ are defined must be coincident in order to compute the trace cross-covariances. In multi-fidelity modeling, this is almost always the case since low-fidelity simulations produce the same type of output data as their high-fidelity counterparts. Another requirement for this modeling strategy to work is that the discrepancies between functional data be mostly in amplitude rather than in phase (figure 6.1 A.). Interpolation of phase shifted functional data is much more complex and it would require modeling with warping functions (Ramsay and Li [1998]) that is beyond the scope of the presented work.

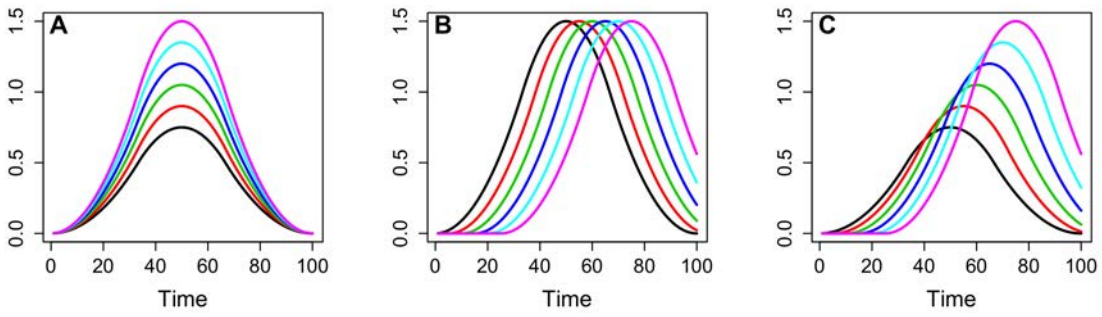


Figure 6.1: ***A*** *- Amplitude shifted ensemble of functions.* ***B*** *- Phase shifted ensemble of functions.* ***C*** *- Phase-amplitude shifted ensemble of functions*

6.2 Projection Based Interpolation of Multivariate Functional Data

Alternatively, one can approach the problem of interpolation of multivariate functional data from a projection perspective. The idea is to assume K basis systems, one for each level of multivariate functional data, and then use the coefficients of basis expansion in combination with multivariate geostatistics to forecast new functions at

some new locations[3].

For instance, consider a sample of bi-variate (2-levels) functional data $\boldsymbol{\mathcal{X}}_{\boldsymbol{s_i}} = \left(\mathcal{X}^{(1)}_{\boldsymbol{s_i}}, \mathcal{X}^{(2)}_{\boldsymbol{s_i}}\right)^T$ fully observed over a set of design points $\boldsymbol{s}_i$ where $i = 1, 2, ..., N$ and where $\boldsymbol{s}$ is a vector in $\mathbb{R}^d$. Let $e^{(1)} = \{\phi^{(1)}_1, \phi^{(1)}_2, ..., \phi^{(1)}_P\}$ and $e^{(2)} = \{\phi^{(2)}_1, \phi^{(2)}_2, ..., \phi^{(2)}_Q\}$ be ortho-normal sets of functional principal components of $\mathcal{X}^{(1)}_{\boldsymbol{s_i}}$'s and $\mathcal{X}^{(2)}_{\boldsymbol{s_i}}$'s, respectively, and let $\xi^{p(k)}_{\boldsymbol{s_i}}$ be a principal component score of $\mathcal{X}^{(k)}_{\boldsymbol{s_i}}$ on $\phi^{(k)}_p$. As in the previous section, we assume that each of the two elements of multivariate data is non-stationary and that it can be decomposed into deterministic mean and globally second order stationary residual

$$\mathcal{X}^{(k)}_{\boldsymbol{s}}(t) = m^{(k)}_{\boldsymbol{s}}(t) + \delta^{(k)}_{\boldsymbol{s}}(t) \tag{6.13}$$

We have shown in chapter 3 that interpolation of non-stationarity in functional data translates into non-stationarity of fpc scores (or basis coefficients):

$$\begin{aligned} \xi^{p(k)}_{\boldsymbol{s_i}} &= \hat{m}^{p(k)}_{\boldsymbol{s}_i} + r^{(k)}_{\boldsymbol{s}_i}; \\ \hat{m}^{p(k)}_{\boldsymbol{s}_i} &= \sum_{l=0}^{d} \beta_l f_l(\boldsymbol{s}), \quad \beta_l \in \mathbb{R}. \end{aligned} \tag{6.14}$$

The new function of k-th level is estimated by forecasting its fpc scores with a linear combination of fpc scores of all already observed functions (across all levels):

$$\xi^{p(k)}_{\boldsymbol{s_0}} = \sum_{i=1}^{N}\sum_{p=1}^{P} \lambda^{(1)}_{i,p}\xi^{p(1)}_{\boldsymbol{s_i}} + \sum_{i=1}^{N}\sum_{q=1}^{Q} \lambda^{(2)}_{i,q}\xi^{p(2)}_{\boldsymbol{s_i}}. \tag{6.15}$$

This is a co-kriging problem that is analogous to the single variate projection based interpolation outlined in chapter 3. Hence, the weights $\lambda^{(k)}_{p,q}$ are found by solving the well-known system of universal co-kriging equations (Chiles and Delfiner [1999], pg. 300):

[3]It is not a huge intellectual leap to develop this extension after the developments presented in chapter 3. As a matter of fact Bohorquez et al. [2016] had the same idea however in different context.

$$\begin{bmatrix} \boldsymbol{C}_{11}^{11} & \boldsymbol{C}_{12}^{11} & \boldsymbol{C}_{11}^{12} & \boldsymbol{C}_{12}^{12} & \boldsymbol{F}_1^1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}_{21}^{11} & \boldsymbol{C}_{22}^{11} & \boldsymbol{C}_{21}^{12} & \boldsymbol{C}_{22}^{12} & \mathbf{0} & \boldsymbol{F}_2^1 & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}_{11}^{21} & \boldsymbol{C}_{12}^{21} & \boldsymbol{C}_{11}^{22} & \boldsymbol{C}_{12}^{22} & \mathbf{0} & \mathbf{0} & \boldsymbol{F}_2^1 & \mathbf{0} \\ \boldsymbol{C}_{21}^{21} & \boldsymbol{C}_{22}^{21} & \boldsymbol{C}_{21}^{22} & \boldsymbol{C}_{22}^{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{F}_2^2 \\ (\boldsymbol{F}_1^1)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{F}_2^1)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{F}_1^2)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (\boldsymbol{F}_2^2)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1^1 \\ \boldsymbol{\lambda}_2^1 \\ \boldsymbol{\lambda}_1^2 \\ \boldsymbol{\lambda}_2^2 \\ \boldsymbol{\sigma}_1 \\ \boldsymbol{\sigma}_2 \\ \boldsymbol{\sigma}_3 \\ \boldsymbol{\sigma}_4 \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_0^{11} \\ \boldsymbol{c}_0^{21} \\ \boldsymbol{c}_0^{12} \\ \boldsymbol{c}_0^{22} \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Where $[\boldsymbol{C}_{pq}^{lm}]_{ij} = Cov(\xi_{\boldsymbol{s_i}}^{pl}, \xi_{\boldsymbol{s_j}}^{qm}) = C_{pq}^{lm}(\boldsymbol{s_i} - \boldsymbol{s_k})$, and $\boldsymbol{F}_p^l(i,.) = \{f_0(\boldsymbol{s}_i), ..., f_L(\boldsymbol{s}_i,)\}$.

Parameter inference. Given that this approach effectively transforms a multivariate functional interpolation problem into a multivariate (vector) interpolation problem, many parameter inference procedures developed in multivariate geostatistics are available. Both variogram fitting procedures with the linear model of coregionalization (LMC, Goovaerts [1997]), as well as automated maximum likelihood approaches are applicable (Gelfand et al. [2004], Fricker et al. [2013], Zhang [2007]). The size of the model depends on the size of the training dataset and the number of kept functional principal components on every level of multivariate functional data. Bohorquez et al. [2016] reported numerical difficulties with the linear model of coregionalization for large numbers of kept principal components.

The range of applicability. The projection-based approach is applicable to a variety of modeling situations. The method is not limited to amplitude shifted curves, instead, it can work with phase, amplitude and phase-amplitude shifted ensembles (figure 6.1). The only consequence that higher complexity in functional data could have on this method is the increase of the dimensionality of the model that ultimately affects parameter inference[4]. One attractive feature of this approach is that it does not require that all levels of multivariate data be defined on the same temporal domain. What's more the components of multivariate data do not even need to be functional. The method can work equally well with proxies that produce outputs of different type (i.e. flow diagnostics (Shahvali et al. [2012]), image processing based proxies, etc.).

[4]The higher the complexity in functional data, the larger is the number of fpcs needed to adequately describe such data.

6.3 Performance Analysis on Synthetic Datasets

In this section, we set out to explore and assess the performance of the previously presented emulation techniques on a purely synthetic numerical model of the subsurface. For this purpose, we developed a homogeneous 3D oil-water reservoir model with 4 producer wells at the top of the reservoir structure, and an aquifer connected at the bottom left corner for pressure support (Figure 6.2 left). The four wells produce two types of fluid, oil and water. Initially, the reservoir is saturated with oil and wells do not produce any water until the reservoir pressure becomes low enough to allow water encroachment from the aquifer. The speed of encroachment is dependent on the reservoir properties and the viscosity of the present fluids. One typical field water production rate (FWPR) response is given in Figure 6.2 right, while the model parameters that are the most influential on FWPR are summarized in Table 6.1.

Given that all of the presented computer code emulation techniques aim to make use of both computationally expensive (high fidelity), and computationally cheap (low fidelity) simulations two levels of numerical abstractions were considered. High fidelity flow simulations, were computed on a finely gridded reservoir volume (150x100x25), while the low fidelity flow simulations were computed on a coarsely gridded reservoir volume (150x100x13). The two solutions produce somewhat different, but highly correlated ($\rho = 0.91$) flow responses (Figure 6.2 - right).

We used the model to develop two datasets for methodological comparisons and assessment. The first dataset considered only two input parameters, PERMZm and PORVm (Table 1). The second dataset considered three input parameters: PERMZm, PORVm, and PERM (Table 1). Both datasets consist of training and testing subsets. The training subsets were produced by latin hypercube sampling and were evaluated with both high and low fidelity flow simulations. The test sets were produced with uniform sampling and were evaluated only with the high fidelity flow solution (the target response). The two datasets are summarized in Table 6.2.

Output data pre-processing The training ensemble of FWPR curves of the two parameter dataset is given in Figure 6.4 - right. What is obvious from this figure is that the data are shifted in both phase and amplitude. Trace based co-kriging is not directly applicable in this case since the method can only work with amplitude shifted data. However, the ensemble of phase-amplitude shifted FWPR curves can

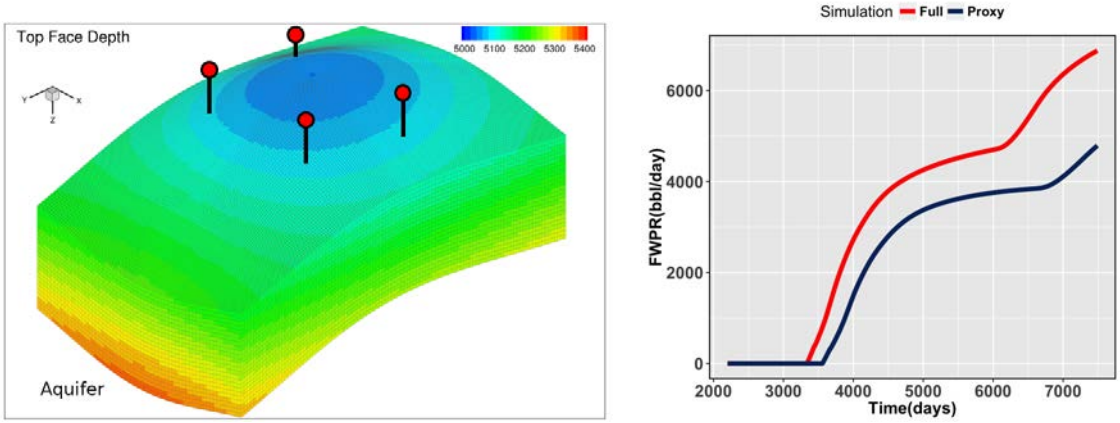


Figure 6.2: *Left - 3D reservoir model; Right - An example of proxy and full solutions.*

Table 6.1: *Simulation parameters*

Parameter	Value	Description
PORVm (-)	1-1000	Aquifer Strength
PERMZm (-)	0-1	Vertical Perm. (K mult.)
K (md)	25	Reservoir permeability
ϕ (frac)	0.2	Reservoir porosity
μ_o (cp)	0.0002	Oil Viscosity
μ_w (cp)	0.00001	Water viscosity

be transformed into an ensemble of amplitude-shifted curves with a simple ad-hoc procedure. For one curve, the procedure consists of identification of the water breakthrough time, followed by a simple regression fit to the early post water breaktrhough rates and substitution of the zero production rates with regressions solution. This procedure is explained visually in Figure 6.3.

6.3.1 Analysis: Computer Experiment with Two Parameters

Our first analysis focuses on the two parameter dataset. In this exercise, only a portion of the available training data was used, namely 50 high fidelity flow (fine) simulations and 150 low fidelity simulations (proxy). The sub-sampled training dataset was used to fit the following models: Universal Trace Kriging (**UTrK**) by Menafoglio

Table 6.2: *Summary of the produced datasets*

Dataset type	# Proxy	# Full	# Test
2 parameter	189	176	400
3 parameter	466	462	400

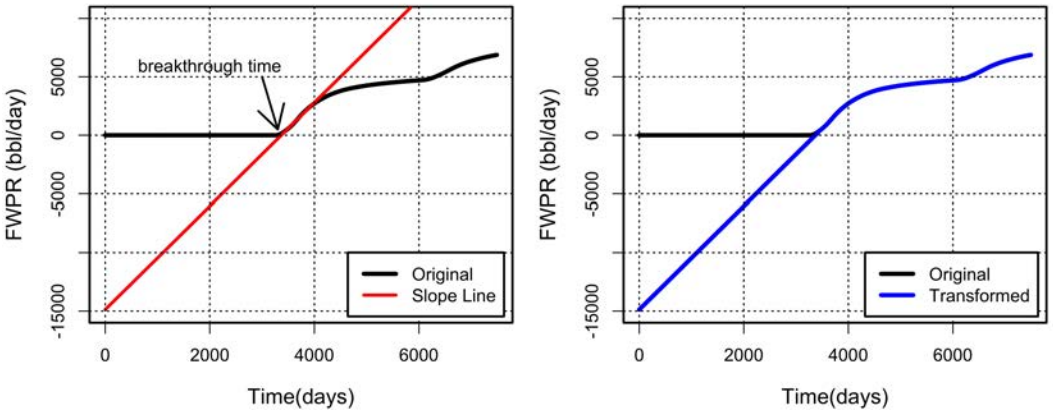


Figure 6.3: *Curve transformation procedure. Left - Original curve with a straight line fitted through the early breakthrough rates. Right - The resulting "transformed" curve.*

et al. [2013], Universal Trace Co-Kriging (**UTrCoK**) introduced in this paper, projection based interpolation for functional data (**UCoK**) by Menafoglio et al. [2016b] (chapter 3), and projection based Universal co-kriging with secondary functional data (**UCoK2**) presented previously. Note that UCoK and UTrK were fitted only on the full physics responses (i.e., without considering the low-fidelity model) since they are univariate functional interpolation methods.

Given that the projection based methods can be fitted with a variable number of principal components, we produced models with two (suffix: ".K2"), and three (suffix: ".K3") leading principal components. For parameter inference, we used variogram fitting and linear model of coregionalization (LMC, Goovaerts [1997]) on omnidirectional variograms computed over the unit cube of re-scaled input parameters[5].

The produced statistical models were then used to predict the test set (400 curves) and summarize the predictions by computing the sum of squared errors (SSE) of each

[5]This common practice was proposed by Sacks et al. [1989]

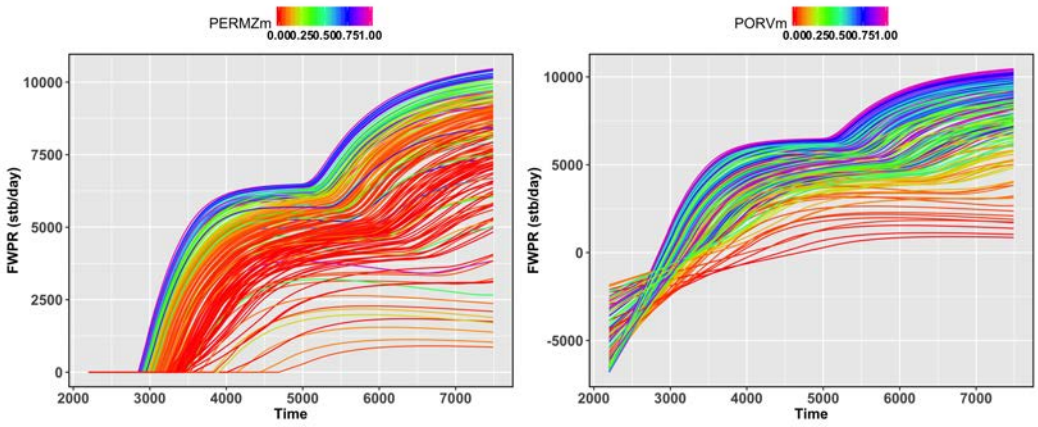


Figure 6.4: *Raw and transformed FWPR curves from 2 parameter dataset. Left - Raw curves colored by PERMZm, Right - Transformed curves colored by PORVm*

prediction.

$$SSE_i = \|\mathcal{X}_i^{(k)}(t) - \hat{\mathcal{X}}_i^{(k)}(t)\|^2 \tag{6.16}$$

To better appreciate the magnitude of the error all SSE's were normalized by the average squared norm of the entire test set (400 simulations).

$$SSE_i^n = \frac{SSE_i}{\frac{1}{400}\sum_{i=1}^{400}\|\mathcal{X}_i^{(k)T}(t) - \mu^{(k)T(t)}\|^2} \tag{6.17}$$

$\mu^{(k)T(t)}$ is the mean of the test set.

Empirical variograms of the trace based co-kriging and universal co-kriging with secondary data are given in Figures 6.5 and 6.6 along with the fits produced with LMC.

Test sets error summary is given in Table 6.3, and visually in Figure 6.8. We observe that trace based methods performed slightly better than projection based approaches, and we also observe that incorporation of the secondary data in a form of proxy solution improved the overall SSE. Examples of forecasts produced with each interpolation approach are shown in figure 6.7 for four design points.

6.3.2 Monte Carlo Analysis

To assess the performance under variable training set sizes and different ratios of full physics to proxy simulations, we set up a Monte Carlo study. For variable numbers

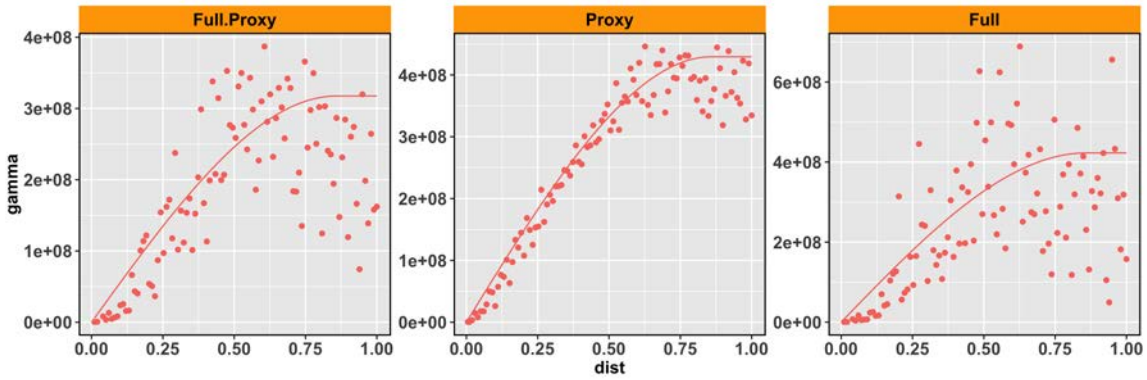


Figure 6.5: *Empirical omni-directional trace variograms and models fitted with the LMC ($Sph(\frac{d}{0.85})$). Left - trace-cross-variogram, middle and right auto trace-variograms*

Table 6.3: *2D dataset - Error Summary Table (SSE)*

	min	p0.25	p0.5	p0.75	max	mean
Projection Methods						
UcoK.K2	0.0019	0.0080	0.0205	0.0535	2.3807	0.0739
UcoK.K3	0.0004	0.0026	0.0052	0.0112	2.2712	0.0483
UcoK2.K2	0.0018	0.0088	0.0181	0.0486	0.9568	0.0482
UcoK2.K3	0.0005	0.0027	0.0049	0.0086	1.0661	0.0239
Trace Methods						
UTrCoK	0.0000	0.0001	0.0005	0.0034	0.4143	0.0175
UTrK	0.0000	0.0001	0.0007	0.0045	2.2030	0.0416

of proxy and full physics simulations we repeated the previous forecasting study one hundred times, and at each step we computed the mean and the median of the test sets SSE's. Distribution of the mean and the median of SSE for each fitting method on the two parameter dataset is shown in Figure 6.9. The same analysis was performed on the three parameter dataset and its results are shown in Figure 6.10.

We observe that the median of the SSE was consistently lower for trace-based methods compared to projection-based methods. We also observe that all methods had similar SSE for a large number of full physics simulations.

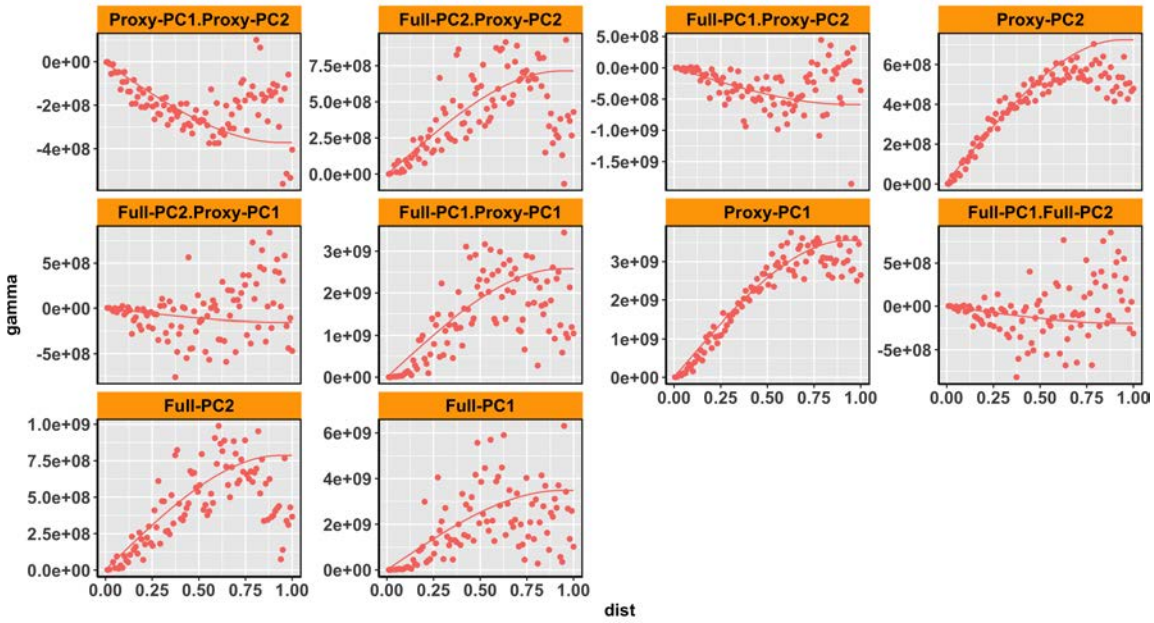


Figure 6.6: *UCoK2: Empirical auto and cross omni-directional variograms and models fitted with the LMC for K=2. (Sph(d/0.94).*

6.4 Case Study: Uranium Contamination Dataset

In this section, we apply and illustrate the presented computer code emulation techniques on a real case study. The case study considers a numerical model of uranium bio-remediation experiment in Rifle Colorado (Yabusaki et al. [2007], Li et al. [2011], Kowalsky et al. [2012]). The experiment consisted of acetate and tracer injection into eleven injection wells and monitoring their concentrations at twelve monitoring wells (Figure 6.11 left). The presence of acetate in the subsurface is known to stimulate biochemical reactions between in-situ bacteria and mobile uranium U(VI) ions (Williams et al. [2011]), producing immobile uranium U(IV) ions. Since there is no direct way of inferring the volumes of immobilized uranium, indirect inference by means of numerical simulation and inversion is necessary. In particular, spatial distributions of immobilized uranium from the numerical models that matched the measured data at monitoring wells can be used to estimate of the immobilized volumes of U(VI).

Numerical modeling of bio-remediation is difficult and computationally expensive. One has to consider both geological and geochemical uncertainties and complex

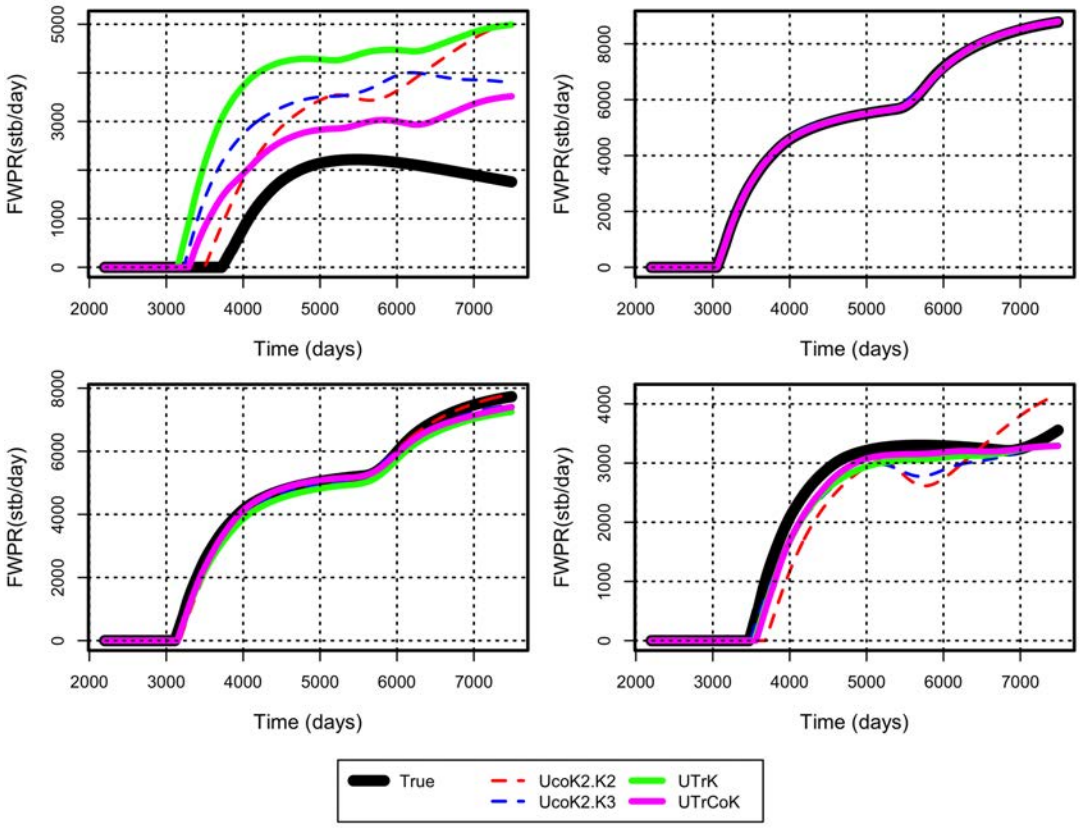


Figure 6.7: *2 parameter dataset: An example of forecasts for four randomly selected design points.*

physics need to be simulated with advanced reactive transport numerical simulators. Simulation models used in this case study were developed with Crunchflow (Steefel et al. [2015]), a reactive transport simulator. The contaminated site is an unconfined aquifer in aluvial floodplane that was modeled as a single layer with 64x68x1 grid blocks with thickness of about 2.5 meters. We used latin hypercube sampling to vary five input parameters: three geological and two geochemical. Geological parameters are: mean log permeability (meanLogK) of the reservoir, correlation length (CorrL) of reservoir permeability and the variance of reservoir permeability (varK), while geochemical parameters are kinetic rates of microbial reactions: ferric rate (FerricRate) and microbial sulfate reduction rate (SRBrate). The parameters and their ranges are summarized in Table 6.4. Geological properties were modeled with sequential

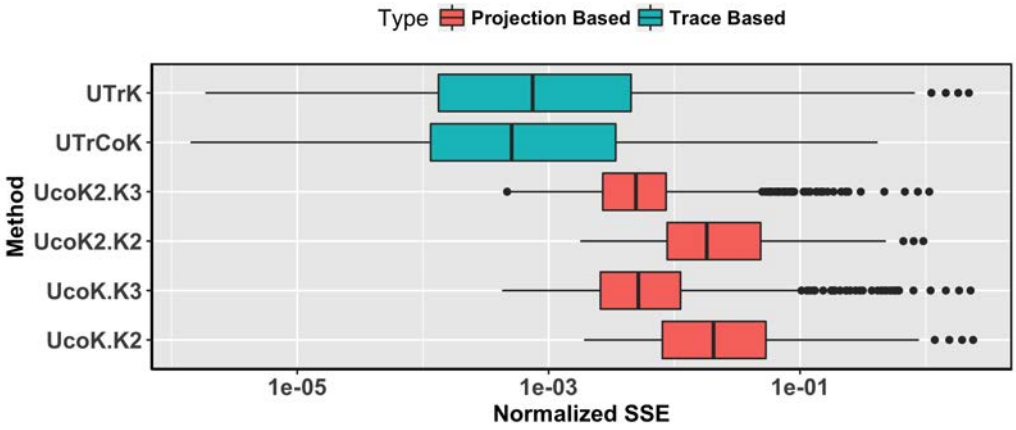


Figure 6.8: *Normalized SSE distribution of each forecasting approach.*

Gaussian co-simulation (coSGS, Verly [1992]), and a total of 500 geological models were developed. While this model is fairly small, one simulation run took around 2 hours due to high complexity of the modeled physics. To demonstrate and evaluate our computer code emulation methodology we upscaled/upgridded the models to produce proxy flow simulations. Upgridded models contained 32x34x1 grid blocks and this simplification reduced simulation time to just about 10 minutes.

Table 6.4: *Uranium contamination model parameters*

Parameter	Range
meanLogK	-10.5 to -10
CorL	3 m - 7 m
varK	0.2 - 0.7
FerricRate	1 - 2
SRBRate	0 - 2

In our analysis, we considered simulated acetate concentration curves from monitoring well number 11 (Figure 6.12). With this data we conducted the same type of Monte Carlo study as we did before on the synthetic reservoir model. The only difference was that in this case we did not have a fixed test set, instead at every iteration we randomly sampled for variable numbers of proxy and full physics reservoir models and a non overlapping test set of size 100. In all models, we used variogram

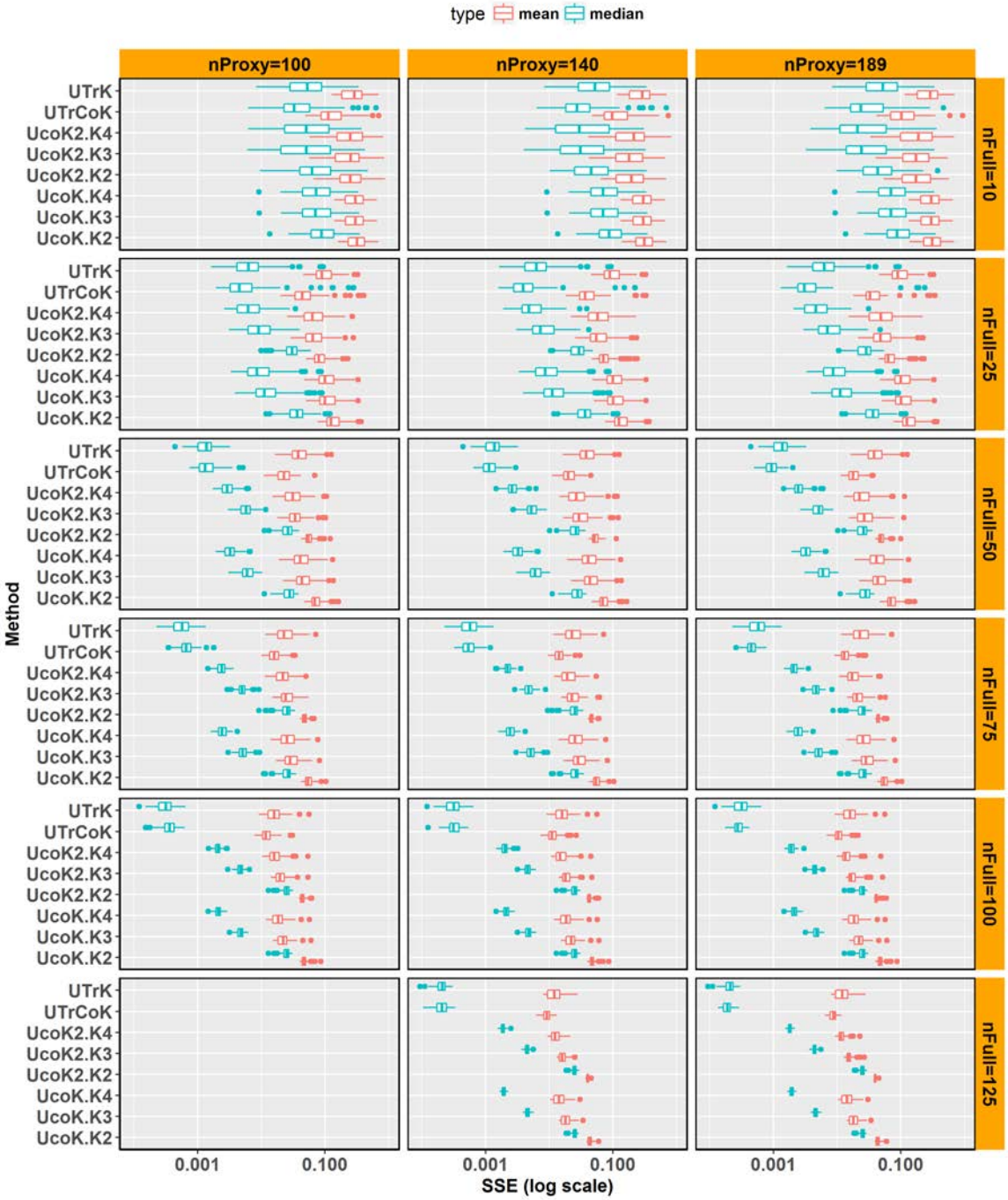


Figure 6.9: *Error analysis of Monte Carlo results on 2 parameter dataset. (Note: mean = mean of means; median = mean of medians accross 100 datasets as varied in MC study)*

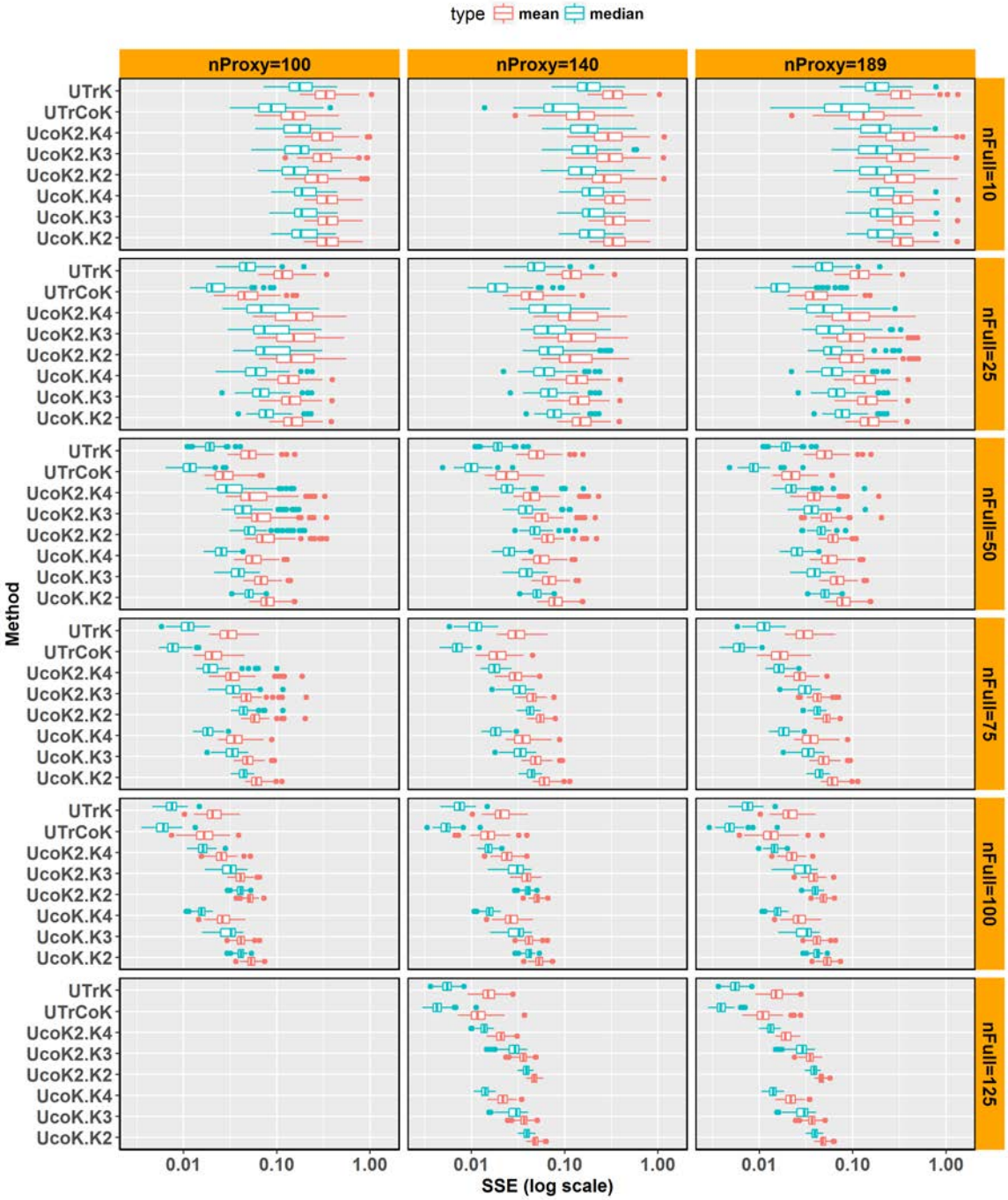


Figure 6.10: *Error analysis of Monte Carlo results on 3 parameter dataset. (Note: mean = mean of means; median = mean of medians accross 100 datasets as varied in MC study)*

fitting procedure for parameter inference, and in the case of projection-based methods we considered five and six principal components since they captured the most of the variance in this data (98%). A few forecasts produced with the trace-based methods on this dataset are given in Figure 6.13, while the results of the Monte Carlo study are given in Figure 6.14.

We observe that the results of the uranium case study are very similar to the results we obtained on synthetic datasets. Trace-based approach slightly outperformed the projection-based approaches, and in this case there was not much difference between single variate projection based approach (UCoK) and multivariate projection based approach (UCoK2).

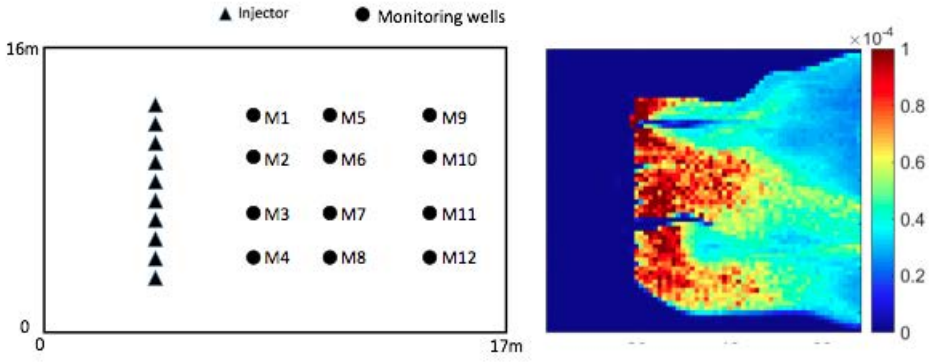


Figure 6.11: *Uranium contamination model. Left - spatial setup (modified from Kowalsky et al. [2012]). Right - A map of immobilized uranium at the end of simulation time*

6.5 Chapter Conclusion

In this chapter we introduced and analyzed trace co-kriging (UTrCoK), a novel and original method for interpolation of multivariate functional data. The method is useful for emulation of functional variables produced by computer codes of variable degrees of fidelity and numerical speed. The proposed method is applicable to situations where all computer codes produce the same type of functional outputs (i.e. rate vs. time), and where discrepancies between the functions are in amplitude rather

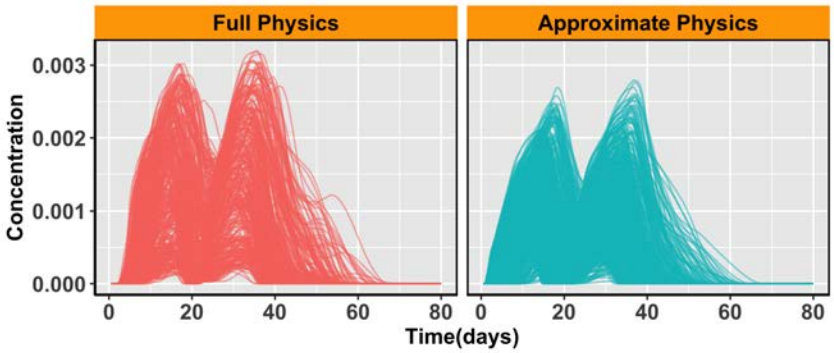


Figure 6.12: *Uranium Dataset: Full physics and approximate physics datasets*

than in phase. In addition, we also presented a projection-based multifidelity computer code emulation technique. The two methods were applied to real and synthetic subsurface flow modeling case studies and their solutions were then compared to the solutions of another two single-variate functional interpolation methods: universal trace kriging (UTrK: Menafoglio et al. [2013], chapter 3) and universal co-kriging for functional data (UCoK: Menafoglio et al. [2016b], chapter 3). To gain deeper understanding about the ranges of applicability of each method, we set up three Monte Carlo studies in which we varied the size of the training sets and the ratios between proxy and full physics simulation runs. Based on the results of our analyses we draw the following conclusions:

- In general UTrCoK performed best out of all considered methods, and particularly better in cases when the number of high fidelity flow simulations was low. This is due to the fact that proxy flow simulations in combination with the linear model of coregionalization (LMC) helped produce better variogram fits.

- UTrCoK requires a much lower modeling effort. Trace variography required LMC fitting over three empirical variograms for two levels of computer code, while the projection-based method on the same data and with only two principal components on each level of computer code required computing and fitting ten variograms. Automated parameter inference procedures in the context of UCoK2 were not attempted in this work.

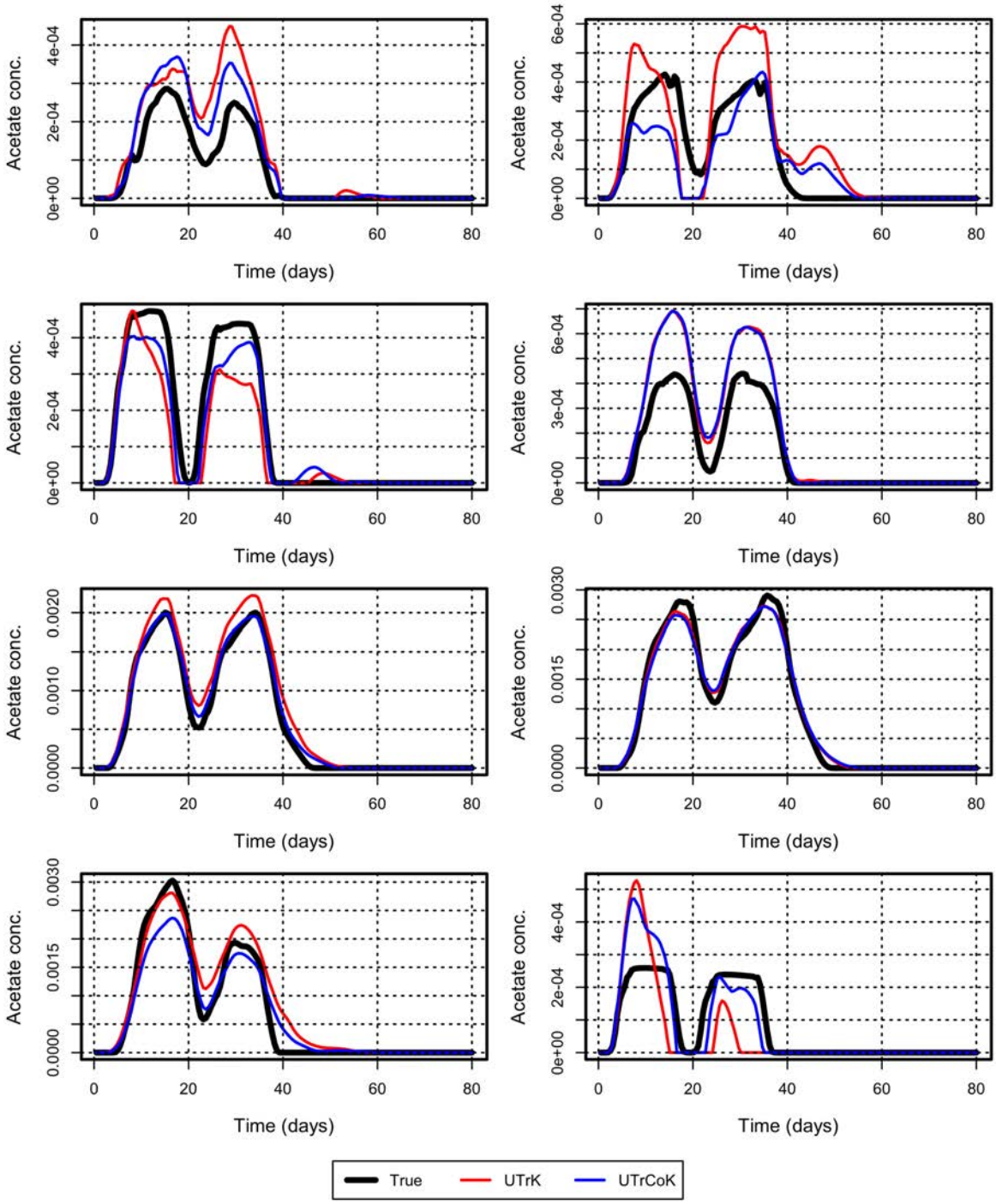


Figure 6.13: *Uranium Dataset: A few forecasts*

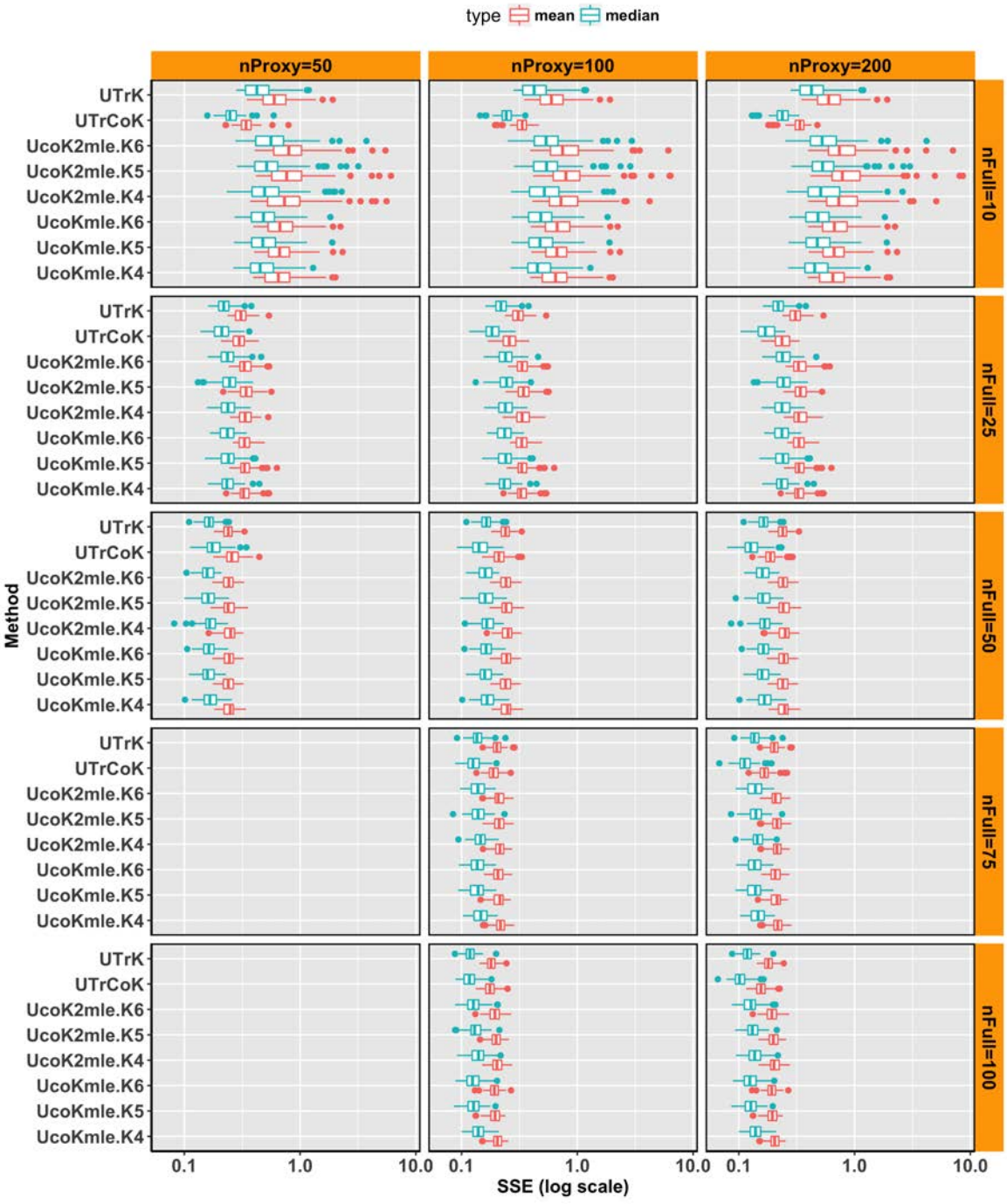


Figure 6.14: *Error analysis of Monte Carlo results on Uranium contamination dataset.*
(Note: mean = mean of means; median = mean of medians accross 100 datasets as varied in MC study)

- All methods, single and multivariate, converged to the same solution for larger numbers of high fidelity flow simulations. This result suggests that proxy flow simulations become useless after a certain number of full physics simulations and that one can approximate the true solution by means of single variate functional interpolation. This result raises an important practical question of how to estimate this critical number of full physics simulations, or when to stop sampling with the proxy?

- Projection-based methods performed worse than trace based methods for low numbers of high fidelity flow simulations. This poor performance is due to difficulties with the estimation of the functional principal components with very low number of training functions. This in combination with the previous item suggests that there is a specific (most likely narrow) range of high fidelity runs for which it is beneficial to use proxy modeling with projection based approach. We hypothesize that this range depends on the complexity of functional data and the dimensionality of the input space. This topic remains to be investigated in future research.

In our analyses we relied on variogram fitting for parameter inference which was our only option in the case of trace based methods. However, we do recognize the need for the development of an automated procedure for parameter inference of trace based methods. This subject will also be in the focus of our future work.

6.5.1 Application to Shale Reservoir Modeling

In the US there is a plethora of information from thousands of vertical wells that were drilled through nowadays developed shale plays. Such information mostly includes well logs and in specific cases production data. For example, Anadarko dataset considered in chapters 4 and 5 also included 2500 well logs from vertical wells drilled in the study area. The challenge is how to incorporate such massive amount of data into data driven shale forecasting framework. We hypothesize that the projection based approach presented in this chapter can be applied with slight modifications (as in chapter 4) to this forecasting problem. This topic is left to be explored in the future research.

Bibliography

Hilbert space. https://en.wikipedia.org/wiki/Hilbert_space. Accessed: 2017-10-12.

Roderick Perez Altamar and Kurt Marfurt. Mineralogy-based brittleness prediction from surface seismic data: Application to the barnett shale. *Interpretation*, 2(2): T255–T271, November 2014. doi: 10.1190/INT-2013-0161.1.

J.J. Arps. Analysis of decline curves. *Transactions of the AIME*, 160(01), December 1945. doi: 10.2118/945228-G. URL https://www.onepetro.org/journal-paper/SPE-945228-G.

Ziv Bar-Joseph, David K. Gifford, and Tommi S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 1(1):1 – 9, 2001. URL https://pdfs.semanticscholar.org/1f9e/1dd39e8d76f6f3169a82211aa83a3a87d1cf.pdf.

M. Bohorquez, R. Giraldo, and J. Mateu. Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, 31(1):53–70, June 2016. ISSN 1436-3259. doi: 10.1007/s00477-016-1266-y. URL http://dx.doi.org/10.1007/s00477-016-1266-y.

Carl de Boor. *A Practical Guide to Splines*. Springer, 1978. ISBN 978-0-387-95366-3. URL http://www.springer.com/us/book/9780387953663.

I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling - Theory and Applications*. Springer, 2005. ISBN 978-0-387-28981-6. URL http://www.springer.com/us/book/9780387251509.

Francesca Bottazzi and Ernesto Della Rossa. A functional data analysis approach to surrogate modeling in reservoir and geomechanics uncertainty quantification.

Mathematical Geosciences, 49(4):517–540, 2017. ISSN 1874-8953. doi: 10.1007/s11004-017-9685-y. URL http://dx.doi.org/10.1007/s11004-017-9685-y.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984.

Leo Breiman. Bagging predictors. Technical report, Department of Statistics, University of California Berkeley, 1994. URL https://www.stat.berkeley.edu/~breiman/bagging.pdf.

Leo Breiman. Random forests - random features. Technical report, Department of Statistics, University of California Berkeley, 1999. URL https://www.stat.berkeley.edu/~breiman/random-forests.pdf.

Leo Breiman and Jerome Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(1):3–54, 1997. doi: 10.1111/1467-9868.00054. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00054.

William Caballero, Ramón Giraldo, and Jorge Mateu. A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment*, 27(7):1553–1563, Feb 2013. ISSN 1436-3259. doi: 10.1007/s00477-013-0691-4. URL http://dx.doi.org/10.1007/s00477-013-0691-4.

Jef Caers. *Modeling Uncertainty in the Earth Sciences*. Wiley, 2011. ISBN 9781119998716. URL https://books.google.com/books?id=ZtGsi4_Tp1UC.

Q. Cao, R. Banerjee, S. Gupta, J. Li, W. Zhou, and B. Jeyachandra. Data driven production forecasting using machine learning. In *SPE Argentina Exploration and Production of Unconventional Resources Symposium*, Buenos Aires, Argentina, 1-3 June 2016. Society of Petroleum Engineers. doi: 10.2118/180984-MS. URL https://www.onepetro.org/conference-paper/SPE-180984-MS.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics - Modeling Spatial Uncertainty*. Wiley, 1999.

Isobel Clark, Karen Basinger, and William Harper. Muck, a novel approach to cokriging. In Bruce Buxton, editor, *87 Conference on Geostatistical, Sensitivity and*

Uncertainty Methods for Ground-water Flow and Radionuclide Transport Modeling, 1987. URL http://drisobelclark.kriging.com/publications/Battelle1987.pdf.

L. De Cesare, D. E. Myers, and D. Posa. Estimating and modeling space-time correlation structures. *Statistics and Probability Letters*, 51:9–14, January 2001. URL http://www.u.arizona.edu/~donaldm/homepage/my_papers/StProbltrs-1.pdf.

Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, April 2010. URL https://projecteuclid.org/euclid.aos/1266586626.

Anh N. Duong. An unconventional rate decline analysis approach for tight and fracture-dominated gas wells. In *Canadian Unconventional Resources and International Petroleum Conference*, Calgary, Alberta, Canada, 19-21 October 2010. Society of Petroleum Engineers. doi: 10.2118/137748-MS. URL https://www.onepetro.org/conference-paper/SPE-137748-MS.

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25): 14863 – 14868, 1998. URL http://www.pnas.org/content/95/25/14863.short?ssource=mfc&rss=1.

Soodabeh Esmaili and Shahab Mohaghegh. Using data driven analytics to assess the impact of design parameters on production from shale. In *SPE Annual Technical Conference and Exhibition*, New Orleans, LA USA, 30 September - 2 October 2013. Society of Petroleum Engineers. doi: 10.2118/166240-MS. URL https://www.onepetro.org/conference-paper/SPE-166240-MS.

Darryl Fenwick, Céline Scheidt, and Jef Caers. Quantifying asymmetric parameter interactions in sensitivity analysis: Application to reservoir modeling. *Mathematical Geosciences*, 46(4):493–511, May 2014. ISSN 1874-8953. doi: 10.1007/s11004-014-9530-5. URL https://doi.org/10.1007/s11004-014-9530-5.

Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis.*

Springer Series in Statistics. Springer-Verlag New York, 2006. ISBN 978-0-387-36620-3. URL http://www.springer.com/us/book/9780387303697.

Thomas Fricker, Jeremy Oakley, and Nathan Urban. Multivariate gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56, 2013. doi: 10.1080/00401706.2012.715835. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.2012.715835.

Alan Gelfand, Alexandra Schmidt, Sudipto Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Sociedad de Estadistica e Investigacion Operativa Test*, 13(2):263–312, 2004. URL https://link.springer.com/article/10.1007/BF02595775.

Ramon Giraldo. *Geostatistical Analysis of Functional Data*. PhD thesis, University of Barcelona, 2009.

Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

Ognjen Grujic and Shahab Mohaghegh. Fast track reservoir modeling of shale formations in the appalachian basin. application to lower huron shale in eastern kentucky. In *SPE Eastern Regional Meeting*, Morgantown, WV, USA, 13-15 October 2010. Society of Petroleum Engineers. doi: 10.2118/139101-MS. URL https://www.onepetro.org/conference-paper/SPE-139101-MS.

Ognjen Grujic, Carla Da Silva, and Jef Caers. Functional approach to data mining, forecasting and uncertainty quantification in unconventional reservoirs. In *SPE Annual Technical Conference and Exhibition*, Houston, TX, USA, October 2015. Society of Petroleum Engineers. doi: 10.2118/174849-MS. URL https://www.onepetro.org/conference-paper/SPE-174849-MS.

Ognjen Grujic, Alessandra Menafoglio, Guang Yang, and Jef Caers. Cokriging for multivariate hilbert space valued random fields: application to multifidelity computer code emulation. *to appear in: Stochastic Environmental Research and Risk Assessment*, 2017. URL http://doi.org/10.1007/s00477-017-1486-9.

Chaohua Guo, Mingzhen Wei, and Hong Liu. Modeling of gas production from shale reservoirs considering multiple transport mechanisms. *PLOS ONE*, 10(12):1–24,

12 2015. doi: 10.1371/journal.pone.0143649. URL https://doi.org/10.1371/journal.pone.0143649.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York, 2012. ISBN 9781461436553. URL https://books.google.com/books?id=OVezLB__ZpYC.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN 9780470016916. URL https://books.google.com/books?id=6Fu3BgAAQBAJ.

L. Josset, D. Ginsbourger, and I. Lunati. Functional error modeling for uncertainty quantification in hydrogeology. *Water Resources Research*, 51(2):1050–1068, February 2015. URL http://onlinelibrary.wiley.com/doi/10.1002/2014WR016028/abstract.

Henry F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, Sep 1958. ISSN 1860-0980. doi: 10.1007/BF02289233. URL https://doi.org/10.1007/BF02289233.

B Karpiński and M Szkodo. Clay minerals–mineralogy and phenomenon of clay swelling in oil & gas industry. *Advances in Materials Science*, 15(1):37–55, 2015. URL https://www.degruyter.com/downloadpdf/j/adms.2015.15.issue-1/adms-2015-0006/adms-2015-0006.pdf.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. URL https://www.jstor.org/stable/2673557.

M. B. Kowalsky, S. Finsterle, K. H. Williams, C. Murray, D. Commer, M. Newcomer, A. Englert, C. I. Steefel, and S. S. Hubbard. On parameterization of the inverse problem for estimating aquifer properties using tracer data. *Water Resources Research*, 48(6):1–25, June 2012. doi: 10.1029/2011WR011203. URL http://onlinelibrary.wiley.com/doi/10.1029/2011WR011203/full.

Randy F. LaFollette and William David Holcomb. Practical data mining: Lessons-learned from the barnett shale of north texas. In *SPE Hydraulic Fracturing Technology Conference*, The Woodlands, TX USA, January 2011. Society of Petroleum Engineers. doi: 10.2118/140524-MS. URL https://www.onepetro.org/conference-paper/SPE-140524-MS.

Li Li, Nitin Gawande, Michael B. Kowalsky, Carl I. Steefel, and Susan S. Hubbard. Physicochemical heterogeneity controls on uranium bioreduction rates at the field scale. *Environmental Science and Technology*, 45(23):9959–9966, October 2011. doi: 10.1021/es201111y. URL https://www.ncbi.nlm.nih.gov/pubmed/21988116.

Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests and randomized trees. Technical report, Department of EE & CS, University of Liege, Belgium, 2014. URL https://orbi.ulg.ac.be/bitstream/2268/155642/1/louppe13.pdf.

A. Menafoglio, A. Guadagnini, and P. Secchi. Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. *Water Resources Research*, 52(8):5708–5726, 2016a. ISSN 1944-7973. doi: 10.1002/2015WR018369. URL http://dx.doi.org/10.1002/2015WR018369.

Alessandra Menafoglio, Piercesare Secchi, and Matilde Dalla Rosa. A universal kriging predictor for spatially dependent functional data of a hilbert space. *Electronic Journal of Statistics*, 7(0):2209–2240, 2013. ISSN 1935-7524. doi: 10.1214/13-ejs843. URL http://dx.doi.org/10.1214/13-EJS843.

Alessandra Menafoglio, Alberto Guadagnini, and Piercesare Secchi. 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, Feb 2014. ISSN 1436-3259. doi: 10.1007/s00477-014-0849-8. URL http://dx.doi.org/10.1007/s00477-014-0849-8.

Alessandra Menafoglio, Ognjen Grujic, and Jef Caers. Universal kriging of functional data: Trace-variography vs cross-variography? application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39 – 55, 2016b. ISSN 2211-6753. doi: http://dx.doi.org/10.1016/j.spasta.2015.12.003. URL http://www.sciencedirect.com/science/article/pii/S2211675315001141.

Shahab Mohaghegh, Ognjen Grujic, Saeed Zargari, and Amirmasoud Kalantari. Modeling, history matching, forecasting and analysis of shale reservoirs performance using artificial intelligence. In *SPE Digital Energy Conference and Exhibition*, The Woodlands, TX, USA, 19-21 April 2011. Society of Petroleum Engineers. doi: 10.2118/143875-MS. URL https://www.onepetro.org/conference-paper/SPE-143875-MS.

Amir Mohammad Nejad, Stanislav Sheludko, Robert Frank Shelley, Hodgson Trey, and Patrick Riley Mcfall. A case history: Evaluating well completions in eagle ford shale using a data-driven approach. In *SPE Hydraulic Fracturing Technology Conference*, The Woodlands, TX, USA, 3-5 February 2015. Society of Petroleum Engineers. doi: 10.2118/173336-MS. URL https://www.onepetro.org/conference-paper/SPE-173336-MS.

David Nerini, Pascal Monestiez, and Claude Manté. Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101(2):409–418, Feb 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2009.03.005. URL http://dx.doi.org/10.1016/j.jmva.2009.03.005.

S.A. Proskin, J.D. Scott, and H.S. Chhina. Current practice in the interpretation of microfrac tests in oil sands. In *SPE California Regional Meeting*, Ventura, California, 4-6 April 1990. Society of Petroleum Engineers. doi: 10.2118/20040-MS. URL https://www.onepetro.org/conference-paper/SPE-20040-MS.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 2005.

James Ramsay, Giles Hooker, and Spencer Graves. *Functional Data Analysis with R and MATLAB*. Springer, 2009.

J.O. Ramsay and Xiaochun Li. Curve registration. *Journal of Statistical Society*, 60:351–363, April 1998. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00129/abstract.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006. URL http://www.gaussianprocess.org.

Lan Ren, Jinzhou Zhao, and Yongquan Hu. Hydraulic fracture extending into network in shale: Reviewing influence factors and their mechanism. *The Scientific World Journal*, 2014(847107):1–9, September-October 2014. doi: 10.1155/2014/847107.

J.B. Roen, B.J. Walker, Appalachian Oil, Natural Gas Research Consortium, West Virginia Geological, and Economic Survey. *The Atlas of Major Appalachian Gas Plays*. Number v. 1 in Publication (West Virginia Geological and Economic Survey). West Virginia Geological and Economic Survey, 1996. URL https://books.google.com/books?id=AtWdcAAACAAJ.

Peter J. Rousseeuw, Ida Ruts, and John W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999. doi: 10.1080/00031305.1999.10474494. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474494.

Jerome Sacks, William Welch, Toby Mitchell, and Henry Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. URL https://projecteuclid.org/euclid.ss/1177012413.

Mark Robert Segal. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418):407–418, 1992. doi: 10.1080/01621459.1992.10475220. URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1992.10475220.

Mohammad Shahvali, Bradley Mallison, Kaihong Wei, and Herve Gross. An alternative to streamlines for flow diagnostics on structured and unstructured grids. *SPE journal*, 17, September 2012. doi: 10.2118/146446-PA. URL https://www.onepetro.org/journal-paper/SPE-146446-PA.

Robert Shelley, Luke Saugier, Wadhah Al-Tailji, Nijat Guliyev, and Koras Shah. Understanding hydraulic fracture stimulated horizontal eagle ford completions. In *SPE/EAGE European Unconventional Resources Conference and Exhibition*, Vienna, Austria, 20-22 March 2012. Society of Petroleum Engineers. doi: 10.2118/152533-MS. URL https://www.onepetro.org/conference-paper/SPE-152533-MS.

Hiroki Sone and Mark D. Zoback. Mechanical properties of shale-gas reservoir rocks - part 1: Static and dynamic elastic propertis and anisotropy. *Geophysics*, 78 (5):D381–D392, September-October 2013. doi: 10.1190/GEO2013-0050.1. URL http://doi.org/10.1190/GEO2013-0050.1.

C.I. Steefel, C.A.J. Appelo, B. Arora, D. Jacques, T. Kalbacher, O. Kolditz, V. Lagneau, P.C. Lichtner, K.U. Mayer, J.C.L. Meeussen, S. Molins, D. Moulton, H. Shao, J. imnek, N. Spycher, S.B. Yabusaki, and G.T. Yeh. Reactive transport codes for subsurface environmental simulation. *Computational Geosciences*, 19(3):445–478, Jun 2015. doi: 10.1007/s10596-014-9443-x. URL https://link.springer.com/article/10.1007/s10596-014-9443-x.

Arthur Thenon, Véronique Gervais, and Mickaële Le Ravalec. Multi-fidelity metamodeling for reservoir engineering - application to history matching. *Computational Geosciences*, 20(6):1231–1250, 2016. ISSN 1573-1499. doi: 10.1007/s10596-016-9587-y. URL http://dx.doi.org/10.1007/s10596-016-9587-y.

Sumeet Trehan, Kevin Carlberg, and Louis J. Durlofsky. Error modeling for surrogates of dynamical systems using machine learning. *International Journal for Numerical Methods in Engineering*, July 2017. ISSN 1097-0207. doi: 10.1002/nme.5583. URL http://dx.doi.org/10.1002/nme.5583.

G. W. Verly. Sequential gaussian cosimulation: A simulation method integrating several types of information. In Amilcar Soares, editor, *Geostatistics Troia 92*, 1992. URL https://link.springer.com/chapter/10.1007/978-94-011-1739-5_42.

Hans Wackernagel. *Multivariate Geostatistics*. Wiley, 2010.

Fred P. Wang and Julia F.W. Gale. Screening criteria for shale-gas systems. In *GCAGS 59th Annual Meeting, Shreveport, Louisiana*, volume 59, pages 779–793. AAPG, November 2009.

Kenneth Williams, Philip E. Long, James Davis, Michael Wilkins, A. Lucie N'Guessan, Carl Steefel, Li Yang, Darrell Newcomer, Frank Spane, Lee Kerkhof, Lora McGuinness, Richard D. Dayvault, and Derek Lovley. Acetate availability and its influence on sustainable bioremediation of uranium contaminated groundwater.

Geomicrobiology Journal, 28(5-6):519–539, July 2011. doi: 10.1080/01490451.2010.520074. URL http://dx.doi.org/10.1080/01490451.2010.520074.

World Energy Council. World energy resources, 2016. URL https://www.worldenergy.org/wp-content/uploads/2016/10/World-Energy-Resources-Full-report-2016.10.03.pdf.

Steven B. Yabusaki, Yilin Fang, Philip E. Long, Charles T. Resch, Aaron D. Peacock, John Komlos, Peter R. Jaffe, Stan J. Morrison, Richard D. Dayvault, David C. White, and et al. Uranium removal from groundwater via in situ biostimulation: Field-scale modeling of transport and biological processes. *Journal of Contaminant Hydrology*, 93(1-4):216–235, Aug 2007. ISSN 0169-7722. doi: 10.1016/j.jconhyd.2007.02.005. URL http://dx.doi.org/10.1016/j.jconhyd.2007.02.005.

Bicheng Yan, Yuhe Wang, and John E. Killough. Beyond dual-porosity modeling for the simulation of complex flow mechanisms in shale reservoirs. *Computational Geosciences*, 20(1):69–91, Feb 2016. doi: 10.1007/s10596-015-9548-x. URL https://doi.org/10.1007/s10596-015-9548-x.

Yan Yu and Diane Lambert. Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8(4):749–762, 1999. doi: 10.1080/10618600.1999.10474847. URL http://amstat.tandfonline.com/doi/abs/10.1080/10618600.1999.10474847.

Hao Zhang. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18(2):125–139, 2007. ISSN 1099-095X. doi: 10.1002/env.807. URL http://dx.doi.org/10.1002/env.807.

M.D. Zoback. *Reservoir Geomechanics*. Cambridge University Press, 2007. ISBN 9780521770699. URL https://books.google.com/books?id=7Ih4MgEACAAJ.