Universitat de Girona

GEOSTATISTICS FOR CONSTRAINED VARIABLES: POSITIVE DATA, COMPOSITIONS AND PROBABILITIES. APPLICATION TO ENVIRONMENTAL HAZARD MONITORING

Raimon TOLOSANA-DELGADO

ISBN: 84-689-6660-6
Dipòsit legal: GI-125-2006

Geostatistics for constrained variables:

positive data, compositions and probabilities.

Application to environmental hazard monitoring

PhD THESIS

director: Dra. Vera Pawlowsky-Glahn

Doctorate Programm: Medi Ambient — Environmental Sciences

Itinerary: Tecnologia i Física Ambientals — Environmental Physics and Technology

Universitat de Girona

Raimon Tolosana-Delgado

Girona, September 2005

Universitat de Girona
Institut de Medi Ambient
Programa de Doctorat: Medi Ambient
Itinerari: Tecnologia i Física Ambientals

La doctora **Vera Pawlowsky-Glahn**, professora catedràtica del departament d'Informàtica i Matemàtica Aplicada de la Universitat de Girona, en qualitat de directora, i el doctor **Josep Daunis i Estadella**, professor titular del departament d'Informàtica i Matemàtica Aplicada i professor del Programa de Doctorat en Medi Ambient de la Universitat de Girona, en qualitat de tutor,

CERTIFIQUEN:

que l'enginyer **Raimon Tolosana Delgado** ha realitzat la tesi titulada "*Geostatistics for constrained variables: positive data, compositions and probabilities*", que es presenta en aquesta memòria per optar al grau de Doctor en Medi Ambient (Itinerari de Tecnologia i Física Ambientals) per la Universitat de Girona.

I perquè així consti als efectes oportuns, signen el present certificat a Girona, el 26 de setembre del 2005

Vera Pawlowsky-Glahn
directora

Josep Daunis i Estadella
tutor

Raimon Tolosana-Delgado
doctorand

Acknowledgements

Some acknowledgments are always due in a PhD Thesis. First, I want to thank my director, Dr. Vera Pawlowsky-Glahn, who has guided me since 1996, from the most basic under-graduate statistical course to the consecution of this Thesis. She is, as Germans say, my *doctormother*. Close to her, I would also like to mention the same long-term guidance of Dr. Juan-José Egozcue. And, as well, the rest of my *doctorfamily* deserves a sincere thank: Carles Barceló-Vidal, Josep Daunis-i-Estadella, Josep-Anton Martin-Fernàndez, Glòria Mateu-Figueras, and Santiago Thió-Henestrosa—the members of the group—and John Aitchison, Heinz Burger, Gerald van den Boogaart, Antonella Buccianti and Hilmar von Eynatten—who have intensely taken part in the discussion network where this Thesis was born—. This professional thanks giving should not finish without mentioning Àngels Canals, Mireia Garreta, Eugènia Martí-Roca, Neus Otero and Albert Soler, who have always driven me to the world of real problems, not letting me go too far in the heaven of impractical methods. Finally, I would like to thank Dr. van den Boogaart and Dr. Wackernagel for their deep and useful reviews of this work.

But a Thesis is not only an intellectual work, but also an intense emotional experience. In this sense, I am specially grateful to Joan-Jordi (for his almost-infinite patience with my humor up-and-downs) and to my elder. No one else knows as they how important is for me such a step like applying to the PhD degree. The rest of the *real family*, specially my brother Arnau and my cousin Laura, come afterwards, jointly with my closest friends (Sílvia, Patrick, Maite, Mireia G., Mariona, Marc, Patrícia, Laura, René, Xavi M., Arnau, Roger, Rosalía, Mireia N., Christian, Maribel, Neus, Meri, Xavi P., Anabel, Andrés, Vicky and Julieta). It is a good luck to have such a long list of strictly-selected beloved people.

Agraïments

Una Tesi no és només el resultat de l'esforç del seu autor, tot i que l'autor sigui l'únic responsable del seu contingut, per a bé i per a mal. Per això calen alguns agraïments. Per començar, voldria agrair la dra. Vera Pawlowsky-Glahn aquests 10 anys de guia, des de l'assignatura bàsica d'estadística a optar a ser doctor. Certament, ella es la meva *doctor-mare*, com diuen els alemanys. De la mateixa manera, el dr. Juan-José Egozcue també ha dirigit la meva formació, a voltes en co-direcció, a voltes a l'ombra. I també voldria agrair la feina de tots els científics que m'han ajudat mentre elaborava aquesta tesi, amb les seves idees i la discussió constant: Carles Barceló-Vidal, Josep Daunis-i-Estadella, Josep-Anton Martin-Fernàndez, Glòria Mateu-Figueras, i Santiago Thió-Henestrosa —els membres del grup de recerca— així com John Aitchison, Heinz Burger, Gerald van den Boogaart, Antonella Buccianti i Hilmar von Eynatten. També vull esmentar l'ajuda d'Àngels Canals, Mireia Garreta, Eugènia Martí-Roca, Neus Otero i Albert Soler, que han estat el meu vincle amb els problemes reals. Finalment, estic profundament agraït al dr. van den Boogaart i el dr. Wackernagel per les seves profundes i útils revisions d'aquest text. Sense les contribucions de tots i cadascun d'aquests homes i dones, aquesta tesi no existiria.

Però una tesi tampoc és només el resultat d'un esforç intel·lectual. És una experiència emocional de primera categoria. En aquest sentit, estic particularment agraït al Joan-Jordi (per la seva quasi infinita paciència amb els meus daltabaixos emocionals dels darrers temps) i als meus pares. Ningú millor que ells sap com és d'important per a mi arribar a optar al títol de doctor. Vull acabar reconeixent la resta de la meva familia, amb mon meu germà Arnau i ma cosina Laura al capdavant, junt amb els meus amics íntims (Sílvia, Patrick, Maite, Mireia G., Mariona, Marc, Patrícia, Laura, René, Xavi M., Arnau, Roger, Rosalía, Mireia N., Christian, Maribel, Neus, Meri, Xavi P., Anabel, Andrés, Vicky i Julieta). És una gran sort comptar amb una llista tan llarga de persones estimades.

Contents

1 Introduction **1**
- 1.1 An environmental motivation . . . 1
- 1.2 Statement of the problem . . . 2
  - 1.2.1 Hazard . . . 3
  - 1.2.2 Regionalized variable . . . 4
  - 1.2.3 Sample space . . . 5
- 1.3 Structure of the Thesis . . . 6
- 1.4 Case studies . . . 7
  - 1.4.1 The Tordera river at Gualba (Spain) . . . 7
  - 1.4.2 Air pollution in the Carpathian Range (Ukraine) . . . 8

2 Preliminary concepts **13**
- 2.1 General basic notation . . . 13
- 2.2 Geometry of the sample space . . . 14
  - 2.2.1 Vector space . . . 14
  - 2.2.2 Linear applications . . . 17
- 2.3 Probability laws on coordinates . . . 19
  - 2.3.1 Measure considerations . . . 19
  - 2.3.2 First and second-order moments . . . 20
  - 2.3.3 Normal probability distributions . . . 22
  - 2.3.4 Regression on $\mathbb{E}$ . . . 23
- 2.4 Inference on coordinates . . . 24
  - 2.4.1 Frequentist estimation . . . 24
  - 2.4.2 Bayesian estimation . . . 25
- 2.5 Case studies . . . 26
  - 2.5.1 Water conductivity . . . 26
  - 2.5.2 Ammonia system . . . 29
  - 2.5.3 Moss pollution . . . 33
- 2.6 Distributions on the Simplex . . . 36
  - 2.6.1 The Dirichlet distribution . . . 37
  - 2.6.2 The Normal distribution on the Simplex . . . 39
  - 2.6.3 The $\mathcal{A}$ distribution . . . 40
- 2.7 Remarks . . . 43

2.8 Addendum: invariance of coordinate mean and variance 45

3 Geostatistics in the real space **47**
3.1 Random function 47
3.2 Structural analysis 50
3.2.1 General aspects 50
3.2.2 Auto-covariance functions 50
3.2.3 Cross-covariance functions 56
3.3 Linear prediction 58
3.3.1 General universal kriging 58
3.3.2 Kriging of the drift 60
3.3.3 Simple kriging 60
3.3.4 Properties of kriging estimators 61
3.4 Bayesian Methods 62
3.4.1 Bayesian kriging 62
3.4.2 Model-based geostatistics 63
3.4.3 Bayesian/maximum entropy geostatistics 64
3.5 Change-of-support problems 65
3.5.1 Relationship between point and block-supports 66
3.5.2 Universal block kriging 68
3.5.3 Global change-of-support 69
3.6 Case study: conductivity 71
3.7 Remarks 78
3.8 Addendum: validity of change-of-support models 80

4 Geostatistics in an arbitrary Euclidean space **81**
4.1 Notation 81
4.2 Random function 82
4.3 Structural analysis 83
4.4 Linear prediction 84
4.4.1 The general case of kriging 85
4.4.2 Simple kriging 87
4.4.3 Universal kriging 90
4.5 Remarks 94

5 Geostatistics in the positive real space **97**
5.1 Lognormal kriging 97
5.2 Positive real line space structure 99
5.3 Kriging in the positive real space 100
5.4 Change-of-support problems 101
5.4.1 Lognormal change-of-support model 101
5.4.2 Normal on $\mathbb{R}_+$ change-of-support model 103
5.5 Case study: ammonia pollution risk 107

5.6 Remarks . . . 117

6 Geostatistics in the Simplex **119**
6.1 Kriging of compositions . . . 119
6.2 Simplex space structure . . . 122
6.3 Structural analysis . . . 123
6.4 Kriging in the Simplex . . . 125
6.5 Case study: air quality index . . . 126
6.6 Remarks . . . 130

7 Geostatistics for probability functions **133**
7.1 Indicator kriging . . . 134
7.1.1 Random function . . . 134
7.1.2 Structural analysis . . . 135
7.1.3 Linear prediction . . . 136
7.1.4 Indicator kriging family techniques . . . 137
7.1.5 Disjunctive kriging . . . 138
7.1.6 Bayesian/maximum entropy method . . . 139
7.2 Kriging in $\mathbb{S}^D$ for probabilities . . . 140
7.2.1 Random function . . . 140
7.2.2 The D-part Simplex, a space for discrete probabilities . . . 140
7.2.3 Linear prediction . . . 141
7.3 Kriging in $\mathbb{S}^D$ for generalized indicators . . . 142
7.3.1 The generalized indicator function . . . 142
7.3.2 Structural analysis . . . 145
7.3.3 Linear prediction . . . 148
7.4 A Bayesian method . . . 156
7.5 Case study: conductivity hazard . . . 159
7.5.1 Kriging in the Simplex for generalized indicators . . . 159
7.5.2 Kriging in the Simplex for a single indicator . . . 169
7.5.3 Bayesian estimation in the Simplex for a single indicator . . . 173
7.6 Remarks . . . 174
7.7 Addendum: Bayesian estimation of probability vectors . . . 175

8 Conclusions **179**
8.1 Discussion of case studies . . . 179
8.1.1 Water pollution . . . 179
8.1.2 Air pollution . . . 181
8.2 Discussion of methods . . . 182
8.3 Future work . . . 184

9 Notation summary **187**

Bibliography **190**

List of Tables

1.1 Water quality categories. 7

2.1 Point and interval estimates of the mean of ammonia coordinate. 31

3.1 Periods of the trigonometric drift functions. 72
3.2 Fitted drift coefficients for classical regression and kriging of the drift. . 77

5.1 Fitted drift coefficients using classical regression for $p\mathrm{NH}_4$, $p\mathrm{H}$, and $p\mathrm{K}_a$ series. 109

6.1 Parameters of variogram models in the moss pollution system. 128

7.1 Estimated deciles for the residual conductivity of July 2003. 160
7.2 Parameters of the auto-covariance functions of the generalized indicator functions. 163
7.3 Estimators of a probability for several loss criteria and prior distributions. 177

8.1 Geometries for each parameter in the Gualba station. 179

List of Figures

1.1 Lithologic map of the Tordera basin. 9
1.2 Map of land uses of the Tordera basin. 9
1.3 Location of the Carpathian range. 10
1.4 Features of the studied region of the Ukrainian Carpathian range. . . . 10

2.1 Histogram of conductivity against the normal distribution. 27
2.2 Joint posterior distribution of μ and σ parameters for conductivity. . . 28
2.3 Predictive distribution of conductivity. 28
2.4 Histogram of ammonia compared with a normal on $\mathbb{R}_+$ distribution. . . 31
2.5 Joint posterior distribution of μ and σ parameters for ammonia. . . . 32
2.6 Predictive distribution of ammonia. 32
2.7 Ternary diagram of $(\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg})$, and the normal distribution on $\mathbb{S}^D$. . . 34
2.8 Diagrams of $(\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg})$, with confidence regions. 34
2.9 Time evolution over July 2002 of all the measured variables. 44

3.1 Elements of variograms and covariance functions. 52
3.2 Some variogram models without sill. 54
3.3 Some variogram models with a sill. 54
3.4 Evolution of conductivity and temperature (years 2002-2003). 73
3.5 Evolution of conductivity and temperature (July 2002/2003). 73
3.6 Frequency spectrum of water temperature. 74
3.7 Scatter plot of observed water temperature against prediction. 74
3.8 Covariance function of conductivity original data set. 75
3.9 Covariance function of conductivity residuals. 75
3.10 Time evolution of residual conductivity. 76
3.11 Time evolution of conductivity. 76
3.12 Time evolution of high conductivity hazard. 78

5.1 Selectivity curves for a standard lognormal distribution. 103
5.2 Comparison of conventional income curves for different supports. 104
5.3 Time series of $p\mathrm{NH}_4$, $p\mathrm{H}$, and $p\mathrm{K}_a$. 107
5.4 Experimental auto- and cross-correlation functions at short range. . . . 108
5.5 Experimental auto- and cross-covariances functions at short range. . . . 111

5.6 Experimental and fitted auto- and cross-covariance functions at long range. 112
5.7 Predicted coordinates of the ammonia system for July 2002 and 2003. . 113
5.8 Hazard of exceeding each threshold defining water quality categories. . 114
5.9 Probability of each water quality category. 115

6.1 Maps of (Hg, Fe, Pb) at the sampled locations. 127
6.2 Omnidirectional variograms and cross-variograms for the moss pollution system. 128
6.3 Maps of predicted values for each coordinate in the moss pollution system. 131
6.4 Predicted values for (Hg, Fe, Pb) in the moss pollution system. 132

7.1 Predicted and true conditional probability as a function of the correlation coefficient. 153
7.2 Predicted and true conditional probability as a function of the parameter α. 154
7.3 Data set and estimated distribution for the first 10 days of July 2003. . 161
7.4 Matrix of generalized indicator auto- and cross-variogram plots. 162
7.5 Generalized indicator variogram plots, with fitted models. 163
7.6 Generalized indicator data and simple kriging prediction. 165
7.7 Final distribution predictions as a function of the parameter α. 166
7.8 Data set and estimated distributions for July 2002 and July 2003. . . . 167
7.9 Estimated distributions of some selected prediction moments. 168
7.10 Estimated hazard of exceeding $1000\mu S/cm$ of conductivity. 168
7.11 Experimental variogram and fitted model of the single cutoff case. . . . 170
7.12 Data set and estimated distributions for July 2002 with a single cutoff. 171
7.13 Estimated distributions for July 2002 with a single cutoff using the global Bayesian method. 172

Chapter 1

Introduction

1.1 An environmental motivation

One of the most important applications of Life and Earth sciences to our everyday life is the assessment of *environment quality*. This concept arose recently, after more than one century of industrial use and abuse of our environment. In a natural environment, pollution of running or still water, air and soil, as well as the damage on living beings, has been the concern of environmentalists since the 70s. In urban areas of developed countries, an interest has also arisen on the quality of the environment, from air pollution to the potability of supplied water. Humans, like all living beings, use the environment as a resource, and demand a minimal quality of it. But at the same time, this very use alters the quality of this resource, most usually lowering it. The resolution of this contradiction, and the quest for a tradeoff between maximal use and minimal quality alteration, is the subject of environmental management policies. A reasonable environmental management policy calls for an assessment of the environmental quality, as well as of the potential uses of an environmental resource.

To provide policy makers with such an assessment, scientists have been developing environmental quality indices for the last 20 years. A quality index is intended as an *objective quantitative* measure of the suitability of a resource for a given use. There are many types of indices, according to which method they use to measure quality: color, chemical elements or microorganism concentration, counts of sensitive macroinvertebrates, etc. to give but a few from the huge variety of physical, chemical, microbial or biological quality indices available. Note the variety of characteristics of the sample space of these variables: some are qualitative (color), some numerical of discrete nature (counts), and other are continuous, either unbounded (temperature) or bounded (concentrations). However, the measures should be simple and give an integrated information on the state of the environment, and this is not common: the most integrated and powerful quality indices are so complex that only an expert can use them as *objective* measures. Statistics have there a tool to offer, which can integrate as many information of any kind as one can model and offers a final quantitative continuous measure of quality: probability.

In natural risk assessment, one uses the concepts of hazard, vulnerability and risk. Hazard of a given event is defined as the probability of occurrence of this event. Vulnerability informs of the loss that will occur if this hazardous event takes place; it is usually measured in terms of money or human lives. Risk is then the expected loss, or the product of hazard by vulnerability. This simple scheme can become so complex as desired, by introducing many simple hazardous events, or events which accept different degrees of danger. By comparing them with environmental quality issues, one might interpret the use of a resource as a vulnerability, while hazard can be seen as an environmental quality measure: the higher the hazard the lower the quality.

Usually, natural hazard is associated to a place and a moment, and so may be also an environmental quality measure. In this way, one may take measures *e.g.*of a given water quality index in a river, both in time at a single place, or along the flow in a given moment. It is then needed to take into account some sort of relation between the measurements, since they are clearly not independent. Summarizing, a probabilistic water quality index calls for statistical techniques of computation of probabilities which take full profit of all the characteristics of the measurements involved, mainly their mutual dependence and their different sample spaces, and with different scales of comparison in them.

1.2 Statement of the problem

The *sample space* of a variable is its set of possible values. Although it is an old statistical concept, its practical importance has been seldom considered in the applied sciences. In particular, it came to the geosciences through the context of Compositional Data Analysis [Aitchison, 1986]. It is nevertheless quite easy to take it into account, when the sample space may be given an Euclidean structure.

Geostatistics [Matheron, 1965] is the name of a series of techniques devised to treat data sets with mutual dependence, something which precludes using on them classical statistical techniques. Geostatistical techniques may be divided in two groups: those predicting the value of a variable, and those estimating its probability distribution. In a conventional framework, both types assume almost always (implicit or explicitly) the variables to have a *real unbounded* sample space. However, many variables do not satisfy this requirement: their sample spaces are either subsets of the real space, or a set of categories. This is not an unimportant issue: by taking into account possible structures for the sample space, new light is cast on old geostatistical problems, *e.g.*those affecting positive variables assumed to be lognormally distributed. In this way, we may estimate better the probability distribution as a tool for hazard/quality assessment.

The goal of this Thesis is to integrate these considerations on the structure of the sample space of the variables into the existing geostatistical techniques, in the case that this sample space can be given a meaningful Euclidean structure. We will show that geostatistical tools and concepts are objects or operations in this space structure,

which means that they have a sense on their own, independently of the way they are represented.

1.2.1 Hazard

Hazard assessment is essentially committed with the estimation of the probability of occurrence of events, which are regarded as dangerous [Hewitt, 1997]. The first step is then the definition of a dangerous event, or a family of ordered events. One might think on the danger of, *e.g.*presence of carbon monoxide in the air (lethal in parts-per-billion), or the concentration of ammonia in a lake. The first case would give us a single dangerous event—being above the toxicity threshold—, while the second one would give rise to a family of events ordered in an increasing degree of damage—being above each of the toxicity thresholds for fishes, for human beings, for agricultural uses, etc.—.

In probabilistic terms, hazard assessment reduces to the estimation of the probability of occurrence of an event. A Bernoulli event is the simplest model to devise: at each *trial*, a probability of *success* is defined as the probability p of occurrence of the desired event, and the probability of *failure* is its complementary $q = 1 - p$. These two probabilities do not change from trial to trial, and the result of each trial is independent from all the others.

There are several philosophical approaches to the concept of probability. Each of them yield different procedures to estimate p or q, being complementary. We will speak about the frequentist and about the Bayesian approach, following the exposition of Leonard and Hsu [1999]. The *frequentist approach* defines the probability of occurrence of an event as the limit of the number of times the event occurred divided by the total number of trials, when this number of trials tends to infinity. The *Bayesian approach* regards the value of p as a subjective reliability of occurrence of the event. In the frequentist case, the estimator $\hat{p}$ of p will be the number of times the event occurred divided by the number of trials, although this number of trials is finite. The Bayesian estimation procedure takes the *prior* reliability assessment of the possible values of p, *updates* it by the information brought by the observed events, and obtains a *posterior* assessment of the reliability of this occurrence. From the differences between them, we highlight: a) the Bayesian approach needs a prior assessment of the possible values of p, and b) it offers a posterior assessment of the possible values of p, while the frequentist approach yields a single estimate which only depends on the data, and not on prior knowledge.

When attending to a family of ordered dangerous events, we face the estimation of a probability distribution function, *i.e.*a function which describes the reliability of each possible outcome of a random variable. This is achieved by assuming a parametric model for that random variable (*e.g.*the normal or Gaussian distribution), and estimating the parameters of the model. The Bernoulli case may be seen as a particular case, where the Bernoulli model is described by the parameter p. Both the frequentist and the Bayesian approaches allow such an estimation, using a *sample*, a

set of independent realizations of the random variable, *i.e.*independent measurements. Once the parameters are estimated, the probability function is completely specified, and the probability of any dangerous event can be computed. In the second example above, once we know the probability function of the ammonia concentration we can compute the probability of being above each of the mentioned thresholds. Obviously, these probabilities are expected to be ordered, as were the thresholds themselves.

1.2.2 Regionalized variable

The term *regionalized variable* was coined by Matheron [1965] to describe those sets of measurements, distributed across time and space, presenting a mutual dependence inherited from the proximity of their sampling locations. Its generalization to regionalized vectors is usually referred to as *coregionalization*. The set of techniques used to analyze regionalized variables and vectors is known as *geostatistics*.

The coregionalized paradigm is used to investigate structural dependencies among spatially-distributed variables, like *e.g.*joint covariation of porosity and log-permeability in an aquifer or of several climatic variables along a mountain range. This spatial structural analysis of covariation is usually followed by an interpolation procedure: the coregionalization assumption allows the estimation of the whole vector, or of some of its components, at any non- or partially-sampled location, jointly with a measure of the incurred error. A further assumption, the joint normality of all variables at all locations, delivers a stronger result: the estimate and its error define the distribution of uncertainty on the true predicted value, conditional on the observations. Using this result, hazard assessments have been conducted for regionalized variables.

When joint multivariate normality is not a valid model assumption, but the probabilities of some hazardous events have to be determined, there are other useful geostatistical techniques. One of them works with indicator functions [Journel, 1983]: these are valued as one at those places where the hazardous event was actually observed, and as zero elsewhere at the sampled locations. Then, these indicator values are considered a coregionalization, and interpolated: the obtained values are finally interpreted as the conditional probability of observing the event. This technique is very frequently used, due to its simplicity and straightforward application, in spite of the rather high frequency in which it delivers results impossible to interpret as a probability.

The other approach used to deal with non-normally distributed coregionalizations is based on transformations: coregionalizations are assumed to follow a joint multivariate normal model *after* application of a specified marginal transformation, *e.g.*a logarithm. Then, classical methods are applied to the transformed scores, and finally interpolations are either used to define the joint model (and to compute hazard estimates) or back-transformed. This approach has a long history of application in positive and compositional coregionalizations, which are expected to become tractable after a logarithmic or a logistic transformation. However, the obtained results after back-transformation are regarded as non-optimal, since we cannot minimize simultaneously the error in the transformed and back-transformed spaces: with an example, if we

interpolate a permeability (or $[H_3O^+]$) as a positive variable by using its logarithm, we cannot simultaneously obtain an optimal estimate of this permeability (or $[H_3O^+]$) which corresponds to an optimal log-permeability (or pH), and viceversa.

1.2.3 Sample space

These considerations on the optimality properties of back-transformed estimates lead us to the keystone of this work. It deals exclusively with data which scale is captured by a particular structure of the sample space: an Euclidean structure. Briefly speaking, the *sample space* of a variable is the set of its possible results; the *scale* of a variable (or a data set) is the analyst's interpretation of how different are its values; finally, the *structure* given to the sample space is a choice of operations, and with them the analysts pretends to adequately describe the scale of the random variable. These concepts are well-known in statistics, but its practical implications in applied sciences have been seldom taken into account, until the work of Pawlowsky-Glahn [2003]. Exploring their applications and implications to the geostatistical case form the keystone of this Thesis.

Most statistical methods (both under independence assumptions, and from the geostatistical perspective) assume the data set to be drawn from the *real unbounded* Euclidean space: this is seldom explicitly said, but comes implicitly when taking as a measure of the prediction error a *squared difference*, or the squared *Euclidean distance* between the prediction and the target. In the case of parametric methods, this is much clearer, because most of them assume the data set to be generated by a normal distribution, which has an unbounded domain. Those methods developed for other models essentially work with transformations of the data, sometimes called *link functions* [Leonard and Hsu, 1999], intended to deliver real unbounded results. An example of such a procedure may be found in the definition of the lognormal distribution.

Usually, strictly positive variables are applied a logarithmic transformation, with the aim to obtain normally-distributed scores. The original variable is said then to follow a lognormal distribution, which takes into account the fact that intervals in the transformed space do not have the same length as in the original one. But the sample space of this distribution, the positive real line, can be given a real vector space structure (indeed, an Euclidean one), and standard algebra can be applied. First, a *basis* (a set of vectors univocally generating the whole space) may be chosen, and then any vector (fixed or random) of this space can be univocally expressed with its coordinates with respect to this basis. By definition, these coordinates are real and unbounded, and the Euclidean distance is well-suited for them, as is any hypothesis of normality. From this point of view, for instance, the arithmetic average of real numbers should be replaced by the geometric average when dealing with variables with a positive scale, as we will see in chapter 5.

Following the same approach, compositional data—positive vectors which components sum up to 100%, or any other fixed constant—can be treated in a new way if their sample space, the *Simplex*, is given an Euclidean space structure [Billheimer et al., 2001, Pawlowsky-Glahn and Egozcue, 2001]. This structure arises if we assume

that compositions only convey information about the relative importance of each part in a total. In this structure, a D-part composition (for instance, the content in Pb, Hg and Fe of some moss species) is expressed as a vector of real unbounded coordinates with respect to a basis formed with $D-1$ compositions. Then statistics can be applied to these real scores, and those results defining a geometric object (a mean, a confidence region, a line) may be applied to the chosen basis to recover a compositional object. Such procedure, for instance, advocates for the closed geometric mean as the best central tendency indicator of a compositional data set [Aitchison, 1982].

1.3 Structure of the Thesis

To address these concepts, this document is structured as follows.

First chapter has already outlined the main concepts involved in this study, and it will end with a presentation of the case studies used to illustrate it.

Second and third chapters are a state-of-the-art, devoted to present the founding ideas and methods used throughout the rest of the document. The second chapter is devoted to sample space considerations, based on the concept of Euclidean space, focusing on its geometric characterization and introducing probability distributions and inference techniques on it, particularly for the Simplex. A preliminary investigation is conducted on the three case studies to illustrate the concepts introduced. The third chapter summarizes existing geostatistical techniques, and develops one of the case studies.

Fourth chapter is the central part of this work, since it generalizes the main geostatistical techniques in order to deal with variables valued on an arbitrary Euclidean space. Three particular cases—with their respective case studies—are included in this Thesis: the third chapter was devoted to real variables, the fifth is centered on positive ones, and the sixth on compositional vectors.

Seventh chapter presents possible applications of the results in the preceding sections (mainly chapter six) to estimate the probability distribution of a broader class of variables, not necessarily valued on an Euclidean space.

The last chapter closes this work with a discussion on the results obtained in the analyzed case studies—focusing on the comparison of the hazard results obtained with each technique—and some methodological conclusions and open issues.

Some of these chapters begin with a short summary of their theoretical content, and all finish with a sort of preliminary conclusions regarding that specific chapter. The idea of these introductory and final summaries is to put each chapter in the general context of the work. Also, some chapters contain a final *addendum*, where some complementary explanations and proofs are included. A summary of the notation is included in a last chapter in the fashion of an appendix.

1.4 Case studies

1.4.1 The Tordera river at Gualba (Spain)

The Tordera River is located in north-eastern Spain, in the Catalan provinces of Barcelona and Girona. It drains a basin of 835 km^2, between three mountain systems (figure 1.1), the Montseny Massif, the Montnegre Range and the Guilleries Range. The studied station is located at UTM 461050.01 easting and 4618241.4 northing, in the municipality of Gualba. It is placed in the upper valley, and its catchment area represents approximately a fourth of the total basin, with contributions from the two first mountain systems. These are similar granitic massifs, with some metamorphic pelitic rocks—from phillite to orthogneiss—and small marble outcrops. The river itself flows through quaternary siliceous infills.

Most of the basin surface (figure 1.2) is occupied with meadows and woods, some of them under special protection plans (Montseny Natural Park and Montnegre-Corredor Natural Park). The Sant Celoni waste-water treating plant dumps its effluents upstream of this station. This village had 14278 registered inhabitants in year 2004; it is a small industrial center, with chemical industries and pharmaceutical facilities [Idescat, 2005, April 15]. These industries have their own waste-water treating plants, which also dump into the river. The human impact on the river is considered as moderate.

The Gualba station belongs to the XACQA (water quality automatic control network), integrated by 33 stations distributed along the main rivers in Catalunya. It is the only one of this kind in the Tordera basin. At these stations, the river water is sampled almost continuously, measuring some parameters to monitor urban and industrial pollution: pH, water temperature, Ammonia content, dissolved Oxygen content, conductivity, cloudiness, etc. These measures are used to define some categories of acceptable water uses [Poch, 1999]. Table 1.1 shows them as function of the parameters studied here, which were kindly provided by Lluís Godé from the *Agència Catalana de l' Aigua* (ACA, Catalan Water Control Agency).

Table 1.1: Water quality categories used by the ACA (Catalan Water Control Agency), as functions of conductivity, pH and ammonium content [Poch, 1999].

		Conductivity		pH		Ammonium	
category	uses	min	max	min	max	min	max
1	quality-demanding uses	0	1000	6.5	8.5	0	0.05
2	general uses	0	1000	6.5	8.5	0.05	1
3	non-demanding uses	0	1000	6.5	8.5	1	4
4	minimal uses	1000	2500	6.5	9.0	4	20
		$\mu S/cm$		—		mg/l	

In this context, focus is mainly put on monitoring Nitrogen, due to its key importance in the eutrophication processes. Eutrophication is an uncontrolled increase of

algae populations in rivers or lakes due to too much availability of Nitrogen and Phosphor, the main nutrients controlling the growth of living beings. In granitic basins, Phosphor is naturally present, and the limiting factor becomes Nitrogen. Independently of its eutrophication action, Ammonia (NH_3) is an interesting parameter in itself: apart from the limits in table 1.1, Spanish legal dispositions order Ammonia (not Ammonium) content to be kept below 0.025ppm, due to its poisonous character [Mapfre, 2000].

Ammonia (NH_3) is very difficult to measure, due to its volatility properties. It is kept in aqueous solution in the form of Ammonium (NH_4^+), which is far less dangerous. Ammonium behaves as a weak acid, and returns to Ammonia form due to the equilibrium equation

$$NH_4^+ + H_2O \rightleftharpoons NH_3 + H_3O^+,$$

characterized by the equilibrium constant

$$\frac{[NH_3] \cdot [H_3O^+]}{\left[NH_4^+\right]} = K_a, \tag{1.1}$$

being $[X]$ the molar concentration or molarity (mol/l) of species X. The equilibrium constant is inversely related to the absolute temperature due to thermodynamic relations. It can nevertheless be reasonably approximated by a polynomial [Martí, 2004, pers. comm.]

$$p\mathrm{K}_a = 4 \cdot 10^{-8} \cdot T^3 + 9 \cdot 10^{-5} \cdot T^2 - 3.56 \cdot 10^{-2} \cdot T + 10.072, \tag{1.2}$$

where $p\mathrm{K}_a = -\log_{10} K_a$, and T is measured in Celsius degrees. Using this decimal logarithm, expression (1.1) becomes

$$p\mathrm{NH}_3 = p\mathrm{K}_a + p\mathrm{NH}_4 - p\mathrm{H}, \tag{1.3}$$

with $pX = -\log_{10}[X]$. Then, although it can be hardly measured, the Ammonia content may be computed using expressions (1.2) and (1.3), once Ammonium content, pH and water temperature are known.

Tolosana-Delgado [2004] showed that the Ammonia system is strongly affected in this river by a periodic drift, mainly of 24h-period, which was suggested to be related to solar radiation through water temperature and dissolved oxygen content. This would imply that fluctuation of chemical parameters in this river may not only be caused by humans, but also be due to the natural dynamics of its ecosystem.

The goal will be the inference of the hazard of Ammonia pollution—exceedance of the 0.025ppm legal threshold—as a function of the measured parameters, as well as the assessment of water quality as a function of these parameters, following table 1.1, and taking into account possible periodic drifts.

1.4.2 Air pollution in the Carpathian Range (Ukraine)

The Carpathian Range crosses eastern Europe describing a 1500 km long arch from the Czech Republic to Romania, crossing Poland and Ukraine, and surrounding Hungary

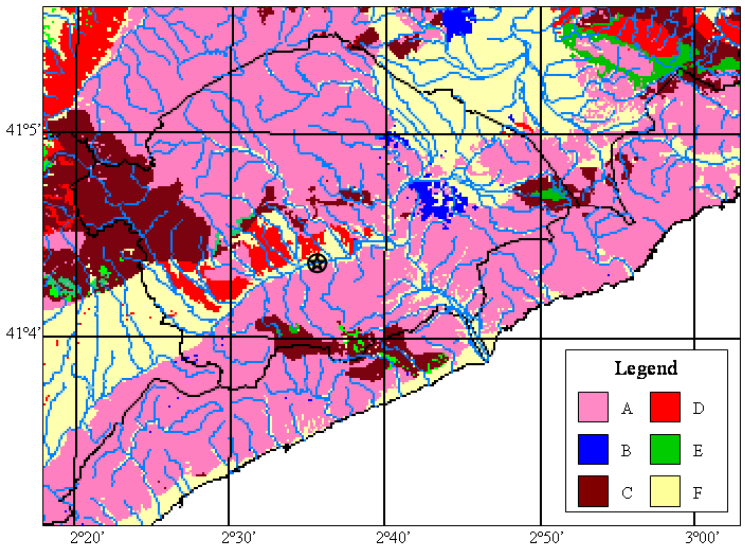

Figure 1.1: Lithologic map of the Tordera basin, distinguishing between: (A) granitic and other acid-intermediate plutonic rocks, (B) other igneous rocks (mainly lava flows), (C) metasedimentary and metamorphic siliceous rocks, (D) siliceous sedimentary series, (E) carbonate and meta-carbonare rocks, and (F) Quaternary infills. Note that (E) materials represent the only major source of HCO_3^-. The star is placed in the studied station.

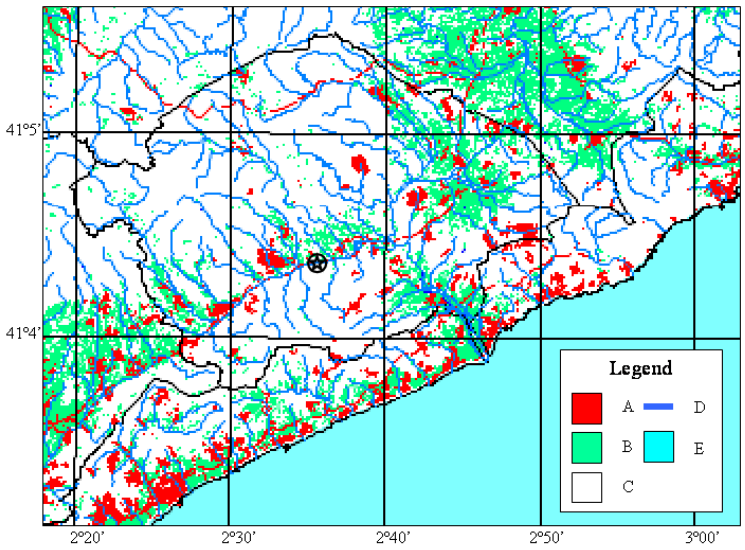

Figure 1.2: Map of land uses of the Tordera basin, distinguishing between: (A) urban areas and structures, (B) agricultural areas, (C) forests and natural areas, (D) rivers and continental water, and (E) the Mediterranean sea. The star is placed in the studied station.

Figure 1.3: Location of the Carpathian range.

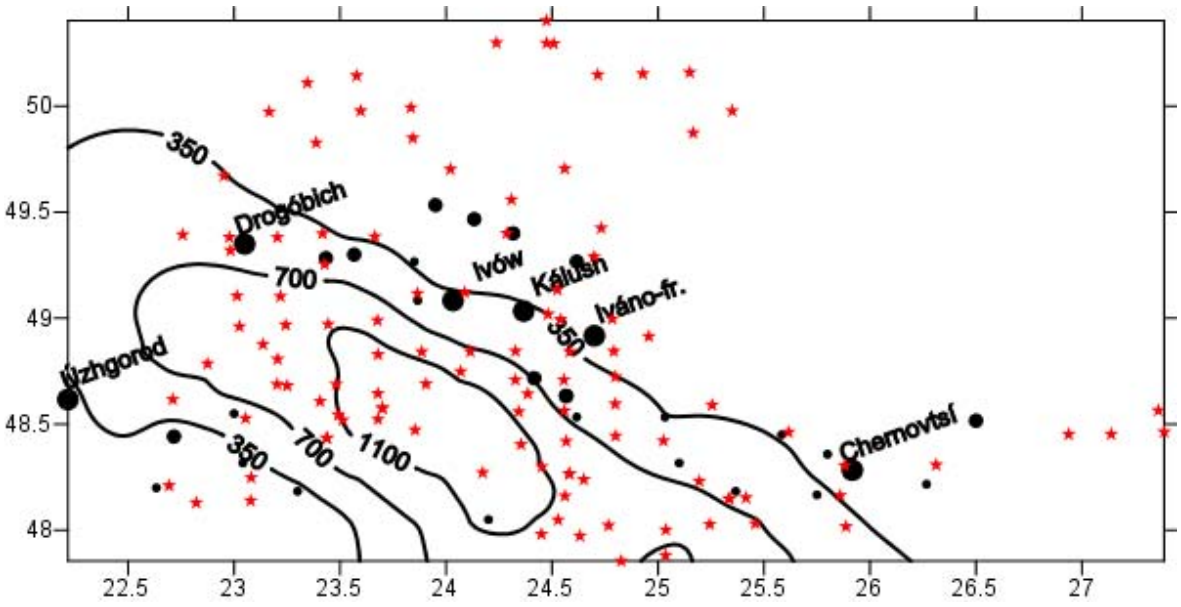

Figure 1.4: Mean features of the studied region of the Ukrainian Carpathian range: height curves, main cities (dots, size shows three categories of cities according to their industrial importance and size) and sampling locations (stars).

(figure 1.3). Although it is not a high range, it has an important influence on the winds of the region and, consequently, on air pollution dynamics. To monitor these dynamics in the Ukrainian part of the Carpathians (figure 1.4), a sample of two species of moss—regarded as proxies for air pollution—was collected: at each sampled location, five whole moss individuals were mixed and processed to analyze their content in several metals: Cd, Pb, Cr, Fe, Hg, etc. Since these plants live several years, results might be interpreted as the average pollution in the last 3-5 years. This data set and the information gathered here were kindly provided by Dr. Tyutyunnik Yulian Gennadievich, with permission from Dr. Blum Oleg Borisovich, chief of the Laboratory of Bio-indication of the National Botanic Garden of the Ukrainian National Science Academy.

From this data set three components have been selected —Pb, Fe and Hg—due to their known connection with major air pollution phenomena in the region. Lead (Pb) is a by-product of combustion of petrol fuels. Regulations about lead content in fuels were not fully implanted in Ukraine during the sampling period. It forms small-size particle aerosols, with a medium transportability, and it is expected to have a strong influence around cities and along the main roads. Iron (Fe) particles—related to open-air corrosion processes—are transported through the air also in aerosols, but of bigger particles: they are consequently more difficult to carry. Iron is then expected to be found mainly around cities. Finally, quicksilver (Hg) is transported through the air dissolved in water vapor, and is deposited with rain. Thus, Hg is highly transportable, up to thousands of kilometers; in fact, in this area it is considered a regional pollutant, originating in the industrial areas of central Europe [Tyutyunnik, 2005, pers. comm.]. The goal in this data set will be to assess the relative influences of these three processes. The relative character is highlighted because the total amount of the studied three parts may be related to exposure time, and to age of collected plants, thus masking absolute intensity of pollution.

Chapter 2

Preliminary concepts from Algebra and Statistics

Real data are uncertain: experimental values are affected by different sources of randomness, models may not take into account important processes, or measurements might be affected by instrumental errors. We usually model such a situation with random variables, from which one wants to extract a central characteristic value and a dispersion indicator. The range of values considered as possible outcomes of such a variable is called the sample space. Most interesting sample spaces can be given a structure, describing meaningful uncertainty-generating processes (or operations). Here and throughout this Thesis, the term "*meaningful*" expresses the subjective assessment of the analyst on how the structure given to a sample space describes the scale considered for the observed data set.

In this chapter the algebraic structure known as Euclidean space is summarized, with its operations and elements. This part can be found in any first-course Algebra textbook, *e.g.* in Rojo [1986]. On an Euclidean space structure, some measure concepts have been introduced, basically those related to random variables and its moments. Furthermore, the Normal distribution on an Euclidean space have been defined [Eaton, 1983, Pawlowsky-Glahn, 2003]. Finally, inference on the parameters of this normal distribution is outlined, closely following the classical inference approach as can be found in any basic bayesian statistics textbook, *e.g.* in Leonard and Hsu [1999]. Three data sets are used to illustrate the concepts of sample space, scale and structure, as well as the estimation of central tendency parameters and hazard probabilities. To close this chapter, a summary of some existing distributions to deal with data in the Simplex is included.

2.1 General basic notation

Throughout this work, a set of objects is represented with double uppercase characters, like $\mathbb{E}$, $\mathbb{F}$ or $\mathbb{S}$; in particular, $\mathbb{R}$ will denote the set of all real numbers, and $\mathbb{R}^D$ the D-dimensional real space. Also, the following notation will be used. The elements of one

of these sets $\mathbf{a} \in \mathbb{E}$ will be denoted by lowercase boldface Latin characters, but the real scalar values $\alpha \in \mathbb{R}$ by lowercase Greek characters. A finite set of elements will be denoted by an uppercase boldface Latin character, $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N\}$. A simple underlining of a lowercase Greek character will represent a vector of real values $\underline{\alpha} \in \mathbb{R}^D$, and a double underlining of a character $\underline{\underline{\varphi}}$, a matrix of real coefficients. Furthermore, an operator $T(\cdot)$ (either a Greek or an uppercase Latin character) acting in an element $\mathbf{a}$ will be represented by $T\mathbf{a} = T(\mathbf{a})$. The next section introduces further notation related to Euclidean spaces. A complete notation summary is included in the last chapter.

2.2 Geometry of the sample space

2.2.1 Vector space

Definition 2.1 (Real vector space) *Let $\mathbb{E}$ be a set, and let $(\mathbb{R}, +, \cdot)$ represent the* scalar field *of real numbers, with classical sum and product defined on it. The set $\mathbb{E}$ equipped with two closed operations, called for convention* inner sum *and* external product *and denoted respectively by $\oplus$ and $\odot$, is called a* real vector space *(or, simply, vector space) if the following properties are satisfied. For any $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathbb{E}$ and $\lambda, \mu \in \mathbb{R}$,*

- *commutative inner sum: $\boldsymbol{a} \oplus \boldsymbol{b} = \boldsymbol{b} \oplus \boldsymbol{a}$,*
- *associative inner sum: $(\boldsymbol{a} \oplus \boldsymbol{b}) \oplus \boldsymbol{c} = \boldsymbol{a} \oplus (\boldsymbol{b} \oplus \boldsymbol{c})$,*
- *existence of a neutral element with respect to the inner sum: $\boldsymbol{a} \oplus \boldsymbol{n} = \boldsymbol{a}$,*
- *existence of an opposite element for any other: $\boldsymbol{a} \oplus \bar{\boldsymbol{a}} = \boldsymbol{n}$; one shall denote alternatively $\bar{\boldsymbol{a}} = \ominus \boldsymbol{a}$, and interpret $\ominus$ as the inverse operation of $\oplus$,*
- *external product distributive with respect to the inner sum: $\lambda \odot (\boldsymbol{a} \oplus \boldsymbol{b}) = \lambda \odot \boldsymbol{a} \oplus \lambda \odot \boldsymbol{b}$,*
- *external product distributive with respect to the sum of the scalar field: $(\lambda + \mu) \odot \boldsymbol{a} = \lambda \odot \boldsymbol{a} \oplus \mu \odot \boldsymbol{a}$,*
- *external product associative with the product of the scalar field: $\lambda \odot (\mu \odot \boldsymbol{a}) = (\lambda \cdot \mu) \odot \boldsymbol{a}$,*
- *existence of a neutral element in the scalar field with respect to the external product: $1 \odot \boldsymbol{a} = \boldsymbol{a}$.*

Then, the elements of $\mathbb{E}$ are called vectors.

The properties stated allow for an extensive and unambiguous treatment of all vectors in $\mathbb{E}$ through the concepts of linear combination and linear independence.

Definition 2.2 *Given a set* $\boldsymbol{E} \subset \mathbb{E}$ *formed with* D *vectors,* $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_D\}$*, and a set of scalar* coefficients $\lambda_1, \lambda_2, \ldots, \lambda_D$*, a linear combination of* $\boldsymbol{E}$ *is a vector* $\boldsymbol{b} \in \mathbb{E}$ *computed as*

$$\boldsymbol{b} = \lambda_1 \odot \boldsymbol{e}_1 \oplus \lambda_2 \odot \boldsymbol{e}_2 \oplus \cdots \oplus \lambda_D \odot \boldsymbol{e}_D = \bigoplus_{i=1}^{D} \lambda_i \odot \boldsymbol{e}_i. \tag{2.1}$$

Using this, the following concepts are introduced:

1. if there are no λ_i which allow to obtain $\mathbf{b}$ with (2.1), then $\mathbf{b}$ is *linearly independent* of $\mathbf{E}$; otherwise, when these λ_i exist, then $\mathbf{b}$ is said to be *linearly dependent* of $\mathbf{E}$,

2. $\mathbf{E}$ is said to be linearly independent if all its vectors are linearly independent of the others,

3. if all the vector in $\mathbb{E}$ are linearly dependent of $\mathbf{E}$, then $\mathbf{E}$ is called a *generating system* of $\mathbb{E}$,

4. if $\mathbf{E}$ is a linearly independent generating system, then it is called a *basis* of $\mathbb{E}$, and the number of vectors in $\mathbf{E}$ is identified with the *dimension* of $\mathbb{E}$.

Definition 2.3 *The unique values* λ_i *needed to obtain a vector* $\boldsymbol{b}$ *as a linear combination of a basis* $\boldsymbol{E}$ *are called the* coordinates of $\boldsymbol{b}$ in the basis $\boldsymbol{E}$.

Definition 2.4 *Let* $\mathbb{F}$ *be a subset of* $\mathbb{E}$*. If any linear combination (2.1) of vectors of* $\mathbb{F}$ *is included in* $\mathbb{F}$*, then* $\mathbb{F}$ *is called a* vector subspace *of* $\mathbb{E}$.

All properties and concepts defined in this section can be equally applied to vector subspaces with the same inner sum and external product operations.

Definition 2.5 (Scalar product) *Let* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ *be elements of* $\mathbb{E}$*, and* $\lambda \in \mathbb{R}$*; then, any function* $\langle \cdot, \cdot \rangle_{\mathbb{E}} : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ *is called a* scalar product *between elements of* $\mathbb{E}$ *if it satisfies the following conditions*

1. *it is symmetric,* $\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\mathbb{E}} = \langle \boldsymbol{b}, \boldsymbol{a} \rangle_{\mathbb{E}}$,

2. *it is linear with respect to the vector space operations,*
   $\langle \boldsymbol{a} \oplus \lambda \odot \boldsymbol{b}, \boldsymbol{c} \rangle_{\mathbb{E}} = \langle \boldsymbol{a}, \boldsymbol{c} \rangle_{\mathbb{E}} + \lambda \langle \boldsymbol{b}, \boldsymbol{c} \rangle_{\mathbb{E}}$,

3. *the scalar product of a vector with itself is positive,* $\langle \boldsymbol{a}, \boldsymbol{a} \rangle_{\mathbb{E}} \geq 0$,

4. *the scalar product of a vector with itself is zero only for the neutral element,* $\langle \boldsymbol{a}, \boldsymbol{a} \rangle_{\mathbb{E}} = 0 \Leftrightarrow \boldsymbol{a} = \boldsymbol{n}$.

Definition 2.6 (Euclidean space) *A set* $\mathbb{E}$ *is a* D*-dimensional Euclidean space if it is a* D*-dimensional real vector space equipped with a scalar product. It is usually denoted by the compact form* $\{\mathbb{E}, \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathbb{E}}\}$.

Definition 2.7 *In an Euclidean space, the* norm $\|\cdot\|_{\mathbb{E}} : \mathbb{E} \to \mathbb{R}_+$ *of a vector is defined as* $\|\boldsymbol{a}\|_{\mathbb{E}} = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle_{\mathbb{E}}}$. *By using the norm of the difference, one can define a distance in an Euclidean space, as* $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} \ominus \boldsymbol{b}\|_{\mathbb{E}}$.

Property 2.1 *Let* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ *be elements of* $\mathbb{E}$, *then the distance function*

$$d(\cdot, \cdot) : \mathbb{E} \times \mathbb{E} \to \mathbb{R}_+$$

computed as $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} \ominus \boldsymbol{b}\|_{\mathbb{E}}$ *satisfies the following conditions:*

1. *it is positive,* $d(\boldsymbol{a}, \boldsymbol{b}) \geq 0$,
2. $d(\boldsymbol{a}, \boldsymbol{b}) = 0 \Leftrightarrow \boldsymbol{a} = \boldsymbol{b}$,
3. *it is symmetric,* $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$,
4. *it satisfies the triangular inequality,* $d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{c}, \boldsymbol{b}) \geq d(\boldsymbol{a}, \boldsymbol{b})$.
5. *it is invariant by translation (or application of the inner sum operation),* $d(\boldsymbol{a} \oplus \boldsymbol{c}, \boldsymbol{b} \oplus \boldsymbol{c}) = d(\boldsymbol{a}, \boldsymbol{b})$,
6. *it is scaled by the application of an external product,* $d(\lambda \odot \boldsymbol{a}, \lambda \odot \boldsymbol{b}) = |\lambda| \cdot d(\boldsymbol{a}, \boldsymbol{b})$.

The cosine of an *angle* θ between two vectors $\mathbf{a}$ and $\mathbf{b}$ is given by

$$\cos(\theta) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{E}}}{\|\mathbf{a}\|_{\mathbb{E}} \cdot \|\mathbf{b}\|_{\mathbb{E}}},$$

from which the angle θ itself is obtained. This allows the introduction of the concepts of

1. *parallelism*: $\mathbf{a}$ and $\mathbf{b}$ are parallel if $\theta = 0$ or, equivalently, $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{E}} = \|a\|_{\mathbb{E}} \cdot \|b\|_{\mathbb{E}}$,
2. *orthogonality*: $\mathbf{a}$ and $\mathbf{b}$ are orthogonal if $\theta = \pi/2$ or, equivalently, $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{E}} = 0$;

Note that parallelism implies linear dependence, whereas two vectors $\mathbf{a}, \mathbf{b} \neq \mathbf{n}$ are linearly independent if they are orthogonal.

Definition 2.8 (Orthogonal basis) *A basis* $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_D\}$ *is said to be* orthogonal *if all its elements are orthogonal to each other.*

Any vector $\mathbf{b} \in \mathbb{E}$ can be univocally expressed as a linear combination (2.1) of the elements of such an orthogonal basis by

$$\mathbf{b} = \bigoplus_{i=1}^{D} \beta_i \odot \mathbf{e}_i; \quad \beta_i = \frac{\langle \mathbf{b}, \mathbf{e}_i \rangle_{\mathbb{E}}}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}}, \tag{2.2}$$

where the values β_i are the coordinates with respect to that basis.

Definition 2.9 (Orthonormal basis) *A basis is said to be* orthonormal *if it is orthogonal and its elements have unitary norm,*

$$\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle_{\mathbb{E}} = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

The use of orthonormal bases further simplifies the computation of the coordinates β_i of a vector $\mathbf{b}$ with respect to a basis $\mathbf{E}$, since $\beta_i = \langle \mathbf{b}, \mathbf{e}_i \rangle_{\mathbb{E}}$ in equation (2.2).

Property 2.2 *Once an orthonormal basis is specified, any element of the space $\mathbb{E}$ is univocally determined by the real vector of its coordinates $\underline{\beta} = (\beta_i) \in \mathbb{R}^D$ in that basis, and all vector operations in the space can be defined as follows:*

1. *if $\boldsymbol{g} = \boldsymbol{a} \oplus \boldsymbol{b}$ then $\underline{\gamma} = \underline{\alpha} + \underline{\beta}$,*
2. *if $\boldsymbol{b} = \lambda \odot \boldsymbol{a}$ then $\underline{\beta} = \lambda \cdot \underline{\alpha}$,*
3. *$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\mathbb{E}} = \langle \underline{\alpha}, \underline{\beta} \rangle_{\mathbb{R}} = \sum_{i=1}^{D} \alpha_i \cdot \beta_i$,*
4. *$\| \boldsymbol{a} \|_{\mathbb{E}} = \| \underline{\alpha} \|_{\mathbb{R}} = \sqrt{\sum_{i=1}^{D} \alpha_i^2}$,*
5. *$d_{\mathbb{E}}(\boldsymbol{a}, \boldsymbol{b}) = d_{\mathbb{R}}(\underline{\alpha}, \underline{\beta}) = \sqrt{\sum_{i=1}^{D} (\alpha_i - \beta_i)^2}$,*

where $\underline{\alpha}, \underline{\beta}$ and $\underline{\gamma}$ represent the vectors of coefficients of respectively $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{g}$ in the given basis. In other words, given an orthonormal basis, any D-dimensional Euclidean space $\{\mathbb{E}, \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathbb{E}}\}$ and its space of coordinates $\{\mathbb{R}^D, +, \cdot, \langle \cdot, \cdot \rangle_{\mathbb{R}}\}$ are completely equivalent.

2.2.2 Linear applications

Definition 2.10 (Linear transformation) *Let $\{\mathbb{E}, \oplus, \odot\}$ and $\{\mathbb{F}, +, \cdot\}$ be two different vector spaces, with different rules of addition and product with a scalar. Then, an application $T(\cdot)$ from $\mathbb{E}$ onto $\mathbb{F}$ is called a linear transformation if and only if it satisfies*

$$T(\boldsymbol{a} \oplus \lambda \odot \boldsymbol{b}) = T\boldsymbol{a} + \lambda \cdot T\boldsymbol{b}$$

for any $\boldsymbol{a} \in \mathbb{E}, \boldsymbol{b} \in \mathbb{F}$ and $\lambda \in \mathbb{R}$.

Some short comments about linear transformations follow.

- With the same vector operations of $\{\mathbb{F}, +, \cdot\}$, the set of linear applications $L(\mathbb{E}, \mathbb{F})$ is itself a vector space, denoted by $\{L(\mathbb{E}, \mathbb{F}), +, \cdot\}$.
- Fixed suitable basis for $\mathbb{E}$ and $\mathbb{F}$ (with respectively dimensions D and C), the linear transformation T is univocally identified by a $C \times D$ matrix of real coefficients $\underline{\underline{T}}$.

- If the vectors $\mathbf{a}$ and $\mathbf{b}$ are represented in these respective basis as $\underline{\alpha}$ and $\underline{\beta}$, column matrices, then
$$\mathbf{b} = T\mathbf{a} \Leftrightarrow \underline{\beta} = \underline{\underline{T}} \cdot \underline{\alpha}.$$
- Linear applications can be composed, and the result is still a linear application. For instance, if $B : \mathbb{V}_1 \to \mathbb{V}_2$ and $C : \mathbb{V}_2 \to \mathbb{V}_3$, then we may define a new linear transformation $(CB) \equiv C \circ B \equiv A : \mathbb{V}_1 \to \mathbb{V}_3$.
- Given suitable basis for the spaces $\mathbb{V}_1, \mathbb{V}_2, \mathbb{V}_3$, the following relation between the matrix representation of these applications is satisfied: $\underline{\underline{A}} = \underline{\underline{C}} \cdot \underline{\underline{B}}$.

Definition 2.11 (Adjoint transformation) *Let $\{\mathbb{E}, \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathbb{E}}\}$ and $\{\mathbb{F}, +, \cdot, \langle\cdot,\cdot\rangle_{\mathbb{F}}\}$ be two different Euclidean spaces. Then, for any linear transformation $T(\cdot) : \mathbb{E} \to \mathbb{F}$, there exists another linear transformation $T^t(\cdot) : \mathbb{F} \to \mathbb{E}$ such that*
$$\langle T\boldsymbol{a}, \boldsymbol{b}\rangle_{\mathbb{F}} = \langle \boldsymbol{a}, T^t\boldsymbol{b}\rangle_{\mathbb{E}}$$
for any $\boldsymbol{a} \in \mathbb{E}, \boldsymbol{b} \in \mathbb{F}$. The linear transformation T^t is called adjoint transformation *of T, and vice-versa.*

Definition 2.12 (Endomorphism on $\mathbb{E}$) *Let $\mathbb{E}$ be a D-dimensional vector space. A linear transformation $T(\cdot)$ from and onto $\mathbb{E}$ is called an* endomorphism.

Regarding endomorphisms, note the following comments.

- Once fixed a basis $\mathbf{E}$, an endomorphism may be represented by a $D \times D$ matrix.
- If we change the basis of representation to $\mathbf{F}$, we may obtain the new matrix expression of the endomorphism as $\underline{\underline{\varphi}}^{-1} \cdot \underline{\underline{T}} \cdot \underline{\underline{\varphi}}$, where $\underline{\underline{\varphi}}$ contains the coordinates of the vectors of $\mathbf{F}$ with respect to the basis $\mathbf{E}$.
- Composition of endomorphisms is an inner operation of the space $L(\mathbb{E}, \mathbb{E})$, since it takes two endomorphisms and returns another endomorphism. Therefore this space has further structure than a simple vector space. Note that it plays de role of a product, and is associative but not commutative, like the product of matrices.

Two particular endomorphisms are interesting, once we introduce the composition operation.

Definition 2.13 (Identity on $\mathbb{E}$) *Let $\mathbb{E}$ be a D-dimensional vector space, and $L(\mathbb{E}, \mathbb{E})$ the space of endomorphisms on $\mathbb{E}$. Then, there exists an application $I(\cdot) \in L(\mathbb{E}, \mathbb{E})$ such that for any $T(\cdot) \in L(\mathbb{E}, \mathbb{E})$, the composite satisfies $IT = TI = T$. The matrix representation of this endomorphism is the identity matrix of D columns and rows.*

Definition 2.14 (Inverse endomorphism) *Let $\mathbb{E}$ be a D-dimensional vector space, and $L(\mathbb{E}, \mathbb{E})$ the space of endomorphisms on $\mathbb{E}$. If a pair of applications $T_1(\cdot), T_2(\cdot) \in L(\mathbb{E}, \mathbb{E})$ satisfy that their composition is $T_1T_2 = T_2T_1 = I$ the identity, they are called* invertible*, and one is considered the inverse of the other, denoted by $T_1^{-1} = T_2$. The matrix representations of these endomorphisms are mutually inverse matrices.*

2.3 Probability laws on coordinates

2.3.1 Measure considerations

Notation

In this section, $\mathbb{E} \subset \mathbb{R}^D$ is taken as a C-dimensional Euclidean space included in the D-dimensional real space, with $C \leq D$. We will identify $\underline{\alpha}$ and $\underline{\beta}$ as the vectors of coordinates of $\mathbf{a}$ and $\mathbf{b}$ in a basis of $\mathbb{E}$, and $\underline{a}$ and $\underline{b}$ as real vectors in $\mathbb{R}^D$.

Lebesgue measure on $\mathbb{R}^D$

The *Lebesgue measure* of an interval $(a, b) \subset \mathbb{R}$ is defined as the length of the interval, or the distance between its extreme points, $\lambda(a, b) = d_{\mathbb{R}}(a, b) = |a - b|$. To extend the concept to higher dimensions, first the D-interval must be defined: it is the hyper-rectangle defined by two extreme points of a diagonal and sides parallel to the axis defined by the basis of $\mathbb{R}^D$. The Lebesgue measure of a D-interval defined by vectors $\underline{a}$ and $\underline{b}$ is simply the product of the length of each side of the hyper-rectangle,

$$\lambda_{\mathbb{R}^D} = \lambda(\underline{a}, \underline{b}) = \prod_{i=1}^{D} |a_i - b_i|.$$

Measure and other related concepts are the concern of Measure Theory, a subject treated by many advanced manuals, see for instance [Nielsen, 1997]. The approach of this Thesis is much simpler, focusing on the definition of alternative measures, how to change between them, and their influence in inference procedures.

Lebesgue measure on $\mathbb{E}$

Definition 2.15 *The Lebesgue measure of a C-interval defined by two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in an arbitrary Euclidean space $\mathbb{E}$ with respect to a basis $\boldsymbol{E}$ is*

$$\lambda_{\mathbb{E}} = \lambda(\boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{C} |\alpha_i - \beta_i|, \tag{2.3}$$

where α_i and β_i are the coordinates of $\boldsymbol{a}$ and $\boldsymbol{b}$.

Finally, the Lebesgue measure of a subset $A \subset \mathbb{E}$ is defined as the Lebesgue measure of the subset of $\mathbb{R}^C$ formed with the coefficients of A, extending the classical definition of $\lambda(A)$. As stated by Pawlowsky-Glahn [2003], this connection between the geometric space structure of $\mathbb{E}$ and its Lebesgue measure does not imply that there is just one measure for $\mathbb{E}$. For instance, when $\mathbb{E} \subset \mathbb{R}^D$ one can apply the measure of $\mathbb{R}^D$ to the elements of $\mathbb{E}$.

Probability measures

The kind of measures most often used throughout this work are probability measures, denoted by $P(\cdot)$. This well-established subject is treated by many textbooks, like in Nielsen [1997], already cited, or more specifically, in Rényi [1976]. Probability considers the measure of the whole space to be one, $P(\mathbb{E}) = 1$. They are interpreted in conjunction with a random vector $\mathbf{Z} \in \mathbb{E}$, and the measure of a subset A is equivalent to the probability $\Pr[\mathbf{Z} \in A] = P(A)$ that the random vector falls inside A. We can therefore define a function $F_{\mathbf{Z}}(\cdot)$ that relates each possible subset $A \subset \mathbb{E}$ with its probability measure: $F_{\mathbf{Z}}(A) = \Pr[\mathbf{Z} \in A] = P(A)$. This function is called a *probability law*, and its relationship to the random vector is explicitly encoded in the subindex $\mathbf{Z}$. It is more useful to characterize the probability measure by its *density function*, a function

$$\begin{array}{rccc} f_{\mathbf{Z}}(\cdot): & \mathbb{E} & \rightarrow & \mathbb{R}_+ \\ & \mathbf{z} & & f_{\mathbf{Z}}(\mathbf{z}), \end{array}$$

so that

$$P(A) = \int_A dP(\mathbf{z}) = \int_A f(\mathbf{z})\lambda(\mathbf{z}),$$

where the measure $P(\cdot)$ is said to be *dominated* by the measure $\lambda(\cdot)$, which means that wherever $\lambda(\mathbf{z}) = 0$ then $P(\mathbf{z}) = 0$. The function $f(\mathbf{z})$ tells us how dense is the probability measure of the random vector $\mathbf{Z}$ around each vector $\mathbf{z} \in \mathbb{E}$, or which set of values is more or less likely to occur. The density function can be obtained from

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{dF_{\mathbf{Z}}(\mathbf{z})}{d\lambda(\mathbf{z})} = f_{\mathbf{Z},\lambda}(\mathbf{z}), \tag{2.4}$$

called the *Radon-Nykodym derivative* of $F_{\mathbf{Z}}(\mathbf{z})$ with respect to the measure λ. If we change the measure used in the space, the probability density will change, whereas the probability law will remain the same. There exists nevertheless an easy relationship between the two densities. For a given pair of measures λ_1 and λ_2 (where λ_2 is dominated by λ_1), the following relationship holds

$$f_{\mathbf{Z},\lambda_1}(\mathbf{z}) = \frac{dF_{\mathbf{Z}}(\mathbf{z})}{d\lambda_1(\mathbf{z})} = \frac{d\lambda_2(\mathbf{z})}{d\lambda_1(\mathbf{z})} \cdot \frac{dF_{\mathbf{Z}}(\mathbf{z})}{d\lambda_2(\mathbf{z})} = \frac{d\lambda_2(\mathbf{z})}{d\lambda_1(\mathbf{z})} \cdot f_{\mathbf{Z},\lambda_2}(\mathbf{z}). \tag{2.5}$$

To avoid carrying both the index for the random vector and for the measure used, a measure will be always specified (either the Lebesgue measure in $\mathbb{R}^D$ or in $\mathbb{E}$) and denote a density of a random vector $\mathbf{Z}$ by $f_{\mathbf{Z}}(\mathbf{z})$.

2.3.2 First and second-order moments

Moments on coordinates

Proposition 2.1 *If $\boldsymbol{Z}$ is a random vector in $\mathbb{E}$ a C-dimensional Euclidean space, then its vector of coordinates $\underline{\zeta}$ is also a random vector but defined in $\mathbb{R}^C$.*

As a random vector, $\mathbf{Z}$ may be characterized through its moments.

Definition 2.16 (Moments on $\mathbb{E}$) *The moment coordinates of a random variable in $\mathbb{E}$ are defined as the equivalent real moments of the coordinate random variable, if they exist [Pawlowsky-Glahn, 2003].*

Pawlowsky-Glahn [2003] suggests to apply these moments to the basis and recover elements of the space, which are called *characteristic elements*. The moments which are not expressed as vectors of the space are called then *characteristic measures*. In particular, we will be interested in characteristic *element of central tendency*, and in *measures of dispersion* or spread.

However, the analyst might ask whether results will not depend on the chosen basis. The next sections define means and variances-covariances as objects in the Euclidean structure of $\mathbb{E}$, following Eaton [1983]. These objects are not basis-dependent. Then, we proof that, given a basis, object definitions may be identified with coordinate ones, proving that the coordinate approach gives the same result whatever basis is used. One should take into account that, although means can be expressed in any basis, variances and covariances should only be expressed in *orthonormal* basis. The discussion of this limitation is beyond the scope of this thesis.

Expectation on $\mathbb{E}$

Definition 2.17 (Expectation on $\mathbb{E}$) *The expectation (or mean) of a random variable $\boldsymbol{Z}$ in a C-dimensional Euclidean space is a vector $\bar{\boldsymbol{z}} \equiv \mathrm{E}_{\mathbb{E}}[\boldsymbol{Z}]$ satisfying for any vector $\boldsymbol{x} \in \mathbb{E}$*

$$\langle \boldsymbol{x}, \bar{\boldsymbol{z}} \rangle_{\mathbb{E}} = \mathrm{E}\left[\langle \boldsymbol{x}, \boldsymbol{Z} \rangle_{\mathbb{E}}\right], \tag{2.6}$$

where $\mathrm{E}[\cdot]$ is a real expectation [Eaton, 1983].

Proposition 2.2 *Let $\boldsymbol{Z}$ be a random variable in $\mathbb{E}$ an Euclidean space. Then, the mean of the coordinates of $\boldsymbol{Z}$ equal the coordinates of the mean of $\boldsymbol{Z}$ on $\mathbb{E}$ with respect to any basis*

A proof can be found in Eaton [1983, p. 72], or in this chapter *addendum*.

Variance and Covariance on $\mathbb{E}$

Definition 2.18 (Variance on $\mathbb{E}$) *The variance of a random variable $\boldsymbol{Z}$ in an Euclidean space is an endomorphism $\Sigma \equiv \mathrm{Var}_{\mathbb{E}}[\boldsymbol{Z}]$ satisfying for any pair of vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{E}$*

$$\langle \boldsymbol{x}, \Sigma \boldsymbol{y} \rangle_{\mathbb{E}} = \mathrm{E}\left[\langle \boldsymbol{x}, \boldsymbol{Z} \ominus \bar{\boldsymbol{z}} \rangle_{\mathbb{E}} \langle \boldsymbol{y}, \boldsymbol{Z} \ominus \bar{\boldsymbol{z}} \rangle_{\mathbb{E}}\right], \tag{2.7}$$

where $\mathrm{E}[\cdot]$ is a real expectation [Eaton, 1983].

Proposition 2.3 *Let $\boldsymbol{Z}$ be a random variable on $\mathbb{E}$ an Euclidean space. Then, the variance matrix $\underline{\underline{\Sigma}}$ of the coordinates of $\boldsymbol{Z}$ with respect to an orthonormal basis equals the matrix representation of the variance (as an endomorphism Σ) of $\boldsymbol{Z}$ on $\mathbb{E}$ with respect to that basis.*

A proof is again included in this chapter *addendum*, as well as in [Eaton, 1983, p. 73]. Summarizing, we may identify the expectation vector and the variance endomorphism of a random vector (Eaton [1983] version) with the expectation and variance of the coordinates of this random vector in an orthonormal basis (Pawlowsky-Glahn [2003] version), thus they do not truly depend on the basis. However, their representation does depend on it, and to keep expressions simple we will consider only orthonormal basis from now on. Note also that this identification may also be extended to covariance between vectors in different spaces, although we introduce here only its definition.

Definition 2.19 (Covariance) *The covariance of a random variable $\boldsymbol{Z}_1$ on another random variable $\boldsymbol{Z}_2$, each in its Euclidean space $\{\mathbb{E}, \oplus, \odot, \langle\cdot,\cdot\rangle_{\mathbb{E}}\}$ and $\{\mathbb{F}, +, \cdot, \langle\cdot,\cdot\rangle_{\mathbb{F}}\}$, is a linear transformation $\Sigma_{12} \equiv \mathrm{Cov}_{\mathbb{E}}[\boldsymbol{Z}_1, \boldsymbol{Z}_2] : \mathbb{F} \to \mathbb{E}$ satisfying for any pair of vectors $\boldsymbol{x} \in \mathbb{E}$ and $\boldsymbol{y} \in \mathbb{F}$*

$$\langle \boldsymbol{x}, \Sigma_{12}\boldsymbol{y}\rangle_{\mathbb{E}} = \mathrm{E}\left[\langle \boldsymbol{x}, \boldsymbol{Z}_1 \ominus \bar{\boldsymbol{z}}_1\rangle_{\mathbb{E}} \cdot \langle \boldsymbol{y}, \boldsymbol{Z}_2 - \bar{\boldsymbol{z}}_2\rangle_{\mathbb{F}}\right], \tag{2.8}$$

where $\mathrm{E}[\cdot]$ is a real expectation, and $\bar{\boldsymbol{z}}_i$ is the expectation of $\boldsymbol{Z}_i$, $i = 1, 2$ in its space [Eaton, 1983, p. 85].

2.3.3 Normal probability distributions

The normal distribution is the standard distribution in real geometry. It has interesting properties, and most basic statistical analysis and tests are based upon the assumption that data are outcomes of it. This section extends this special distribution and defines its equivalent on other spaces.

Again, two different definitions will be given for the normal probability distribution on an Euclidean space $\mathbb{E}$. One of them is defined on the coordinates, and takes coordinate parameters (Pawlowsky-Glahn [2003] version). The other is defined using projections, and takes as parameters a vector and an endomorphism (Eaton [1983] version). We state here that both definitions are equivalent, given the identification of means and variances done in the last section (propositions 2.2 and 2.3).

Definition 2.20 (Normal distribution on $\mathbb{E}$ with coordinate parameters) *The random vector $\boldsymbol{Z}$ is said to follow a* normal distribution on $\mathbb{E}$ *(denoted $\boldsymbol{Z} \sim \mathcal{N}_{\mathbb{E}}(\underline{\mu}, \underline{\underline{\Sigma}})$) if its vector of coefficients follow a (multivariate) normal distribution on $\mathbb{R}^C$ with coordinate mean a real vector $\underline{\mu}$ and coordinate covariance matrix a real positive-definite symmetric matrix $\underline{\underline{\Sigma}}$ [Pawlowsky-Glahn, 2003].*

Definition 2.21 (Normal distribution on $\mathbb{E}$ with object parameters) *The random vector $\boldsymbol{Z}$ is said to follow a* normal distribution on $\mathbb{E}$ *(denoted $\boldsymbol{Z} \sim \mathcal{N}_{\mathbb{E}}(\boldsymbol{m}, \Sigma)$) for a given mean vector $\boldsymbol{m}$ and a positive definite symmetric endomorphism Σ on $\mathbb{E}$ if for any testing vector $\boldsymbol{x}$ the random projection $\langle \boldsymbol{x}, \boldsymbol{Z}\rangle_{\mathbb{E}}$ follows a classical univariate normal distribution with expectation $\langle \boldsymbol{x}, \boldsymbol{m}\rangle_{\mathbb{E}}$ and variance $\langle \boldsymbol{x}, \Sigma\boldsymbol{x}\rangle_{\mathbb{E}}$.*

If the variance of $\mathbf{Z}$ is invertible, the density of this normal distribution with respect to the Lebesgue measure on $\mathbb{E}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) \equiv f_{\zeta}(\underline{\zeta}) = |\underline{\underline{\Sigma}}|^{-1} (2\pi)^{-C/2} \exp\left[-\frac{1}{2}(\underline{\zeta} - \underline{\mu})^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot (\underline{\zeta} - \underline{\mu})\right]. \tag{2.9}$$

In the object notation, this density will read

$$f_{\mathbf{Z}}(\mathbf{z}) = |\Sigma^{-1}| (2\pi)^{-C/2} \exp\left[-\frac{1}{2}\langle \mathbf{z} \ominus \mathbf{m}, \Sigma^{-1}(\mathbf{z} \ominus \mathbf{m})\rangle_{\mathbb{E}}\right].$$

If the density with respect to any other measure is sought, then equation (2.5) is enough to compute it. The normal distribution on $\mathbb{E}$ inherits all the nice properties of the normal distribution on $\mathbb{R}^C$. For instance, the characteristic measures of central tendency and dispersion of a multivariate normal random vector are respectively $\underline{\mu}$ and $\underline{\underline{\Sigma}}$. Equation (2.6) gives the characteristic element of central tendency. As is expected in a normal distribution, this element coincides with the mode or most-frequent value of (2.9), and when $C = 1$, also with its median. Further properties can be found in Pawlowsky-Glahn [2003].

A comment is due with regard positive definiteness. Symmetric matrices are known to be positive definite when their eigenvalues are all positive (or zero, in case of semi-definiteness). Positive definite endomorphisms have not been defined in this work. However, it is also known that changes of basis do not change eigenvalues of a matrix. Therefore, a positive definite endomorphism may be taken as that for which its representation in any basis is a positive definite matrix.

2.3.4 Regression on $\mathbb{E}$

The concept of regression on $\mathbb{E}$ helps in understanding the posterior developments of this Thesis. In particular, we will look for a linear prediction of a random vector $\mathbf{Z}$ on $\mathbb{E}$ either using as predictand another random vector $\mathbf{Y}$ on $\mathbb{E}$, or using a real random vector $\underline{X}$ on $\mathbb{R}^A$. Therefore, the general case will be explained following Eaton [1983], and proofs of these assertions are omitted. An alternative approach yielding the same result was presented by Daunis-i-Estadella et al. [2002].

Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be two random vectors, respectively taking values in $\{\mathbb{F}, +, \cdot\}$ and $\{\mathbb{E}, \oplus, \odot\}$. Let $\mathbf{m}_i$ and Σ_{ii} be the vector expectation and operator variance of $\mathbf{Z}_i$ (for $i = 1, 2$) given by definitions 2.17 and 2.18. Let finally Σ_{12} and Σ_{21} be the operator covariances (definition 2.19) respectively of $\mathbf{Z}_1$ on $\mathbf{Z}_2$, and of $\mathbf{Z}_2$ on $\mathbf{Z}_1$.

Let $L(\mathbb{F}, \mathbb{E})$ be the set of linear transformations from $\mathbb{F}$ to $\mathbb{E}$. We look for a linear transformation $B \in L(\mathbb{F}, \mathbb{E})$ and a constant vector $\mathbf{b}_0 \in \mathbb{E}$ such that the affine linear transformation

$$\hat{\mathbf{Z}}_2 = \mathbf{b}_0 \oplus B\mathbf{Z}_1 \tag{2.10}$$

gives predictions of $\mathbf{Z}_2$ with minimal error ϵ, defined as

$$\epsilon = \mathrm{E}\left[\|\mathbf{Z}_2 \ominus \hat{\mathbf{Z}}_2\|_{\mathbb{E}}\right]. \tag{2.11}$$

With these definitions, the optimal predictions are obtained by using the linear transformation

$$B = \Sigma_{21}\Sigma_{11}^{-1} \tag{2.12}$$

and constant

$$\mathbf{b}_0 = \mathbf{m}_2 \ominus \Sigma_{21}\Sigma_{11}^{-1}\mathbf{m}_1. \tag{2.13}$$

Recall that Σ_{11}^{-1} is the inverse application of Σ_{11}.

The joint vector $(\mathbf{Z}_1, \mathbf{Z}_2)$ has as sample space the cartesian product $\mathbb{F} \times \mathbb{E}$, with an Euclidean structure inherited from the structures of $\mathbb{F}$ and $\mathbb{E}$. Then, the definition of a joint normal is suitable. In this case, regression yields the conditional distribution of $\mathbf{Z}_2$ given $\mathbf{Z}_1 = \mathbf{z}_1$. This distribution is a normal one on $\mathbb{E}$, with conditional mean $\mathbf{m} = \mathbf{m}_2 \oplus \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{z}_1 \ominus \mathbf{m}_1)$ and conditional variance $\Sigma_{22} \ominus \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Fixed suitable orthonormal bases for all the spaces involved, these properties may be expressed in matrices, yielding exactly the same results obtained with classical real multivariate regression and real normal distributions, as can be found for instance in Fahrmeir and Hamerle [1984].

2.4 Inference on coordinates

2.4.1 Frequentist estimation

Despite the dual definitions of the last chapters, the estimation of the characteristic measures of distributions on $\mathbb{E}$ is done on the coordinate space $\mathbb{R}^C$ [Pawlowsky-Glahn, 2003]. If $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ is a (random, independent) sample from a $\mathcal{N}_{\mathbb{E}}(\underline{\mu}, \underline{\underline{\Sigma}})$, its likelihood will be

$$\begin{aligned} \mathrm{L}\left(\underline{\mu}, \underline{\underline{\Sigma}}; \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\right) &= \Pr\left[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N; \underline{\mu}, \underline{\underline{\Sigma}}\right] = \prod_{n=1}^{N} \Pr\left[\mathbf{z}_n; \underline{\mu}, \underline{\underline{\Sigma}}\right] = \\ &= \left(2\pi|\underline{\underline{\Sigma}}|\right)^{-(N\cdot C)/2} \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(\underline{\zeta}_n - \underline{\mu})^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot (\underline{\zeta}_n - \underline{\mu})\right]. \end{aligned} \tag{2.14}$$

where $\underline{\zeta}_n$ is the vector of coordinates of $\mathbf{z}_n$ with respect to the basis of $\mathbb{E}$ used to express $\underline{\mu}$ and $\underline{\underline{\Sigma}}$. Note that the first equality is only valid when samples are independent. In these conditions, the maximum likelihood is attained at the values

$$\hat{\underline{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\underline{\zeta}_N = \bar{\underline{\zeta}}, \quad \text{and} \quad \hat{\underline{\underline{\Sigma}}} = (\hat{\sigma}_{ij}) \quad \text{where} \quad \hat{\sigma}_{ij} = \frac{1}{N}\sum_{n=1}^{N}\zeta_{ni}\cdot\zeta_{nj} - \bar{\zeta}_i \cdot \bar{\zeta}_j, \tag{2.15}$$

and ζ_{ni} is the i-th coordinate of the n-th observation.

The estimators (2.15) are *maximum likelihood estimators in* $\mathbb{E}$, and satisfy the same properties as the classical maximum likelihood estimators of the parameters of a normal distribution, as summarized in most statistical textbooks, *e.g.*Fahrmeir and Hamerle

[1984]. In particular, the mean is the best linear unbiased estimator of $\mathrm{E}\left[\underline{\zeta}\right]$, the first moment of the real coordinates of $\mathbf{Z}$.

Being functions of a random sample, these estimators are random themselves too. It is useful then to replace the *point estimations*, specially for central tendency measure, by *confidence region estimations* for a given error probability α. Using the fact that

$$Q = \frac{N(N-D)}{D(N-1)}(\bar{\underline{\zeta}} - \underline{\mu})^t \cdot \widehat{\underline{\underline{\Sigma}}}^{-1} \cdot (\hat{\underline{\zeta}} - \underline{\mu}) \sim \mathcal{F}(D, N-D), \tag{2.16}$$

it is possible to define an hyper-ellipsoid centered on the estimator $\bar{\underline{\zeta}}$ which contains the true value of $\underline{\mu}$ with a confidence of $1-\alpha$,

$$\Pr\left[(\bar{\underline{\zeta}} - \underline{\mu})^t \cdot \widehat{\underline{\underline{\Sigma}}}^{-1} \cdot (\hat{\underline{\zeta}} - \underline{\mu}) \leq \frac{D(N-1)}{N(N-D)} \cdot \mathcal{F}_\alpha(D, N-D)\right] = 1-\alpha, \tag{2.17}$$

being $\mathcal{F}(D, N-D)$ Fisher's $\mathcal{F}$ distribution with D and $N-D$ degrees of freedom, and $\mathcal{F}_\alpha(D, N-D)$ its upper tail α quantile. These concepts and equalities are extracted from Fahrmeir and Hamerle [1984].

Finally, if we are interested on point or confidence-region estimations of a central tendency characteristic element, we can apply the results of equations (2.15) and (2.17) to the basis of $\mathbb{E}$. The obtained element is the best linear unbiased estimator of the first moment of $\mathbf{Z}$ regarding the geometry of $\mathbb{E}$ [Pawlowsky-Glahn and Egozcue, 2001, Pawlowsky-Glahn, 2003].

2.4.2 Bayesian estimation

If we assume that the parameters themselves $\underline{\mu}$ and $\underline{\underline{\Sigma}}$ are also random, as the estimations resulting from (2.15), there is another way to estimate them from a sample $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N \sim \mathcal{N}_{\mathbb{E}}(\underline{\mu}, \underline{\underline{\Sigma}})$. This uses Bayes' Theorem. This is again a classical issue, and many textbooks may be found which treat it. We loosely follow the exposition by Leonard and Hsu [1999]. The next section presents examples of such an estimation.

The first step in the Bayesian estimation procedure is the collection and encoding of all the available *prior information* on the sought parameters. This encoding assigns a prior distribution to these parameters, denoted by $\Pr\left[\underline{\mu}, \underline{\underline{\Sigma}}\right]$. In this way, the analyst explains which of their values are more or less likely in her opinion.

The second step is the computation of the likelihood of the sample using (2.14). The final step is the combination of both through Bayes' Theorem,

$$\Pr\left[\underline{\mu}, \underline{\underline{\Sigma}} | \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\right] \propto \Pr\left[\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N | \underline{\mu}, \underline{\underline{\Sigma}}\right] \cdot \Pr\left[\underline{\mu}, \underline{\underline{\Sigma}}\right], \tag{2.18}$$

to obtain the so-called *posterior distribution* of $(\underline{\mu}, \underline{\underline{\Sigma}})$. Expression (2.18) integrates the information and uncertainties on the knowledge of the values of the parameters *before* looking at the sample, and the information drawn *from* the sample. If instead of the whole distribution we look for a single value of the parameters, either the mode, the mean or the median (in single-parameter cases) of their joint posterior distribution

could be chosen, and it is equally possible to define a suitable posterior confidence region.

However, in this work interest lies in the computation of probabilities that $\mathbf{Z}$ lies in certain hazardous regions. In this situation, we want to take into account all the information available, even the information on *how uncertain* is the obtained prediction. Given a region $A \subset \mathbb{E}$ for each value of $(\underline{\mu}, \underline{\underline{\Sigma}})$ we can compute a hazard probability

$$p(A) = \Pr\left[\mathbf{Z} \in A | \underline{\mu}, \underline{\underline{\Sigma}}\right] = \int_A f_{\mathbf{Z}}(\mathbf{z}; \underline{\mu}, \underline{\underline{\Sigma}}) d\mathbf{z},$$

which will appear in a proportion $\Pr\left[\underline{\mu}, \underline{\underline{\Sigma}} | \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\right]$. Since these proportions sum up to one, they represent the probability density of $p(A)$. Again, we can characterize this density either by confidence intervals, quantiles, or through its mean value,

$$\bar{p}(A) = \int \Pr\left[\mathbf{Z} \in A | \underline{\mu}, \underline{\underline{\Sigma}}\right] \cdot \Pr\left[\underline{\mu}, \underline{\underline{\Sigma}} | \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\right] d(\underline{\mu}, \underline{\underline{\Sigma}}). \tag{2.19}$$

Expression (2.19) gives the *predictive distribution* of $\mathbf{Z}$. Pay attention to the fact that here $\mathbf{Z}$ is a *random* vector and $\mathbf{z}_i$ are *known vectors*. Finally, the predictive density can be readily obtained using the Radon-Nykodym derivative (eq. 2.4), although it is seldom used in this work.

2.5 Case studies

2.5.1 Water conductivity as a real random variable

Before heading to more complicated (and interesting) cases, let us consider the conductivity variable of the Gualba data set. Let $\mathbf{Z}$ be the random variable *measured conductivity*. Conductivity is defined as the ability of a medium (water) to allow the flow of electrons. In this case, it is measured in $\mu S/cm$. Water conductivity is the additive result of the contributions of each one of the present ions, weighted in function of their electric charge. Thus, it seems reasonable to consider that differences among samples of conductivity should be measured using substraction. This points out to the use of an additive scale for conductivity, although its sample space is the set of positive real numbers (denoted by $\mathbb{R}_+$).

The real line with an additive scale structure (denoted by $\{\mathbb{R}, +, \cdot\}$) is the classical Euclidean space. Given $\mathbf{a} = (a), \mathbf{b} = (b) \in \mathbb{E} = \mathbb{R}$ and $\lambda \in \mathbb{R}$, its main operations are defined as the classical ones:

inner sum: sum of components, $\mathbf{a} \oplus \mathbf{b} = (a + b)$

external product: product of the components by the scalar, $\lambda \odot \mathbf{a} = (\lambda \cdot a)$

distance: absolute difference of the components,

$$d(\mathbf{a}, \mathbf{b}) = |a - b| \tag{2.20}$$

scalar product: direct product of the components, $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbb{E}} = a \cdot b$.

With these operations, the following vectors of $\mathbb{R}$ deserve to be mentioned:

neutral element: $\mathbf{n} = (0)$

basis: either $\mathbf{e} = (1)$ or $\bar{\mathbf{e}} = (-1)$ can be considered as (ortho)normal basis of $\mathbb{R}$ as an Euclidean space; if we take the first one, the coordinate of a vector $\mathbf{a} \in \mathbb{R}$ in this basis will be

$$\alpha = \langle \mathbf{a}, \mathbf{e} \rangle_{\mathbb{E}} = a.$$

Thus, if $\mathbb{E} = \mathbb{R}$ we identify as the same thing the *vector* $\mathbf{a}$, its *value* (a), and its *coordinate* α. But this will not be the case in other spaces.

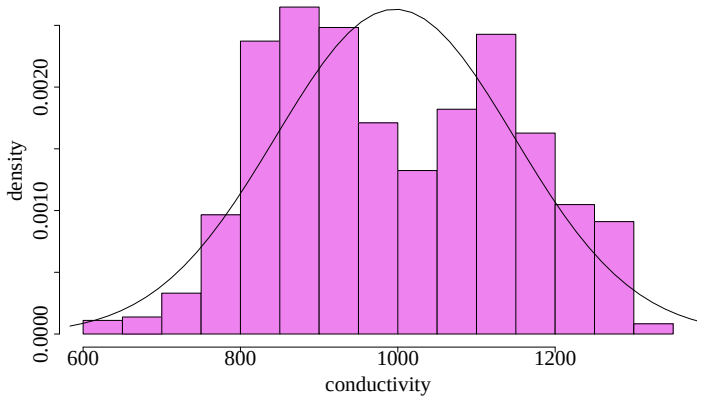

Figure 2.1: Histogram of conductivity data set against the normal distribution, with maximum likelihood estimates of the parameters. This data set was taken at the Gualba station, during July 2002, and it contains 725 samples. Note the clear bi-modality of the data set, which we will ignore in this step. The analysis of this data set is completed in section 3.6.

The normal distribution on $\mathbb{R}$ can be found in all textbooks on statistics, since it is the classical normal distribution. Figure 2.1 shows the shape of this distribution compared with the histogram of the conductivity data set. Note that the histogram classes have the same length according to (2.20). Note also that the fit of the model is unacceptable, since it does not capture the clear bi-modality of the data. However, at this step of the analysis, we use this data set only for illustration. A further discussion will follow in future chapters (section 3.6 is devoted to its analysis). The estimated parameters of this normal distribution are $\underline{\mu} = (\tilde{\mu})$ and $\underline{\underline{\Sigma}} = (\tilde{\sigma}^2)$ are the maximum likelihood estimates of the sample,

$$\tilde{\mu} = 996.309 \mu S/cm \quad \text{and} \quad \tilde{\sigma}^2 = 22992.3 \, (\mu S/cm)^2 .$$

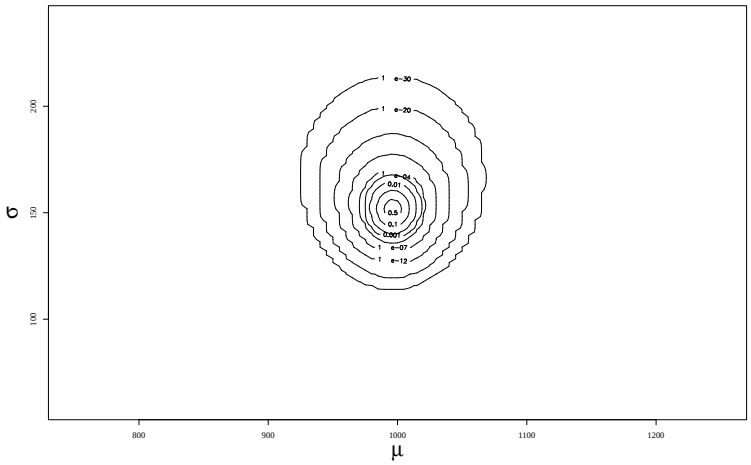

Figure 2.2: Joint posterior distribution of μ and σ parameters for conductivity of the Tordera river at Gualba, measured during July 2002. Some isodensity levels—from 10^{-30} to 0.5—are shown. The only information provided by the prior distribution are the limits of this map. The shape is inherited from the likelihood of the sample.

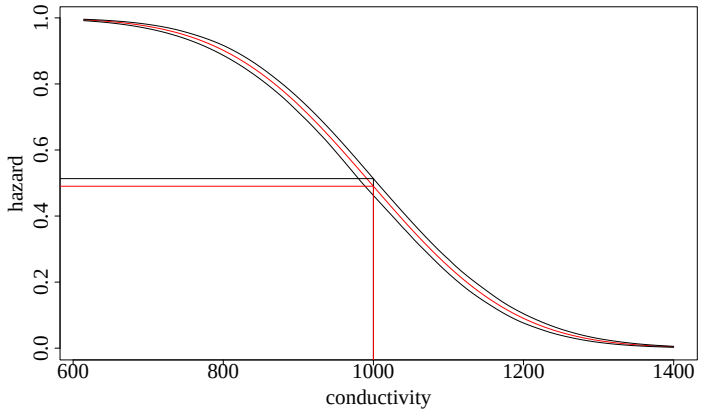

Figure 2.3: Complementary of the predictive distribution of conductivity at Gualba for the month of July 2002, jointly with the 5% and 95% quantiles of the distribution of hazard of exceedance. For a certain threshold, the predictive probability and confidence intervals for it can be read from the plot.

A confidence interval for the mean as a characteristic measure can be easily computed attending to the fact that for $D = 1$ and $N = 725$, the distribution of $\sqrt{Q} \sim \mathtt{t}(N-1)$ in (2.16) is a Student's $\mathtt{t}$, which implies that expression (2.17) is simplified to

$$\Pr\left[\left|\frac{\tilde{\mu}-\mu}{\tilde{\sigma}}\right| \leq \mathtt{t}_{0.975}(N-1)\right] = 0.95,$$

and, provided that $\mathtt{t}_{0.975}(N-1) \approx \mathcal{N}(0,1) = 1.96$, we can be 95% confident that the true mean μ will satisfy

$$\mu \in (\tilde{\mu} - 1.96 \cdot \tilde{\sigma}, \tilde{\mu} + 1.96 \cdot \tilde{\sigma}) = (699.11, 1293.51)\, \mu S/cm.$$

To try a Bayesian approach, we must first define a prior joint distribution for both μ and σ. After considering the characteristics of the basin, essentially the small proportion of carbonate landscape and the low-medium human presence, we believe that the mean conductivity will be reasonably below $1250 \mu S/cm$, and above $750 \mu S/cm$. Without more information, the central quarter of this range seems to be a good range of variation for μ. If we consider the mean to be in the center of the guessed range of variation, and this range to be roughly equivalent to a 95% interval for conductivity, this would imply an approximate σ equal to the range divided by four, thus $125 \mu S/cm$. A possible range of variation of σ could be to take the half and the double of this value. Lacking more information, the prior distribution is a flat distribution between these values.

Updating this distribution by the likelihood of the sample of conductivity through (2.18), we obtain the joint posterior distribution of μ and σ represented in figure 2.2.

However, we finally want to compute the hazard that $\mathbf{Z} > z$, at each interesting threshold z. Table 1.1 summarizes the thresholds of the ACA (Catalan Water Control Agency), to determine to which uses a water mass may be devoted, as a function of its conductivity. Figure 2.3 shows two quantiles of hazard of exceeding 1000 $\mu S/cm$, and the predictive distribution (2.19) of conductivity, which tells that one can be 95% confident that the hazard of exceeding the level $1000 \mu S/cm$ is below a probability of 0.51, with a predictive probability of 0.49.

In conclusion, if we would like to assess the water quality of the Tordera river at Gualba station according to the hazard of exceeding the threshold $1000 \mu S/cm$, we would say that this hazard is at most 0.51 probable. Whether this is finally an acceptable level or not will obviously depend on the water management policy and the expected water use.

2.5.2 Ammonia system as a positive random vector

In the Gualba data set we have the information to characterize the ammonia chemical system of this river. Ammonia (NH_3) is not directly measurable, but it can be computed by using the equilibrium constant equation (1.3). The content in the phases of the system plus the equilibrium constant K_a can be regarded as a random vector $\mathbf{Z} = \left(H_3O^+, K_a, NH_4^+, NH_3\right)$ with positive components. Attending to the fact that acidity is

always accounted for through pH instead of by direct H_3O^+ content, we will consider that the sample space for $\mathbf{Z}$, the positive orthant of $\mathbb{R}^4$ (writen as $\mathbb{R}^4_+$), may be given a relative scale where comparison between vectors should be done on their component logarithms.

To build an Euclidean space structure on $\mathbb{R}^4_+$ describing this logarithmic scale, we need the following operations

inner sum: product of vectors component-wise, $\mathbf{a}\oplus\mathbf{b} = (a_1 \cdot b_1, a_2 \cdot b_2, a_3 \cdot b_3, a_4 \cdot b_4)$

external product: component-wise power of the vector by the scalar,

$$\lambda\odot\mathbf{a} = (a_1^\lambda, a_2^\lambda, a_3^\lambda, a_4^\lambda),$$

distance: square root of the squared difference among logarithms of components,

$$d(\mathbf{a},\mathbf{b}) = \sqrt{\log_{10}^2 \frac{a_1}{b_1} + \log_{10}^2 \frac{a_2}{b_2} + \log_{10}^2 \frac{a_3}{b_3} + \log_{10}^2 \frac{a_4}{b_4}} \tag{2.21}$$

scalar product: sum of component-wise product of logarithms,

$$\langle \mathbf{a},\mathbf{b}\rangle_{\mathbb{E}} = \log_{10} a_1 \log_{10} b_1 + \log_{10} a_2 \log_{10} b_2 + \log_{10} a_3 \log_{10} b_3 + \log_{10} a_4 \log_{10} b_4.$$

where $\mathbf{a} = (a_1, a_2, a_3, a_4), \mathbf{b} = (b_1, b_2, b_3, b_4) \in \mathbb{R}^4_+$ and $\lambda \in \mathbb{R}$. With these operations, the vectors of $\mathbb{R}^4_+$ with a special meaning are:

neutral element: $\mathbf{n} = (1, 1, 1, 1)$

basis: a set of four vectors, with 1 in all components, except by 1/10 in one, like $\mathbf{e}_2 = (1, 1/10, 1, 1)$. There are four vectors of this kind in this orthonormal basis, and the i-th coordinate of any vector $\mathbf{a} \in \mathbb{R}^4_+$ will be

$$\alpha_i = \langle \mathbf{a}, \mathbf{e}\rangle_{\mathbb{E}} = \log_{10} 0.1 \cdot \log_{10} a_i = -1 \cdot \log_{10} a_i = -\log_{10} a_i,$$

which corresponds to the *chemical potential* of the i-th component. Thus, the real vector of potentials $\underline{\zeta} = (p\mathrm{H}, pKa, p\mathrm{NH_4}, p\mathrm{NH_3})$ is exactly the vector of coordinates of $\mathbf{Z} = \left(\mathrm{H_3O^+}, \mathrm{K}_a, \mathrm{NH_4^+}, \mathrm{NH_3}\right)$. From this point of view, the equilibrium equation of Ammonia (1.3) define a 3-dimensional vector subspace of $\mathbb{R}^4_+$ or, alternatively, a 1-dimensional space, that of ammonia content, a random variable $\mathbf{Z}_4 \in \mathbb{R}_+$ as $\mathbf{Z}_4 = [\mathrm{NH_3}]$.

The normal distribution on $\mathbb{R}_+$ was defined by Mateu-Figueras et al. [2002]. It is compared in figure 2.4 with the histogram of the computed $\mathbf{Z}_4$ values. Note that the classes of this histogram have the same length according to the distance in $\mathbb{R}_+$ (2.21). We assume the coordinate ζ_4 to be normally distributed in $\mathbb{R}$, so that we can apply the same procedures of section 2.5.1 to estimate its parameters. These estimates are summarized in table 2.1. If interest lies on computing a mean value of ammonia

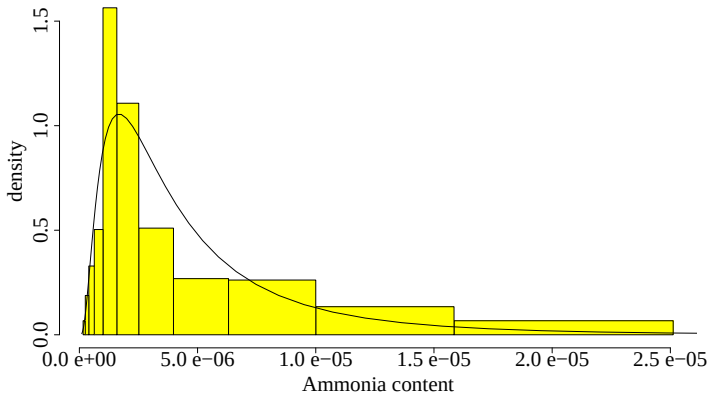

Figure 2.4: Histogram of ammonia content (ζ_4), compared with a normal on $\mathbb{R}_+$ distribution using maximum likelihood estimates for the parameters. This data set was taken at the Gualba station, during July 2002, and it contains 745 samples. Note that the horizontal scale is represented in molarity units

variable	model	mean	95% confidence interval
ζ_4	Normal on $\mathbb{R}$	5.77	(5.035 , 6.526)
$\mathbf{Z}_4$	Normal on $\mathbb{R}_+$	$16.75 \cdot 10^{-7}$	$(3.04, 92.17) \cdot 10^{-7}$
Z_4	Lognormal (on $\mathbb{R}$)	$19.74 \cdot 10^{-7}$	-

Table 2.1: Point and interval estimates of the mean of ammonia coordinate ζ_4, as well as ammonia mean value, considered as a normal variable on $\mathbb{R}_+$, and as a lognormal variable on $\mathbb{R}$. Note that, when taken as a vector of an Euclidean space, $\mathbf{Z}$ is written using a boldface latin character, whereas we use a normal character when it is considered as a real value. The data set was taken at the Gualba station, during July 2002, and it contains 745 samples.

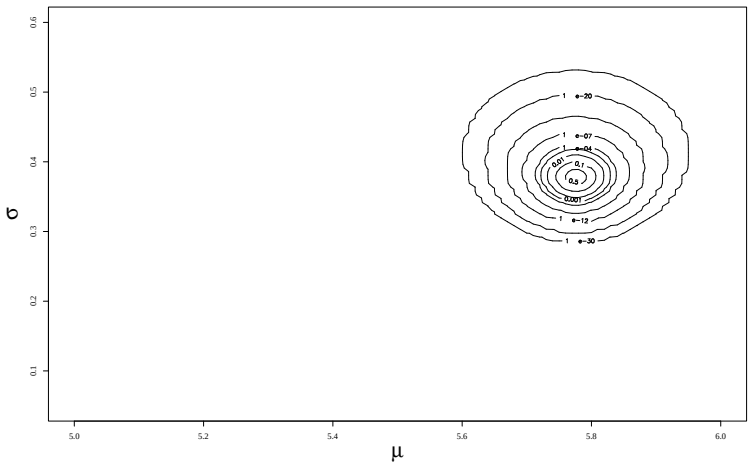

Figure 2.5: Joint posterior distribution of μ and σ parameters for ammonia content in the Tordera river at Gualba during July 2002. Some isodensity levels—from 10^{-30} to 0.5—are shown. The only information provided by the prior distribution are the limits of this map. The shape is provided by the likelihood of the sample.

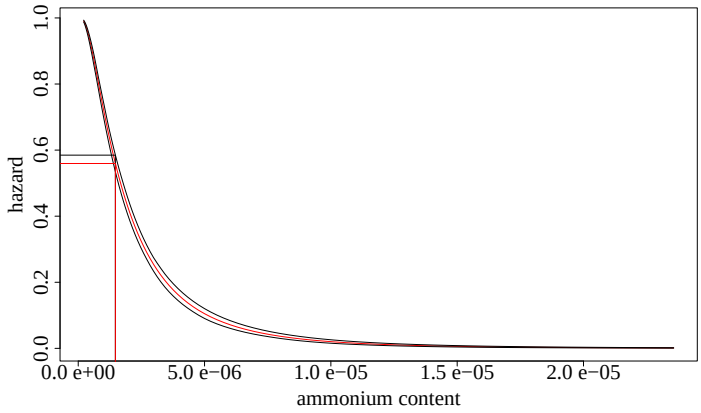

Figure 2.6: Complementary of the predictive distribution of ammonia content in the Tordera river at Gualba during July 2002, jointly with the 5% and 95% quantiles of the distribution of hazard of exceedance. For a certain threshold, the predictive probability and confidence intervals for it can be read from the plot. Note that the horizontal scale is expressed in molarities.

content, then we must apply the central tendency estimate to the basis of the sub-space of $\mathbf{Z}_4$,

$$\mathrm{E}_+ [\mathbf{Z}_4] = \tilde{\mu} \odot \mathbf{e}_4 = \left(\frac{1}{10} \right)^{\tilde{\mu}},$$

which is a valid procedure for both point and interval estimates (table 2.1). The normal distribution in $\mathbb{R}_+$ has exactly the same probability law as the lognormal distribution [McAlister, 1879, Aitchison and Brown, 1957]. However, they are defined on different spaces, respectively $\mathbb{R}_+$ and $\mathbb{R}$, and their densities are different, as well as their expectations. The expectation of Z_4 as a lognormal variable is defined in $\mathbb{R}$ as

$$\mathrm{E} [Z_4] = \left(\frac{1}{10} \right)^{\left(\tilde{\mu} + \frac{1}{2} \tilde{\sigma}^2 \right)}. \tag{2.22}$$

Table 2.1 includes also this lognormal expectation, without confidence intervals, because there exists no clear way to build them, according to Mateu-Figueras et al. [2002]. These authors give also a detailed comparison of these two distributions, their moments and properties.

We followed a Bayesian estimation procedure defined on the coordinate ζ_4 as we did for conductivity in the last section. Again, the prior distribution was considered uniform between a set of limits chosen *a priori*. Updating through (2.18) by the likelihood of the sample of $\mathbf{Z}_4$, ammonia content, we obtained the posterior map (figure 2.5). The shape of this posterior distribution is essentially inherited from the likelihood, while the limits were informed through the prior distribution.

Using this posterior distribution, we computed the hazard associated with each possible threshold value of ammonia content. The distribution (figure 2.6) of this hazard is expressed through some quantiles and the predictive distribution for $\mathbf{Z}_4$ given the observed sample. In this plot, one reads the hazard of exceeding the level of $0.025 mg/l$ (molarity $\sim 1.5 \cdot 10^{-6}$) of ammonia content as being below 0.58 with a 95% of confidence, or with a predictive probability of 0.56.

Therefore, if we want a water quality index to measure the hazard of exceeding the threshold of $0.025 mg/l$ of ammonia content, in the Gualba station we would say that this hazard is slightly below 0.6 probable in the period comprised between July 1, 2002 and July 31, 20002. Again, this will surely be an unacceptable level, but it depends on the water management policy and use.

2.5.3 Moss pollution as a random composition

In the Ukrainian Carpathian Range data set we are interested in the concentration of three heavy metals in moss samples: content in iron (Fe), lead (Pb) and quicksilver (Hg). These heavy metals are either given in parts per million (ppm), proportions (parts per one) or in relative mass percentage (%), which indicates that they are compositional data. The sample space of compositional data is the D-part Simplex, denoted by $\mathbb{S}^D$. A meaningful scale for compositional data takes into account that

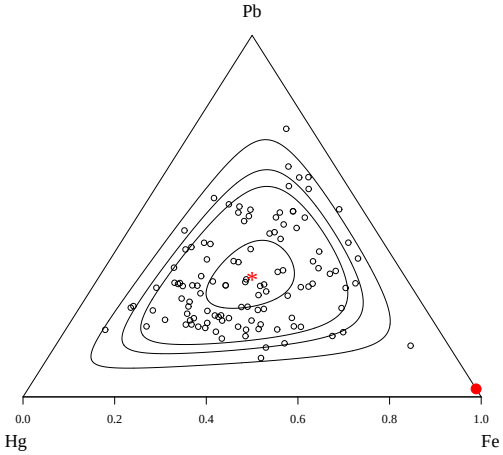

Figure 2.7: Ternary diagram of the composition $(\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg})$, compared with some isodensity levels (corresponding to 50%, 90%, 95% and 99% probability regions) of a normal distribution on $\mathbb{S}^3$ distribution using maximum likelihood estimates for the parameters. The red star indicates the center of the diagram, and the red solid circle the true mean of the data set. The sample (black circles) is perturbed, so that its mean coincides now with the center of the diagram.

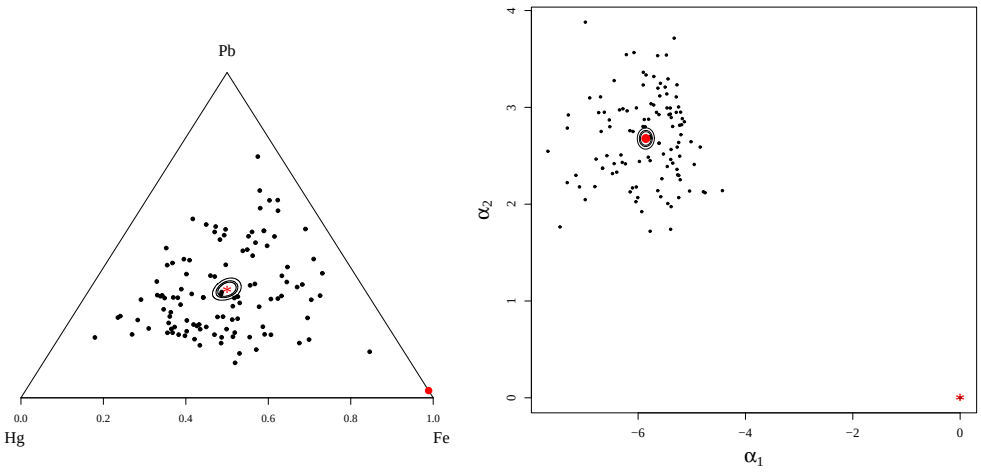

Figure 2.8: Ternary diagram of the composition $(\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg})$ (left) and scatter-plot of their coordinates (right), with 90%, 95% and 99% confidence regions drawn on the center (red star) and the mean value (red solid circle) respectively. Note that the coordinate diagram shows the true position of the sample (black dots), around its mean, whereas in the ternary diagram the sample has been perturbed to the center.

the total sum of a composition is not a relevant information: either the composition has been artificially forced to sum up to a constant (it has been *closed*), or this total amount is conditioned by the sampling procedure and does not inform us about the studied process. A compositional scale considers also information only from relative importance of components.

Let $\mathbf{Z} = (\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg})$ be the random composition indicating the proportion of these three elements in each sample. Its sample space is $\mathbb{E} = \mathbb{S}^3$, and it can be given an Euclidean structure satisfying its relative scale [Billheimer et al., 2001, Pawlowsky-Glahn and Egozcue, 2001] through the following set of operations:

inner sum: *perturbation* [Aitchison, 1982], closed product of vectors component-by-component, $\mathbf{a} \oplus \mathbf{b} = \mathcal{C}\left(a_1 \cdot b_1, a_2 \cdot b_2, a_3 \cdot b_3\right)$

external product: *power operation* [Aitchison, 1986], closed component-wise power of the vector by the scalar, $\lambda \odot \mathbf{a} = \mathcal{C}\left(a_1^\lambda, a_2^\lambda, a_3^\lambda\right)$

distance: proportional to the square root of the squared difference among all possible log-ratios of components [Aitchison, 1982]:

$$d_A(\mathbf{a}, \mathbf{b}) = \sqrt{\frac{1}{D} \sum_{i<j}^{D} \left(\ln \frac{a_i}{a_j} - \ln \frac{b_i}{b_j} \right)^2}, \tag{2.23}$$

with $D = 3$, the number of parts;

scalar product: sum of component-wise product of log-ratios [Aitchison, 1984],

$$\langle \mathbf{a}, \mathbf{b} \rangle_A = \sum_{i=1}^{3} \log \frac{a_i}{\sqrt[3]{a_1 \cdot a_2 \cdot a_3}} \log \frac{b_i}{\sqrt[3]{b_1 \cdot b_2 \cdot b_3}}.$$

where $\mathbf{a} = (a_1, a_2, a_3), \mathbf{b} = (b_1, b_2, b_3) \in \mathbb{S}^3$ are two composition, $\lambda \in \mathbb{R}$ is a real value, and $\mathcal{C}(\cdot)$ represents the closure operation, which divides each part in the composition by their sum in order to obtain proportions. With these operations, the following vectors of $\mathbb{S}^3$ have a special meaning:

neutral element: $\mathbf{n} = \mathcal{C}(1, 1, 1) = (1/3, 1/3, 1/3)$

basis: one possible orthonormal basis is [Egozcue et al., 2003]

$$\mathbf{e}_1 = \mathcal{C}\left(\exp \frac{2}{\sqrt{6}}, \exp \frac{-1}{\sqrt{6}}, \exp \frac{-1}{\sqrt{6}}\right) \quad \mathbf{e}_2 = \mathcal{C}\left(1, \exp \frac{1}{\sqrt{2}}, \exp \frac{-1}{\sqrt{2}}\right)$$

and the corresponding coordinates of any vector $\mathbf{a} \in \mathbb{S}^3$ will be computed using (2.2), which gives the expressions

$$\alpha_1 = \frac{1}{\sqrt{6}} \ln \frac{\mathrm{Fe}^2}{\mathrm{Pb} \cdot \mathrm{Hg}}, \quad \alpha_2 = \frac{1}{\sqrt{2}} \ln \frac{\mathrm{Pb}}{\mathrm{Hg}}. \tag{2.24}$$

Since the dimension of the D-part Simplex is always $D-1$, the basis has only two elements. Consequently, three-part compositions have only two degrees of freedom, and can be meaningfully plotted in 2-D plots, usually called *ternary diagrams*. We assume these two coordinates (α_1, α_2) of the random composition $\mathbf{Z}$ to follow a joint normal distribution, thus $\mathbf{Z}$ follows a normal distribution in the simplex [Mateu-Figueras et al., 2003]. A formal definition is included in section 2.6. Figure 2.7 represents the studied sample, previously centered and with some isodensity levels of the fitted normal distribution on the Simplex. The *centering* operation perturbs the sample to the center of the plot, without altering the spread structure of the data set and enhancing its interpretability [von Eynatten et al., 2002]. The normal distribution appearing in figure 2.7 has as parameters the maximum likelihood estimates of mean and variance matrix for the data set coordinates, which are

$$\underline{\tilde{\mu}} = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{pmatrix} = \begin{pmatrix} -5.8567 \\ 2.6782 \end{pmatrix}, \quad \underline{\underline{\tilde{\Sigma}}} = \begin{pmatrix} \tilde{\sigma}_{11} & \tilde{\sigma}_{12} \\ \tilde{\sigma}_{12} & \tilde{\sigma}_{22} \end{pmatrix} = \begin{pmatrix} 0.4332 & 0.0048 \\ 0.0048 & 0.2067 \end{pmatrix}.$$

If we want to translate this characteristic measures to the Simplex as characteristic elements, we can apply directly the maximum likelihood mean estimate to the vectors of the basis, and obtain the so-called (metric) *center* of the composition [Aitchison, 1982, Pawlowsky-Glahn and Egozcue, 2001],

$$\mathrm{cen}(\mathbf{Z}) = \tilde{\mu}_1 \odot \mathbf{e}_1 \oplus \tilde{\mu}_2 \odot \mathbf{e}_2 = \left(1.13 \cdot 10^{-4}, 0.977, 2.21 \cdot 10^{-2}\right).$$

As was stated in section 2.4, this characteristic element of central tendency is an unbiased estimator of the mean of $\mathbf{Z}$ in the Simplex, $\mathbf{n} = \mathrm{E}_A\left[\mathrm{cen}(\mathbf{Z}) \ominus \mathrm{E}_A\left[\mathbf{Z}\right]\right]$. The same operation can be applied to the confidence regions drawn with help of expression (2.17). Figure 2.8 shows the data set, its center and some confidence regions in both the space of coordinates and the Simplex.

Though the high dimension of the problem, a Bayesian approach would be inherently no more difficult in this case as it was for conductivity or ammonia content in the other case studies. After specifying a joint prior distribution for the five parameters of the system, the updating of the prior by the likelihood of expression (2.14) would be a straightforward task. The only problems would involve computation time and representation of the posterior maps, since the five dimensions cannot be represented at a time. Also, its usefulness in this problem is low, because here we are not interested in computing probabilities of hazardous events.

2.6 Distributions on the Simplex

This section summarizes some useful distributions of random vectors on the Simplex, which are afterwards used in chapters 6 and 7.

Let $\mathbb{S}^D$ be the D-part Simplex, the set of vectors $\mathbf{z} = (z_1, z_2, \ldots, z_D)$ so that $z_i \geq 0$ and $\sum_{i=1}^{D} z_i = 1$. Aitchison et al. [2002] showed that the Simplex, jointly with the operations of perturbation and powering, is a vector space, and Billheimer et al. [2001]

and Pawlowsky-Glahn and Egozcue [2001] that it may be indeed given an Euclidean space. This structure is detailed in section 2.5.3. Taking profit of it, Pawlowsky-Glahn [2003] defines a Lebesgue measure on the Simplex (2.3), which we will denote as $\lambda_{\mathbb{S}^D}$. The relationship between this measure and the classical Lebesgue measure $\lambda_{\mathbb{R}^{D-1}}$ on $\mathbb{R}^{D-1}$ embedding $\mathbb{S}^D$ is given by

$$\frac{d\lambda_{\mathbb{S}}(\mathbf{z})}{d\lambda_{\mathbb{R}^{D-1}}(\mathbf{z})} = \frac{1}{\sqrt{D} z_1 \cdot z_2 \cdots z_D}.$$

The following distributions will be defined with respect to either one of these measures. The other definition follows immediately by application of equation (2.5).

2.6.1 The Dirichlet distribution

Definition 2.22 *Let $\boldsymbol{Z}$ be a random vector on the Simplex. Then it has a Dirichlet distribution with a vector of positive parameters $\underline{\theta} = (\theta_1, \theta_2 \ldots, \theta_D)$ with respect to the classical Lebesgue measure $\lambda_{\mathbb{R}}$ if its density function $f(\boldsymbol{z})$ can be expressed as*

$$\boldsymbol{Z} \sim \mathcal{D}(\underline{\theta}) \Leftrightarrow f(\boldsymbol{z}) = \frac{1}{z_1 \cdot z_2 \cdots z_D} \cdot \frac{\Gamma(\theta_0)}{\prod_{i=1}^{D} \Gamma(\theta_i)} \cdot \prod_{i=1}^{D} z_i^{\theta_i} I\left\{\boldsymbol{z} \in \mathbb{S}^D\right\}, \tag{2.25}$$

where $\Gamma(\theta)$ is the gamma function, and $\theta_0 = \sum_{i=1}^{D} \theta_i$ [Abramovitz and Stegun, 1965].

This distribution is continuous, and completely bounded in the Simplex $\mathbb{S}^D$. In the case $D = 2$, it corresponds to the Beta distribution [Abramovitz and Stegun, 1965].

Property 2.3 *With respect to the classical Lebesgue measure $\lambda_{\mathbb{R}}$, the characteristic descriptors of a Dirichlet variable are:*

1. *mode:* $\boldsymbol{z}_\delta = \mathcal{C}(\theta_1 - 1, \theta_2 - 1, \ldots, \theta_D - 1) = \frac{1}{\theta_0 - D}(\theta_1 - 1, \theta_2 - 1, \ldots, \theta_D - 1)$,
2. *mean:* $\bar{\boldsymbol{z}} = \mathrm{E}[\boldsymbol{Z}] = \mathcal{C}(\theta_1, \theta_2, \ldots, \theta_D) = \frac{1}{\theta_0}(\theta_1, \theta_2, \ldots, \theta_D)$,
3. *variances:* $\sigma_i^2 = \mathrm{Var}[Z_i] = \frac{\theta_i(\theta_0 - \theta_i)}{\theta_0^2(\theta_0+1)}, i = 1, 2, \ldots D$,
4. *covariances:* $\sigma_{ij} = \mathrm{Cov}[Z_i, Z_j] = -\frac{\theta_i \theta_j}{\theta_0^2(\theta_0+1)}, i, j = 1, 2, \ldots D$,
5. *covariance matrix:* $\underline{\underline{\Sigma}} = \mathrm{Var}[\boldsymbol{Z}] = \frac{1}{\theta_0+1}\left(\mathrm{diag}[\bar{\boldsymbol{z}}] - \bar{\boldsymbol{z}} \cdot \bar{\boldsymbol{z}}^t\right)$

with $\mathcal{C}(\cdot)$ denoting the closure operation, and considering $\bar{\boldsymbol{z}}$ as a column-vector.

Actually, the density (2.25) is obtained as the closure of a vector of random variables following independent equally-scaled gamma distributions. The correlation structure exhibited by the parts of a vector $\mathbf{Z} \sim \mathcal{D}(\underline{\theta})$ is particularly weak due to this parental independence, and it is completely inherited from the closure operation. This same

closure is also responsible for the fact that the covariance matrix $\underline{\underline{\Sigma}}$ of $\mathbf{Z}$ (in property 2.3.5) is singular.

The Dirichlet distribution has some interesting properties of pseudo-marginalization with respect to the parts. To summarize them we need the following definitions, due to Aitchison [1986].

Definition 2.23 *A selection matrix $\underline{\underline{S}}$ is a $(C \times D)$ matrix $(1 \leq C \leq D)$ with C elements equal to one, exactly one of them in each row and not more than one in each column. The remaining $C(D-1)$ elements are zero.*

Definition 2.24 *An amalgamation matrix $\underline{\underline{A}}$ is a $(C \times D)$ matrix $(1 \leq C \leq D)$ with D elements equal to one, exactly one of them in each column and not less than one in each row. The remaining $D(C-1)$ elements are zero.*

These properties are stated in the form of propositions. Formal proofs can be found in Haas and Formery [2002], jointly with other properties.

Proposition 2.4 *If $\boldsymbol{Z} \sim \mathcal{D}(\underline{\theta})$, for any selection matrix $\underline{\underline{S}}$, the vector $\mathcal{C}\left(\underline{\underline{S}} \cdot \boldsymbol{Z}\right)$ has as its sample space the C-dimensional Simplex, and as distribution a Dirichlet distribution on $\mathbb{S}^C$ with a vector of parameters $\underline{\underline{S}} \cdot \underline{\theta}$.*

Proposition 2.5 *If $\boldsymbol{Z} \sim \mathcal{D}(\underline{\theta})$, for any amalgamation matrix $\underline{\underline{A}}$ with two rows, the vector $\mathcal{C}\left(\underline{\underline{A}} \cdot \boldsymbol{Z}\right)$ has as its sample space the 2-dimensional Simplex, and as distribution a Dirichlet distribution on $\mathbb{S}^2$ (or a beta distribution) with a vector of parameters $\underline{\underline{A}} \cdot \underline{\theta}$.*

Definition 2.25 *The Dirichlet probability density function with respect to the Lebesgue measure in the Simplex is*

$$f(\boldsymbol{z}) = \sqrt{D}\frac{\Gamma(\theta_0)}{\prod_{i=1}^{D}\Gamma(\theta_i)} \cdot \prod_{i=1}^{D} z_i^{\theta_i} I\left\{\boldsymbol{z} \in \mathbb{S}^D\right\}, \tag{2.26}$$

which corresponds to the application of equation (2.5) to change the measure of representation of a probability density.

As far as we know there is no closed analytical expression for the mean vector and the covariance matrix of the Dirichlet distribution with respect to the geometry of the Simplex.

Property 2.4 *The mode of the distribution is*

$$\boldsymbol{z}_\delta = \mathcal{C}(\theta_1, \theta_2, \ldots, \theta_D) = \frac{1}{\theta_0}(\theta_1, \theta_2, \ldots, \theta_D),$$

and it coincides with the value of the mean in the $\mathbb{R}^D$ geometry.

Proof The equations of the densities (2.26) with parameters θ_i and (2.25) with parameters $\alpha_i = \theta_i + 1$ are proportional. Replacing this second set of parameters in the expression of the mode in property 2.3 yields directly the result. □

2.6.2 The Normal distribution on the Simplex

Definition 2.26 (Normal distribution on the Simplex) *Let $\boldsymbol{Z}$ be a random vector on the Simplex. Then it has a Normal distribution on the Simplex [Mateu-Figueras et al., 2003] with respect to the Lebesgue measure on the Simplex, $\lambda_{\mathbb{S}}$, if its density function $f(\boldsymbol{z})$ can be expressed as*

$$\begin{aligned}\boldsymbol{Z} \sim \mathcal{N}_{\mathbb{S}^D}(\underline{\mu}, \underline{\underline{\Sigma}}) \Leftrightarrow f(\boldsymbol{z}) = & \frac{I\left\{\boldsymbol{z} \in \mathbb{S}^D\right\}}{(2\pi\cdot)^{(D-1)/2}\left\|\underline{\underline{\Sigma}}\right\|^{1/2}} \times \\ & \times \exp\left(-\frac{1}{2}\left(\log\frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right)^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \left(\log\frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right)\right),\end{aligned}$$

with parameters $\underline{\mu} = (\mu_1, \mu_2 \ldots, \mu_{D-1})$ and $\underline{\underline{\Sigma}} = (\sigma_{ij})$, a positive definite symmetric matrix. The vector $\boldsymbol{z}_{-D}$ stands for the vector $\boldsymbol{z}$ without the last component.

Alternatively, this density can also be expressed by

$$f(\mathbf{z}) \propto \prod_{i=1}^{D} e^{\theta_i \log z_i} \cdot \prod_{i=1}^{D-1}\prod_{j=1}^{D-1} e^{\phi_{ij} \log\frac{z_i}{z_D}\log\frac{z_j}{z_D}},$$

with $\underline{\theta}_{-D} = \underline{\underline{\Sigma}}^{-1} \cdot \underline{\mu}$, $\theta_D = -\sum_{j=1}^{D-1}\theta_j$ and $\underline{\underline{\phi}} = -\underline{\underline{\Sigma}}^{-1}/2$. Mateu-Figueras et al. [2003] defined this distribution in terms of the coordinates of $\mathbf{Z}$ with respect to an orthonormal basis of the Simplex, and showed the equivalence between it and definition 2.26. This equivalence is also implied by the specification of the normal distribution using a vector mean and an endomorphism variance, as in definition 2.21.

Property 2.5 *With respect to the Lebesgue measure in the Simplex, the characteristic descriptors of a normally-distributed (on the Simplex) random vector are:*

1. *mode:* $\boldsymbol{z}_\delta = \mathcal{C}(\exp\mu_1, \exp\mu_2, \ldots, \exp\mu_{D-1}, 1) = \mathcal{C}(\exp\underline{\mu}, 1)$
2. *mean:* $\bar{\boldsymbol{z}} = \mathrm{E}_A[\boldsymbol{Z}] = \mathcal{C}(\exp\underline{\mu}, 1)$
3. *covariance matrix:* $\mathrm{Var}_A[\boldsymbol{Z}] = \underline{\underline{\Sigma}}$.

A proof follows immediately from the properties of the normal distribution and the definitions of mean and variance of section 2.3.2. It can be found also in Mateu-Figueras et al. [2003].

Definition 2.27 *The additive-logistic-Normal distribution function (with respect to the classical Lebesgue measure) is*

$$\begin{aligned}\boldsymbol{Z} \sim \mathcal{ALN}(\underline{\mu}, \underline{\underline{\Sigma}}) \Leftrightarrow f(\boldsymbol{z}) = & \frac{I\left\{\boldsymbol{z} \in \mathbb{S}^D\right\}}{\prod_{i=1}^{D} z_i \left(2\pi \cdot \left\|\underline{\underline{\Sigma}}\right\|\right)^{(D-1)/2}} \times \\ & \times \exp\left(-\frac{1}{2}\left(\log\frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right)^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \left(\log\frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right)\right),\end{aligned}$$

which corresponds to the application of equation (2.5) to change the measure of representation of a normal on the Simplex density (definition 2.26).

Proposition 2.6 *If $\boldsymbol{Z} \sim \mathcal{N}_{\mathbb{S}^D}\left(\underline{\mu}, \underline{\underline{\Sigma}}\right)$, and $\underline{\underline{S}}$ is a selection matrix taking at least the last element, the vector $\mathcal{C}\left(\underline{\underline{S}}^* \cdot \boldsymbol{Z}\right)$ has as its sample space the C-dimensional Simplex, and as distribution a Normal distribution on $\mathbb{S}^C$ with a vector of means $\underline{\underline{S}}^* \cdot \underline{\mu}$, and a variance matrix $\underline{\underline{S}}^* \cdot \underline{\underline{\Sigma}}^* \cdot \underline{\underline{S}}^{*t}$, where $\underline{\underline{S}}^*$ is $\underline{\underline{S}}$ without the last row and column. An equivalent property holds for additive-logistic-normal distributions.*

2.6.3 The $\mathcal{A}$ distribution

Definition 2.28 (Aitchison's $\mathcal{A}$ distribution) *Let $\boldsymbol{Z}$ be a random vector on the Simplex. Then it has an Aitchison's $\mathcal{A}$ distribution with respect to the Lebesgue measure on the Simplex, $\lambda_{\mathbb{S}}$, if its density function $f(\boldsymbol{z})$ can be expressed as*

$$\boldsymbol{Z} \sim \mathcal{A}(\underline{\theta}, \underline{\underline{\phi}}) \Leftrightarrow \log f(\boldsymbol{z}) = \kappa(\underline{\theta}, \underline{\underline{\phi}}) + \sum_{i=1}^{D} \theta_i \log z_i + \sum_{i=1}^{D-1}\sum_{j=1}^{D-1} \phi_{ij} \log \frac{z_i}{z_D} \log \frac{z_j}{z_D}, \quad (2.27)$$

with parameters a vector $\underline{\theta}$ and a matrix $\underline{\underline{\phi}}$. It is useful also to call $\theta_0 = \sum_{i=1}^{D} \theta_i$. Note that $\kappa(\underline{\theta}, \underline{\underline{\phi}})$ is here an accessory function, closing the density to integrate to one. The conditions on the parameters to obtain a proper distribution are either:

1. *the symmetric negative definite character of $\underline{\underline{\phi}}$ (thus its invertibility) and $\theta_0 \geq 0$,*
2. *the non-positive-definite character of $\underline{\underline{\phi}}$, and $\theta_i > 0$ for all i.*

Property 2.6 (Density decomposition and quasi-moments) *If the matrix $\underline{\underline{\phi}}$ is invertible, the $\mathcal{A}$ distribution can be parametrized using θ_0, a matrix $\underline{\underline{\Sigma}} = -2\underline{\underline{\phi}}^{-1}$, and a vector $\underline{\mu} = \underline{\underline{\Sigma}} \cdot \underline{\theta}^*$, with $\underline{\theta}^* = \underline{\theta}_{-D} - \theta_0/D$, which gives*

$$\log f(\boldsymbol{z}) = \kappa(\theta_0, \underline{\mu}, \underline{\underline{\Sigma}}) + \frac{\theta_0}{D}\sum_{i=1}^{D} \log z_i - \frac{1}{2}\left(\log \frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right)^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \left(\log \frac{\boldsymbol{z}_{-D}}{z_D} - \underline{\mu}\right). \quad (2.28)$$

Under the first set of conditions for properness of the $\mathcal{A}$ distribution, its density is proportional to a normal in the Simplex density multiplied by a symmetric Dirichlet density. When $\theta_0 = 0$, the Dirichlet contribution is a uniform, and the Aitchison distribution becomes the Normal distribution on the Simplex. Thus the parameters $\underline{\mu}$ and $\underline{\underline{\Sigma}}$ give approximations to the measures of central tendency and dispersion of the distribution for small θ_0, and they are exactly these parameters when $\theta_0 = 0$, in accordance with property 2.5.

Proof: Call $\underline{\zeta} = \log \frac{\mathbf{z}_{-D}}{z_D} = \left(\log \frac{z_i}{z_D}\right)$ to simplify the expressions. To proof the identity of expressions (2.27) and (2.28), it is enough to take $\underline{\underline{\phi}} = -\underline{\underline{\Sigma}}/2$, and develop the products by components:

$$\begin{aligned}
\log f(\mathbf{z}) &= \kappa(\theta_0, \underline{\mu}, \underline{\underline{\Sigma}}) + \frac{\theta_0}{D}\sum_{i=1}^{D} \log z_i + \left(\log \frac{\mathbf{z}_{-D}}{z_D} - \underline{\mu}\right)^t \cdot \underline{\underline{\phi}} \cdot \left(\log \frac{\mathbf{z}_{-D}}{z_D} - \underline{\mu}\right) = \\
&= \kappa + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \underline{\zeta}^t \cdot \underline{\underline{\phi}} \cdot \underline{\zeta} + \underline{\mu}^t \cdot \underline{\underline{\phi}} \cdot \underline{\mu} - \left(\underline{\mu}^t \cdot \underline{\underline{\phi}} \cdot \underline{\zeta} + \underline{\zeta}^t \cdot \underline{\underline{\phi}} \cdot \underline{\mu}\right) = \\
&= (\kappa + \underline{\mu}^t \cdot \underline{\underline{\phi}} \cdot \underline{\mu}) + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \underline{\zeta}^t \cdot \underline{\underline{\phi}} \cdot \underline{\zeta} - \left(\frac{-1}{2}\underline{\theta}^{*t} \cdot \underline{\zeta} + \underline{\zeta}^t \cdot \underline{\theta}^* \cdot \frac{-1}{2}\right) = \\
&= \kappa^* + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \underline{\zeta}^t \cdot \underline{\underline{\phi}} \cdot \underline{\zeta} + \underline{\zeta}^t \cdot \underline{\theta}^* = \\
&= \kappa^* + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \sum_{i=1}^{D-1}\sum_{j=1}^{D-1} \zeta_i \phi_{ij} \zeta_j + \sum_{i=1}^{D-1} \zeta_i \theta_i^* = \\
&= \kappa^* + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \sum_{i=1}^{D-1} \theta_i^* \log \frac{z_i}{z_D} + \sum_{i=1}^{D-1}\sum_{j=1}^{D-1} \log \frac{z_i}{z_D} \phi_{ij} \log \frac{z_j}{z_D} = \\
&= \kappa^* + \sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i + \sum_{i=1}^{D} \theta_i^* \log z_i + \sum_{i=1}^{D-1}\sum_{j=1}^{D-1} \log \frac{z_i}{z_D} \phi_{ij} \log \frac{z_j}{z_D},
\end{aligned}$$

using the symmetry of $\underline{\underline{\phi}}$, and the equivalence $\underline{\underline{\phi}} \cdot \underline{\mu} = -\underline{\underline{\Sigma}}/2 \cdot \underline{\mu} = -\underline{\theta}^*/2$ introduced in definition 2.26. Note that $\theta_D^* = -\sum_{i=1}^{D-1} \theta_i^*$. The sought identity directly results by taking $\theta_i = \theta_i^* + \theta_0/D$.

Taking exponentials of the vector (2.28), we obtain

$$f(\mathbf{z}) \propto \exp\left(\sum_{i=1}^{D} \frac{\theta_0}{D} \log z_i\right) \cdot \exp\left(-\frac{1}{2}\left(\log \frac{\mathbf{z}_{-D}}{z_D} - \underline{\mu}\right)^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \left(\log \frac{\mathbf{z}_{-D}}{z_D} - \underline{\mu}\right)\right),$$

clearly showing that the density of the $\mathcal{A}$ distribution (under the assumptions of definition 2.28.1) is proportional to the product of a normal in the Simplex distribution, and a symmetric Dirichlet distribution. This property will have its importance when simulating samples from an $\mathcal{A}$ distribution. □

Proposition 2.7 (Posterior density) *The updating of a normal distribution on the Simplex by a multinomial likelihood delivers a posterior distribution following an Aitchison's $\mathcal{A}$ distribution with respect to a Lebesgue measure on the Simplex. In general, the Aitchison's $\mathcal{A}$ distribution is a conjugate prior of the multinomial distribution.*

Proof: This proposition follows immediately from definitions 2.26 (normal distribution on the Simplex) and 2.28 ($\mathcal{A}$ distribution on the Simplex), the fact that a multinomial likelihood is proportional to a Dirichlet distribution, and property 2.6. □

Definition 2.29 *The $\mathcal{A}$ density function [Aitchison, 1986] with respect to the classical Lebesgue measure is*

$$\log f(\boldsymbol{z}) = \kappa(\underline{\theta}, \underline{\underline{\phi}}) + \sum_{i=1}^{D}(\theta_i - 1)\log z_i + \sum_{i=1}^{D-1}\sum_{j=1}^{D-1}\phi_{ij}\log\frac{z_i}{z_D}\log\frac{z_j}{z_D},$$

which corresponds to the application of equation (2.5) to change the measure of representation of an $\mathcal{A}$ density (definition 2.28).

This is the original definition given by Aitchison [1986], with a slightly different parametrization of ϕ.

Property 2.7 (Maximum) *The $\mathcal{A}$ distribution has its mode at the value $\boldsymbol{z}$ satisfying the non-linear system of equations*

$$\underline{0} = \underline{\theta}_{-D} - \theta_0 \cdot \boldsymbol{z}_{-D} + 2\underline{\underline{\phi}} \cdot \log\frac{\boldsymbol{z}_{-D}}{z_D},$$

under the Lebesgue measure on the Simplex.

Proof: Using $\underline{\zeta} = \log\frac{\mathbf{z}_{-D}}{z_D}$, we want to maximize expression (2.27) with respect to $\underline{\zeta}$. This is achieved by taking the first derivative and equating it to zero:

$$\underline{0} = \frac{d\log f(\mathbf{z})}{d\underline{\zeta}} = \frac{d\log\mathbf{z}}{d\underline{\zeta}} \cdot \underline{\theta} + 2\underline{\underline{\phi}} \cdot \underline{\zeta},$$

where

$$\frac{d\log z_i}{d\zeta_j} = \frac{1}{z_i}\frac{dz_i}{d\zeta_j} = \left\{ \begin{array}{ll} 1 - z_j & i = j; \\ -z_j, & i \neq j. \end{array} \right\} = \delta_{ij} - z_j,$$

given that

$$\frac{dz_i}{d\zeta_j} = \frac{d}{d\zeta_j}\frac{e^{\zeta_i}}{1 + \sum_{k=1}^{D-1} e^{\zeta_k}} = \left\{ \begin{array}{ll} \frac{e^{\zeta_j}\left(1+\sum_{k=1}^{D-1} e^{\zeta_k}\right) - e^{\zeta_j}e^{\zeta_j}}{\left(1+\sum_{k=1}^{D-1} e^{\zeta_k}\right)^2} = z_j - z_j^2, & i = j; \\ -\frac{e^{\zeta_i}e^{\zeta_j}}{\left(1+\sum_{k=1}^{D-1} e^{\zeta_k}\right)^2} = -z_i z_j, & i \neq j. \end{array} \right. \tag{2.29}$$

Note that the particular case $i = D$ is also included, by considering $\zeta_D = 0$. Then, for the j-th equation we need to know

$$\frac{d\log\mathbf{z}}{d\zeta_j} \cdot \underline{\theta} = \sum_{i=1}^{D}\frac{d\log z_i}{d\zeta_j}\theta_i = \sum_{i=1}^{D}(\delta_{ij} - z_j)\theta_i = \theta_j - z_j\sum_{i=1}^{D}\theta_i = \theta_j - z_j\theta_0$$

which yields for each equation

$$0 = \theta_j - z_j\theta_0 + 2\sum_{i=1}^{D-1}\phi_{ji}\log\frac{z_i}{z_D},$$

thus obtaining the desired expression. To be sure that it is a maximum, we take a second derivative, write expression (2.29) in matrix form, and obtain

$$\frac{d}{d\underline{\zeta}}\left(\underline{\theta}_{-D} - \theta_0\cdot\mathbf{z}_{-D} + 2\underline{\underline{\phi}}\cdot\underline{\zeta}\right) = 2\underline{\underline{\phi}} - \theta_0\cdot\left(\mathrm{diag}\left[\mathbf{z}_{-D}\right] - \mathbf{z}_{-D}\cdot\mathbf{z}_{-D}^t\right).$$

This is a non-positive definite matrix ($\underline{\underline{\phi}}$) *minus* a positive definite matrix (being a full-rank minor of the variance of a Dirichlet-distributed variable, in accordance with property 2.3.5), and consequently the function is convex everywhere and has a maximum. □

2.7 Remarks

Up to here, we have been considering the effect of choosing an Euclidean structure to describe the scale and the sample space of a random vector on some classical statistical issues. We have shown through three examples that many sample spaces can be meaningfully structured as Euclidean spaces. This allows us to express the data as coordinates in a given basis, and work on these coordinates as real numbers, using a classical approach on them. This is what we could call the *principle of working on coordinates* [Pawlowsky-Glahn, 2003].

After realizing the importance of taking into account a meaningful geometry of the sample space, we sought a way to monitor the uncertainty affecting the estimations. Since we are interested in hazard estimates, *i.e.*in probabilities of being above certain toxic thresholds, we applied conventional Bayesian methods to estimate it, and obtained both their central estimates (the predictive) and their whole distribution (as a family of quantiles).

These hazard estimates were used to quantify the quality of water in the studied river. The conductivity quality index had a value of ~ 0.50 which corresponded with the 95% upper bound of the estimates of the probability of exceeding the threshold $1000\mu S/cm$. Equivalently, the ammonia quality index gave a value of 0.6, associated with the threshold of $0.025ppm$. Given the relatively low human impact on the basin, the high value of these indices is unexpected, and could be related to a kind of uncertainty we have not monitored up to now: time dependence.

The methodology applied is based on the assumption that the different samples are independent of each other, explicitly stated in section 2.4. However, from the observation of time evolution (figure 2.9) of these measurements it is evident that they are mutually strongly dependent, with a clear daily drift.

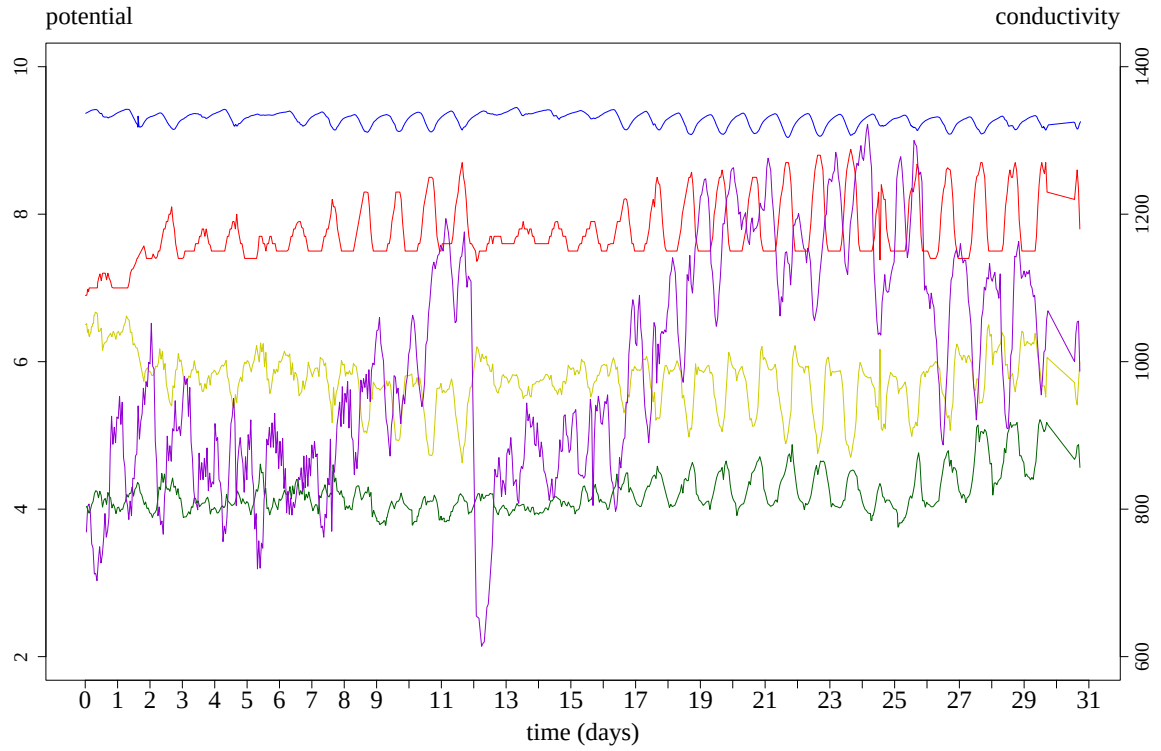

Figure 2.9: Time evolution over July 2002 of all the measured variables, expressed in coordinates in their respective sample spaces. From top to bottom: $p\mathrm{K}_a$, $p\mathrm{H}$, conductivity (in μS/cm with reference to the right scale), $p\mathrm{NH}_3$ and $p\mathrm{NH}_4$.

2.8 Addendum: invariance of coordinate mean and variance

We include here the proofs of the identification between means and variances both in coordinates and as vectors/endomorphisms, or, in other words, that the definitions of these moments given by Eaton [1983] and Pawlowsky-Glahn [2003] are consistent.

Proposition 2.2 *Let $\boldsymbol{Z}$ be a random variable in $\mathbb{E}$ an Euclidean space. Then, the mean of the coordinates of $\boldsymbol{Z}$ equal the coordinates of the mean of $\boldsymbol{Z}$ on $\mathbb{E}$ with respect to any basis (page 21)*

Proof: Let $\mathbf{E}$ be an orthogonal basis of $\mathbb{E}$ (definition 2.8), and $\mathbf{e}_i$ its i-th vector. Then, using the definition of mean on $\mathbb{E}$ (2.6) one may write

$$\langle \mathbf{e}_i, \bar{\mathbf{z}} \rangle_{\mathbb{E}} = \mathrm{E}\left[\langle \mathbf{e}_i, \mathbf{Z} \rangle_{\mathbb{E}}\right]$$

Dividing both sides by $\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}$, one obtains

$$\frac{\langle \mathbf{e}_i, \bar{\mathbf{z}} \rangle_{\mathbb{E}}}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}} = \frac{\mathrm{E}\left[\langle \mathbf{e}_i, \mathbf{Z} \rangle_{\mathbb{E}}\right]}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}}.$$

Note that this expression is always valid because $\mathbf{e}_i \neq \mathbf{n}$, and thus $\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}} > 0$. Taking into account that $\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}$ is constant with respect to $\mathbf{Z}$, one can exchange the division with the expectation,

$$\frac{\langle \mathbf{e}_i, \bar{\mathbf{z}} \rangle_{\mathbb{E}}}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}} = \mathrm{E}\left[\frac{\langle \mathbf{e}_i, \mathbf{Z} \rangle_{\mathbb{E}}}{\langle \mathbf{e}_i, \mathbf{e}_i \rangle_{\mathbb{E}}}\right].$$

which thanks to expression (2.2) ensures us that the coordinates of the mean of $\mathbf{Z}$ on $\mathbb{E}$ will be the mean of the coordinates of $\mathbf{E}$ in an orthogonal basis. If we denote by $\underline{Z}$ the coordinates of $\mathbf{Z}$, and by $\underline{\zeta}$ those of $\bar{\mathbf{z}}$, we may write this last expression as

$$\underline{\bar{\zeta}} = \mathrm{E}\left[\underline{Z}\right]. \tag{2.30}$$

Now consider $\mathbf{F}$ an arbitrary basis. To change coordinates from $\mathbf{E}$ to $\mathbf{F}$ we need the matrix of change of basis $\underline{\underline{\varphi}}$, containing in each column the coordinates of an element of $\mathbf{E}$ with respect to the basis $\mathbf{F}$. Note that both this matrix and the vectors of expression (2.30) are real, and they can be operated with standard real algebra. In particular, to change the basis in which vectors are represented we use matrix multiplication,

$$\underline{\underline{\varphi}} \cdot \underline{\bar{\zeta}} = \underline{\underline{\varphi}} \cdot \mathrm{E}\left[\underline{Z}\right],$$

and since this product is a linear operation, it commutes with the expectation operator, giving

$$\underline{\underline{\varphi}} \cdot \underline{\bar{\zeta}} = \mathrm{E}\left[\underline{\underline{\varphi}} \cdot \underline{Z}\right],$$

thus

$$\underline{\bar{\zeta}}_{\mathbf{F}} = \mathrm{E}\left[\underline{Z}_{\mathbf{F}}\right],$$

where the subindex $\cdot_{\mathbf{F}}$ denotes the basis used, which now is an arbitrary one. □

Proposition 2.3 *Let $\boldsymbol{Z}$ be a random variable on $\mathbb{E}$ an (C-dimensional) Euclidean space. Then, the variance matrix $\underline{\underline{\Sigma}}$ of the coordinates of $\boldsymbol{Z}$ with respect to an orthonormal basis equals the matrix representation of the variance (as an endomorphism Σ) of $\boldsymbol{Z}$ on $\mathbb{E}$ with respect to that basis.*

Proof: If we work in an orthonormal basis, we may use property 2.2 to express each of the vectors and operators from (2.7) in coordinates with respect to this basis, yielding

$$\langle \underline{\xi}, \underline{\underline{\Sigma}} \cdot \underline{\upsilon} \rangle = \mathrm{E}\left[\langle \underline{\xi}, \underline{Z} - \underline{\bar{\zeta}} \rangle \langle \underline{\upsilon}, \underline{Z} - \underline{\bar{\zeta}} \rangle \right],$$

which can be developed as

$$\sum_{i=1}^{C} \xi_i \sum_{j=1}^{C} \Sigma_{ij} \upsilon_j = \mathrm{E}\left[\sum_{i=1}^{C} \xi_i \left(Z_i - \bar{\zeta}_i \right) \sum_{j=1}^{C} \upsilon_j \left(Z_j - \bar{\zeta}_j \right) \right],$$

and reordered to

$$\sum_{i=1}^{C} \sum_{j=1}^{C} \xi_i \upsilon_j \Sigma_{ij} = \sum_{i=1}^{C} \sum_{j=1}^{C} \xi_i \upsilon_j \mathrm{E}\left[\left(Z_i - \bar{\zeta}_i \right) \left(Z_j - \bar{\zeta}_j \right) \right].$$

This expression must be true for all $\underline{\xi}$ and $\underline{\upsilon}$, which implies that the coefficients must be equal,

$$\Sigma_{ij} = \mathrm{E}\left[\left(Z_i - \bar{\zeta}_i \right) \left(Z_j - \bar{\zeta}_j \right) \right]. \tag{2.31}$$

Note that there is a conceptual difference between both sides of the equality. The left-hand side is the (i, j) element of the matrix associated with the endomorphism Σ once it is expressed in an orthonormal basis. The right-hand side is the (real classical) covariance of the (i, j) coordinates of the random variable $\mathbf{Z}$ expressed in the same basis. In other words, we have proven that Pawlowsky-Glahn [2003] definition of characteristic element of dispersion is equivalent to the definition of variance on $\mathbb{E}$ as an operator given by Eaton [1983]

□

Chapter 3

Geostatistics in the real space

This chapter presents the basics of the theory of regionalized variables, or geostatistics, originally developed by Matheron [1965]. These techniques allow the treatment of samples which are non-independent due to their spatial proximity. Recall that classical statistical methods call for an independent, identically-distributed sample. Instead, geostatistics assumes some sort of spatial stationarity, and a known model of dependence of the regionalized variable. This is usually the variogram, which explains how different become two samples as the distance between their sampling locations increase. Using the variogram and the stationarity assumption, both inference (estimation of the mean) and prediction (estimation of the value at an unsampled location) can be achieved with kriging techniques. Kriging yields best linear unbiased estimators and predictors, and allows to describe the uncertainty attached to the prediction with an error variance. These results give the probability function of the sought variable, in the case of a Gaussian model. Geostatistics deals also with support effects: a property is always measured over samples of a given volume (area, length, duration). We can define a different regionalized variable for each support, and model the relationship between them. The final interest of such an approach is the inference of spatial (or time) averages on given blocks (*e.g.*ten-minutes average of conductivity) from data measured in smaller supports (*e.g.*five-seconds average of conductivity). Apart from the seminal work by Matheron [1965], a classical comprehensive reference on geostatistics is Journel and Huijbregts [1978], and introductory ones Clark [1979] and Clark and Harper [2000]. An introductory approach focused on multivariate cases can be found in Wackernagel [1998]. We closely follow the exposition of Chilès and Delfiner [1999], a recent comprehensive textbook. All results summarized in this chapter are extracted from this last book, if not stated otherwise.

3.1 Random function

A look at figure 2.9 shows that each series of the represented measurements present a mixture of two patterns. From one side, there is a clear oscillatory trend. From the other, a quite homogeneous noise blurs these series. The observations present

a mixture of randomness and dependence. Geostatistics allows to do inference with the random character of this sample, exploiting its non-independence through a pre-specified structure of dependence. This section introduces the basic concepts, the next one deals with the specification of these dependence structures, and section 3.3 presents the inference procedure.

Let $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$ be a *point* in a domain $\mathcal{D}$ of the space-time real space $\mathbb{R}^p$, with typically $p \in \{1, 2, 3, 4\}$. We denote by $\mathbf{Z}(\vec{x}) \in \mathbb{R}^D$ a vector-valued function of the location $\vec{x}$ which image has as sample the D-dimensional real space, denoted $\mathbb{R}^D$.

Definition 3.1 (Random function) *Let the function $\boldsymbol{Z}(\vec{x})$ have as image a random vector for any $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$. Then $\boldsymbol{Z}(\mathcal{D})$ is called a* random function *(abbreviated as* RF*).*

Definition 3.2 (Realization) *Any outcome $\boldsymbol{z}(\vec{x})$ of this RF can be viewed as a mapping $\boldsymbol{z}(\cdot) : \mathbb{R}^p \to \mathbb{E}$, called a* realization *(or sample function).*

In accordance with the notation introduced in chapter 2, a given realization at any location will be denoted by lowercase characters, $\mathbf{z}(\vec{x}) = (\zeta_1(\vec{x}), \zeta_2(\vec{x}), \ldots, \zeta_D(\vec{x})) = \underline{\zeta}(\vec{x})$, since, being a real-valued random vector, its coordinates coincide with its values. Uppercase characters $\mathbf{Z}(\vec{x}) = (Z_1(\vec{x}), Z_2(\vec{x}), \ldots, Z_D(\vec{x})) = \underline{Z}(\vec{x})$ denote consequently a RF. Recall that an underlining (*e.g.*$\underline{\zeta}$) will represent a real-valued vector, and a double underlining a matrix of real coefficients.

Most usually, the domain $\mathcal{D}$ is either a bounded continuous region, *e.g.*a volume of the physical space, or an infinite series of time moments, *e.g.*extending from the present to the future. A RF may be seen as an infinite collection of random vectors, where each random vector is linked to a given position $\vec{x}$ in the domain $\mathcal{D}$. In the case of RFs on continuous bounded domains, this collection has uncountably many elements. The realization is then an infinite collection of fixed values, forming a mapping on the domain $\mathcal{D}$. The infinite nature of RFs and realizations preclude any direct observation of them as a whole: instead one can only observe a given regionalized sample, the observed values of the RF at some locations. To estimate characteristics of the RF, or predict its value at unsampled locations, one need some further theoretical assumptions, discussed in detail in *e.g.*Chilès and Delfiner [1999]. We will only concern about stationarity.

Definition 3.3 (Stationarity) *Let $\boldsymbol{Z}(\vec{x})$ be a RF with domain $\mathcal{D} \subset \mathbb{R}^p$ and image $\mathbb{R}^D$. Then it is called*

1. strongly stationary*, when for any set of $B_n \subset \mathbb{R}^D$ and for all set of locations $\{\vec{x}_n\} \in \mathcal{D}$, the following probability is invariant by translation $\vec{h}$:*

$$\Pr\left[(Z_1(\vec{x}+\vec{h}) \in B_1) \cap (Z_2(\vec{x}+\vec{h}) \in B_2) \cap \ldots \cap (Z_N(\vec{x}+\vec{h}) \in B_N)\right] =$$

$$= \Pr\left[(Z_1(\vec{x}) \in B_1) \cap (Z_2(\vec{x}) \in B_2) \cap \ldots \cap (Z_N(\vec{x}) \in B_N)\right];$$

2. second-order stationary*, when for any pair of locations $\vec{x}_n, \vec{x}_m \in \mathcal{D}$, the first two moments are translation-invariant, or*

$$\mathrm{E}\left[\boldsymbol{Z}(\vec{x}_n)\right] = \underline{\mu} \quad \text{and} \quad \mathrm{Cov}\left[\boldsymbol{Z}(\vec{x}_n), \boldsymbol{Z}(\vec{x}_m)\right] = \underline{\underline{C}}(\vec{x}_m - \vec{x}_n) = \underline{\underline{C}}_{nm};$$

3. intrinsic*, when for any pair $\vec{x}_n, \vec{x}_m \in \mathcal{D}$, the increments $(\boldsymbol{Z}(\vec{x}_m) - \boldsymbol{Z}(\vec{x}_n))$ have zero mean and stationary variance:*

$$\mathrm{E}\left[\boldsymbol{Z}(\vec{x}_m) - \boldsymbol{Z}(\vec{x}_n)\right] = \underline{0} \quad \text{and} \quad \mathrm{Var}\left[\boldsymbol{Z}(\vec{x}_m) - \boldsymbol{Z}(\vec{x}_n)\right] = \underline{\underline{\gamma}}(\vec{x}_m - \vec{x}_n);$$

Definition 3.4 (Gaussian RF) *A RF $\boldsymbol{Z}$ is called* Gaussian *if for any sample $\{\vec{x}_n\}$ inside its domain $\mathcal{D}$, the joint distribution is a multivariate normal,*

$$\boldsymbol{Z}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N) \sim \mathcal{N}\left(\underline{\mu}, \underline{\underline{\Sigma}}\right)$$

with parameters

$$\underline{\mu} = \begin{pmatrix} \mathrm{E}\left[\boldsymbol{Z}(\vec{x}_1)\right] \\ \mathrm{E}\left[\boldsymbol{Z}(\vec{x}_2)\right] \\ \vdots \\ \mathrm{E}\left[\boldsymbol{Z}(\vec{x}_N)\right] \end{pmatrix} \quad \text{and} \quad \underline{\underline{\Sigma}} = \begin{pmatrix} \underline{\underline{C}}_{11} & \underline{\underline{C}}_{12} & \cdots & \underline{\underline{C}}_{1N} \\ \underline{\underline{C}}_{21} & \underline{\underline{C}}_{22} & \cdots & \underline{\underline{C}}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\underline{C}}_{N1} & \underline{\underline{C}}_{N2} & \cdots & \underline{\underline{C}}_{NN} \end{pmatrix}$$

with

$$\underline{\underline{C}}_{nm} = \underline{\underline{C}}(\vec{x}_n, \vec{x}_m) = \mathrm{E}\left[\left(\boldsymbol{Z}(\vec{x}_n) - \mu(\vec{x}_n)\right) \cdot \left(\boldsymbol{Z}(\vec{x}_m) - \mu(\vec{x}_m)\right)^t\right]. \tag{3.1}$$

A RF is called intrinsic Gaussian *if for any sample $\{\vec{x}_n\}$ inside its domain $\mathcal{D}$, the joint distribution of the increments $\boldsymbol{Y}(\vec{x}_i) = \boldsymbol{Z}(\vec{x}_i) - \boldsymbol{Z}(\vec{x}_N)$ is a centered multivariate normal,*

$$\boldsymbol{Y}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{N-1}) \sim \mathcal{N}\left(\underline{0}, \underline{\underline{\gamma}}\right),$$

with $\underline{\underline{\gamma}}$ a block matrix equivalent to $\underline{\underline{\Sigma}}$, but containing variances $\underline{\underline{\gamma}}_{nm}$.

When $\vec{x}_n = \vec{x}_m$, $\underline{\underline{C}}_{nm}$ gives the variance-covariance matrix of the function $\mathbf{Z}(\vec{x}_n)$ at location $\vec{x}_n$, whereas for $\vec{x}_n \neq \vec{x}_m$ one obtain the *cross-covariance* matrix between the random vectors located at $\vec{x}_n$ and $\vec{x}_m$.

For intrinsic RFs, knowledge of the mean of the RF is not available, even it could not exist. Thus expression (3.1) cannot be used to encode the relationship between variables at different locations. In these situations, an auxiliary function matrix is used, which is called *cross-variogram* $(\gamma(\vec{x}_n, \vec{x}_m))$, and is defined as

$$\underline{\underline{\gamma}}_{nm} = \underline{\underline{\gamma}}(\vec{x}_n, \vec{x}_m) = \mathrm{E}\left[\left(\mathbf{Z}(\vec{x}_m) - \mathbf{Z}(\vec{x}_n)\right) \cdot \left(\mathbf{Z}(\vec{x}_m) - \mathbf{Z}(\vec{x}_n)\right)^t\right]. \tag{3.2}$$

From now on, it will be considered that the RF is a *Gaussian function*. Gaussian RFs are a very usual assumption, and the theory of regionalized variables was originally developed for them. They have many interesting, simplifying properties, *e.g.*they are completely specified by their means and covariances (3.1). This implies that strong stationarity of a Gaussian RF is ensured with second-order stationary.

3.2 Structural analysis

3.2.1 General aspects

Under second-order stationarity, the covariance matrix (3.1) between the random vectors $\mathbf{Z}(\vec{x}_n)$ and $\mathbf{Z}(\vec{x}_m)$, located respectively at $\vec{x}_n$ and $\vec{x}_m$, does not depend on these exact locations, but on their relative position $\vec{h} = \vec{x}_m - \vec{x}_n$. This implies that it can be expressed as

$$\underline{\underline{C}}(\vec{h}) = \begin{pmatrix} C_{11}(\vec{h}) & C_{12}(\vec{h}) & \cdots & C_{1D}(\vec{h}) \\ C_{21}(\vec{h}) & C_{22}(\vec{h}) & \cdots & C_{2D}(\vec{h}) \\ \vdots & \vdots & \ddots & \vdots \\ C_{D1}(\vec{h}) & C_{D2}(\vec{h}) & \cdots & C_{DD}(\vec{h}) \end{pmatrix} = \left(C_{ij}(\vec{h})\right). \tag{3.3}$$

Note that this is not necessarily a symmetric matrix. The elements of the diagonal, $C_{ii}(\vec{h})$ are called *(auto)-covariance functions*, while those outside the diagonal, $C_{ij}(\vec{h})$ are called *cross-covariance functions*. The *structural analysis* characterizes the RF through the study of these functions of $\vec{h}$.

From a practical point of view, these functions are *a priori* never known, and they must be estimated from the available data. When the mean $\underline{\mu}$ is known, and for any direction $\vec{h}$, the covariance between the variables Z_i and Z_j is estimated using

$$\hat{C}_{ij}(\vec{h}) = \frac{1}{2N(\vec{h})} \sum_{\vec{x}_n - \vec{x}_m \approx \vec{h}}^{N(\vec{h})} (\zeta_i(\vec{x}_n) - \mu_i) \cdot (\zeta_j(\vec{x}_m) - \mu_j), \tag{3.4}$$

where $N(\vec{h})$ represents the number of pairs of observed locations $\vec{x}_n, \vec{x}_m$ such that their difference is approximately equal to $\vec{h}$, with a certain tolerance [Chilès and Delfiner, 1999, p. 36]. When the mean is not known, the variogram can still be estimated as

$$\hat{\gamma}_{ij}(\vec{h}) = \frac{1}{2N(\vec{h})} \sum_{\vec{x}_m - \vec{x}_n \approx \vec{h}}^{N(\vec{h})} (\zeta_i(\vec{x}_m) - \zeta_i(\vec{x}_n)) \cdot (\zeta_j(\vec{x}_m) - \zeta_j(\vec{x}_n)). \tag{3.5}$$

3.2.2 Auto-covariance functions

Definition 3.5 (Auto-covariance function) *For a given real univariate RF $Z_i(\vec{x}) \in \mathbb{R}, \vec{x} \in \mathcal{D}$, the* auto-covariance function *is defined, according to (3.1), as*

$$C_{ii}(\vec{h}) = \mathrm{E}\left[(Z_i(\vec{x}) - \mu_i)) \cdot (Z_i(\vec{x} + \vec{h}) - \mu_i)\right],$$

provided that the mean μ_i exists and is stationary.

Property 3.1 *The covariance function $C_{ii}(\vec{h})$ explains the degree of similarity of two measurements taken at two different locations at a lag distance $\vec{h}$. It satisfies:*

1. *it is a continuous function, everywhere except at the origin $\vec{h} = \vec{0}$;*
2. *it is an even function, $C_{ii}(\vec{h}) = C_{ii}(-\vec{h})$;*
3. *it tends to zero with increasing $\vec{h}$, $\lim_{\vec{h}\to\infty} C_{ii}(\vec{h}) = 0$;*
4. *it is a positive definite function*

Note: item 3 requires the RF to be ergodic [see Chilès and Delfiner, 1999, pag 19-22, for a proper definition].

Condition 4 is maybe the most important, because it implies that any combination of Z_i taken at different locations will have a valid positive variance; it has two practical implications:

- the covariance function has a maximum at the origin, $|C_{ii}(\vec{h})| < C_{ii}(\vec{0}) = \mathrm{Var}\,[Z(\vec{x})]$; this allows us to define an *auto-correlation function*,

$$\rho_i(\vec{h}) = \frac{C_{ii}(\vec{h})}{C_{ii}(\vec{0})}; \tag{3.6}$$

- the spectral representation of the covariance is always strictly positive, and approaches zero when $\vec{u}$ tends to infinity. Under certain regularity conditions [Chilès and Delfiner, 1999, p. 64, 325-326], this spectral representation is given by

$$F_{jj}(\vec{u}) = \int_{\mathbb{R}^p} \exp\left(-2\pi i \left\langle \vec{h}|\vec{u} \right\rangle\right) \cdot C_{jj}(\vec{h}) \cdot d\vec{h} > 0, \tag{3.7}$$

  where "i" is the imaginary unit, and the symbol $\left\langle \vec{h}|\vec{u} \right\rangle$ represents the classical scalar product of the vectors $\vec{h}$ and $\vec{u}$. If $Z_{jj}(\vec{x} = t)$ is a stochastic process in time, then $F_{jj}(u)$ is interpreted as the energy carried by each frequency u.

Definition 3.6 (Variogram) *The* variogram *of the RF is defined as*

$$\gamma_{ii}(\vec{h}) = \mathrm{Var}\left[Z_k(\vec{x} + \vec{h}) - Z_k(\vec{x})\right].$$

The variogram is frequently used as a structural tool, even under the second-order stationarity assumption, because it does not demand the mean to be known. Provided that both exist, there is an easy relationship between covariance functions and variograms,

$$\gamma(\vec{h}) = C(\vec{0}) - C(\vec{h}). \tag{3.8}$$

Even when the covariance does not exist, this expression can be equally applied with $C(\vec{0})$ an arbitrarily-chosen upper bound, and $C(\vec{h})$ an *equivalent covariance function*. The only condition imposed to the variogram is that this equivalent covariance function satisfies the conditions stated above for true covariance functions. Thus, variograms are even positive functions, and are continuous everywhere except at the origin, where they must have a zero value $\gamma_{ii}(\vec{0}) = 0$. In general, towards infinity, the variogram can increase infinitely. However, when the covariance exists, or the RF is second-order stationary, then it has an upper bound at $\gamma_{ii}(\vec{h}) < 2 \cdot C(\vec{0})$, and towards infinity it must tend to $\lim_{\vec{h}\to\infty} \gamma_{ii}(\vec{h}) = C(\vec{0})$; this value is called the *sill* of the variogram.

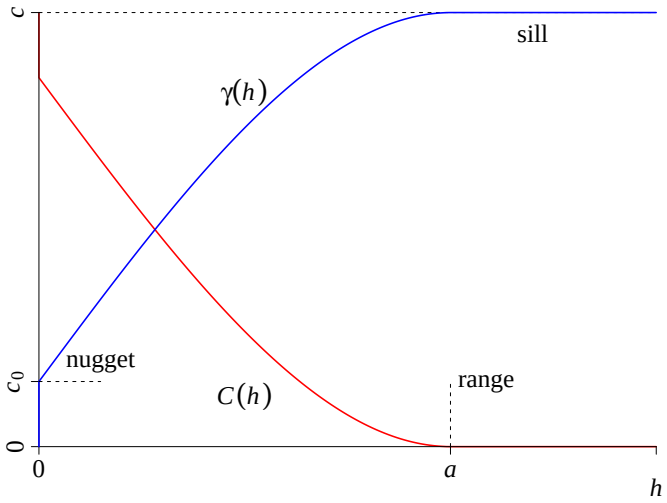

Figure 3.1: Synthetic representation of a covariance and its associated variogram, with their most common features.

Variograms (and hence covariance functions) are described through the following characteristic elements (figure 3.1).

Sill: when it exists, the sill is the value around which the variogram stabilizes for long distances, $\lim_{\vec{h}\to\infty} \gamma_{ii}(\vec{h}) = C(\vec{0})$, and corresponds to the theoretical variance of the RF.

Range: when the sill exists, the range is the distance at which at least 95% of its value is attained; in terms of covariance function, it is the distance at which the covariance drops to (almost) zero, and is thus interpreted as the radius of influence of a location.

Nugget effect: the variogram (and the covariance) can be discontinuous at the origin; though $\gamma_{ii}(\vec{0}) = 0$, it can happen that $\lim_{\vec{h}\to\vec{0}} \gamma_{ii}(\vec{h}) = c_0 \neq 0$. This discontinuity is called the *nugget effect*, and its value is usually represented by c_0.

Behavior around the origin: the shape of the variogram near the origin informs of the degree of continuity of the RF. This shape might be approximated by a curve like $\gamma(\vec{h}) \propto \|h\|^\alpha$ with $0 < \alpha < 2$ (figure 3.2), with the special cases: $\alpha = 0$ gives a nugget effect and implies discontinuity of the RF; $\alpha = 1$ is linked with a (piecewise) continuous function; $\alpha = 2$ is a parabolic behavior, which ensures that the RF is at least piecewise differentiable, and highly regular [Chilès and Delfiner, 1999, p. 51].

Hole effect: this characteristic is identified as a significant oscillation in the variogram or covariance at relatively long distances. This effect indicates a tendency of high values in the RF to be surrounded by low values, and viceversa. *Periodicity* and quasi-periodicity, particularly in time, are special cases of hole effects. A hole effect in more than one dimension must be forcefully *dampened*, in order to approach the sill as the distance increases. Note that an non-dampened hole effect does not approach the sill, but it is still a valid covariance model in one dimension [Chilès and Delfiner, 1999, p. 92-93]. Figure 3.3 shows how a hole effect variogram looks like.

Anisotropy: when the shape of the variogram depends only on the length of $\|\vec{h}\|_{\mathbb{R}} = h$ and not on its direction, we call the structure an *isotropic* one. Naturally, an *anisotropic* variogram depends on this direction. Essentially there are two types of anisotropy: geometric (the most commonly considered) and zonal anisotropy. *Geometric anisotropy* is present when the shape of the variogram in all directions is the same, and the only change is in the ranges; it can be transformed to an isotropic variogram by rotating and scaling the system of reference. Zonal anisotropy implies even different shapes and sill values in each direction, and it is by far much more difficult to deal with.

The classical approach to estimate variogram and covariance functions implies computing their experimental versions (3.4-3.5), and fitting to them a valid model, which must have some of these characteristics. Standardized isotropic classical models—with sill equal to one, thus representing correlation functions of equation (3.6)—are plotted in figures 3.2 and 3.3 and are defined as follows.

Definition 3.7 (Generalized linear or power-law model)

$$\gamma(h) = h^\alpha \quad h > 0;\ 0 < \alpha < 2.$$

For $\alpha = 0$, it is a pure nugget effect. For $\alpha = 1$, it is a linear model, which gives its name to the family. The generalized linear model has no sill, and both the variogram and the associated RF present properties of self-similarity and fractal character: the Gaussian RFs with such a variogram are Brownian motions.

Definition 3.8 (de Wijsian or logarithmic model)

$$\gamma(h) \approx \log\left(\frac{h}{a}\right) + \frac{3}{2}, \quad h > 0;\ a > 0.$$

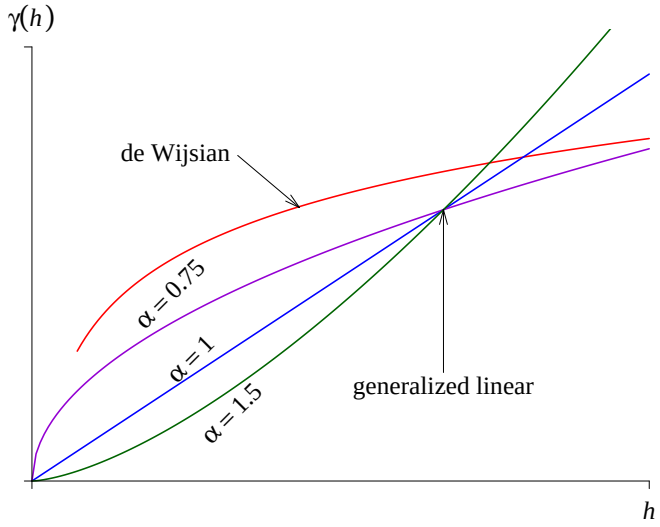

Figure 3.2: Some variogram models without sill.

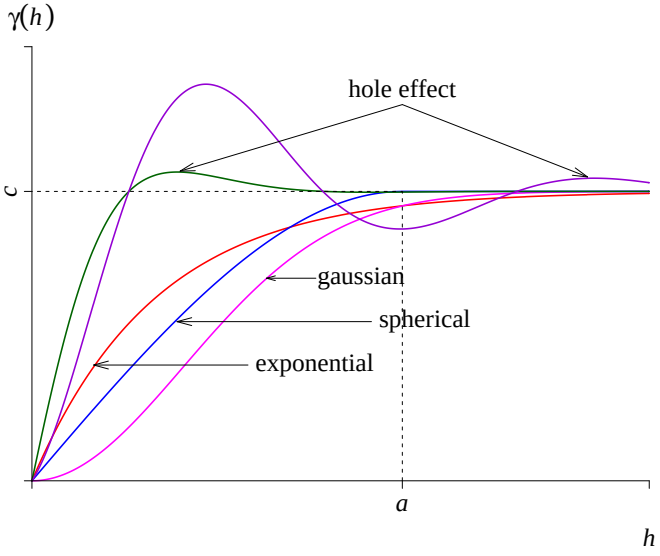

Figure 3.3: Some variogram models with a sill.

For a regularized support, this variogram model has no analytical expression. However, this equation is an approximation for $h > 2a$. This variogram model presents no sill. It has strong ties with the lognormal distribution, and presents also a fractal character.

Definition 3.9 (Spherical)

$$\gamma(h) = \begin{cases} 1 - \frac{3h}{2a} + \frac{r^3}{2a^3}, & \text{if } r \leq a \\ 1, & \text{if } r \geq a \end{cases} \qquad h > 0;\ a > 0. \tag{3.9}$$

It is a valid variogram and covariance model for any positive value of a, and this value coincides with the range of the model: at greater distance than a, the covariance is identically 0. It is classically linked to diffusion phenomena with a limited area of influence. It is one of the most-frequently used.

Definition 3.10 (Generalized exponential or Stable)

$$\gamma(h) = 1 - \exp\left(-\left(\frac{3h}{a}\right)^{\alpha}\right), \quad h > 0;\ 0 < \alpha \leq 2;\ a > 0. \tag{3.10}$$

*This family reaches the sill asymptotically, and the range is defined as the distance at which the correlation coefficient drops to 5%. Two members of this family are commonly used. The exponential model ($\alpha = 1$) is very similar to the spherical one, and is associated to diffusion processes with infinite area of influence. The Gaussian model ($\alpha = 2$) is highly continuous at the origin, which makes its corresponding RF to be infinitely differentiable, exceptionally regular, and almost deterministic; it is a good model for potential fields (*e.g.*gravity) and, in general, deterministic phenomena [Chilès and Delfiner, 1999, p. 85,90].*

Definition 3.11 (Hole effect)

$$\gamma(h) = 1 - \exp\left(-\frac{3h}{a_d}\right) \cdot \cos\left(2\pi\frac{h}{a_t}\right), \quad h > 0;\ a_d, a_t > 0 \tag{3.11}$$

where a_t is the period of the variogram, and a_d the dampening range: at a distance greater than a_d, the hole effect has been reduced to 5% of its original importance, thus considered as zero. In one dimensional problems, any positive values for a_t, a_d are allowed, but for high dimensions the dampening range must be significantly smaller than the period. It describes periodic behavior of variograms and covariance functions.

Definition 3.12 (Composed models) *Valid variogram models for any variable can be defined as a linear combination of the previous models for correlograms (thus with unit sill), each one multiplied by a certain constant, showing its contribution to the total variance,*

$$\gamma_{ii}^{T}(h) = c_0 + \sum_{k}^{K} c_{(k)} \cdot \gamma_{(k)}(h) \quad h > 0;\ c_0, c_{(k)} > 0.$$

An identical expression exists for covariance functions, whenever the models used are linked to valid covariance models,

$$C(h) = c_0 + \sum_{k}^{K} c_{(k)} \cdot \left(1 - \gamma_{(k)}(h)\right). \tag{3.12}$$

Usually, composed models have at least the nugget effect term (denoted by c_0), apart from a correlogram model.

3.2.3 Cross-covariance functions

Definition 3.13 (Cross-covariance function) *For a pair of univariate RFs $(Z_i(\vec{x}), Z_j(\vec{x})) \in \mathbb{R}^2, \vec{x} \in \mathcal{D}$, their cross-covariance function is defined, according to (3.1), as*

$$C_{ij}(\vec{h}) = \mathrm{E}\left[(Z_i(\vec{x}) - \mu_i)) \cdot (Z_j(\vec{x} + \vec{h}) - \mu_j)\right],$$

provided that the means (μ_i, μ_j) exist and are stationary.

Generalizing the auto-covariance function concept, the cross-covariance function explains the mutual linear information of a measurement of two different variables taken respectively at two different locations (separated by a lag vector $\vec{h}$). Remember that in general the cross-covariance functions are not symmetric, although

$$C_{ij}(\vec{h}) = C_{ji}(-\vec{h}).$$

This means that cross-covariance functions are neither odd nor even functions. They are bounded by the auto-covariance functions through Cauchy-Schwarz's inequality,

$$C_{ij}(\vec{h}) \leq \sqrt{C_{ii}(0) \cdot C_{jj}(0)}.$$

Cross-covariances can be described through the same elements used with auto-covariances (figure 3.1), plus a *delay* or *off-set effect*: maximum correlation between two different variables does not necessarily occur at the same location, but displaced a certain lag $\vec{h}$. In particular, hole effects apply also to cross-covariances. Also, ranges in both positive and negative directions can be identified, where cross-covariance vanishes. Finally, cross-covariances can have nugget effects elsewhere, not only at the origin [Chilès and Delfiner, 1999]. However, cross-covariance functions are not modelled by using only particular models, since they can seldom ensure that the resulting set of covariance functions (3.3) is a valid joint covariance model.

Property 3.2 (Cramér, 1940) *Under certain regularity conditions, the set of auto- and cross-covariance functions $C_{jk}(\vec{h})$ form a valid model if*

$$F_{jk}(\vec{u}) = \int_{\mathbb{R}^p} \exp\left(-2\pi i \left\langle \vec{h} | \vec{u} \right\rangle\right) \cdot C_{jk}(\vec{h}) \cdot d\vec{h}, \tag{3.13}$$

i.e. *the set of spectral densities $F_{jk}(\vec{u})$ associated to all $C_{jk}(\vec{h})$, form a positive definite Hermitian matrix for all frequencies $\vec{u}$ [Chilès and Delfiner, 1999]. Recall that here i represents the imaginary unit, and $\left\langle \vec{h}|\vec{u} \right\rangle$ is the classical scalar product between vectors $\vec{h}$ and $\vec{u}$.*

This formulation of Cramér criterion is more restrictive than the original one, which does not demand the densities to exist, but only their measures. However, this approach is enough in the scope of this Thesis.

Definition 3.14 (Cross-variogram) *Cross-variograms are defined as*

$$\gamma_{ij}(\vec{h}) = \mathrm{E}\left[\left(Z_i(\vec{x}) - Z_i(\vec{x}+\vec{h})\right) \cdot \left(Z_j(\vec{x}) - Z_j(\vec{x}+\vec{h})\right)\right],$$

Cross-covariograms present the following relationship with cross-covariances

$$\gamma_{ij}(\vec{h}) = C_{ij}(\vec{0}) - \frac{1}{2}\left(C_{ij}(\vec{h}) + C_{ij}(-\vec{h})\right), \tag{3.14}$$

Note that cross-variograms do not capture asymmetric features of the covariance structure. Cressie [1991] introduces a measure of joint variation which keeps this information. His definition uses increments between different variables, which may have no physical sense. On the other hand, van den Boogaart and Brenning [2001] shows that a generalized cross-covariance can be computed with an estimated mean value, and that prediction results this generalized covariance will yield are equivalent (up to the addition of a constant, as is explained in next section) to those obtained with the true cross-covariance.

Equation (3.14) implies that cross-variograms are even functions, and satisfy $\gamma_{ij}(\vec{0}) = 0$. Using variograms and cross-variograms, a *coefficient of codispersion* [Wackernagel, 1998] is defined,

$$R_{ij} = \frac{\gamma_{ij}(\vec{h})}{\sqrt{\gamma_{ii}(\vec{h}) \cdot \gamma_{jj}(\vec{h})}}. \tag{3.15}$$

The coefficient of codispersion is bounded to $|R_{ij}(\vec{h})| \leq 1$, since by Cauchy-Schwarz inequality

$$\gamma_{ij}(\vec{h}) \leq \sqrt{\gamma_{ii}(\vec{h}) \cdot \gamma_{jj}(\vec{h})}.$$

These conditions on cross- and auto- correlations and variograms are necessary conditions to ensure that the joint model (3.2) or (3.3) is a valid one. But the necessary and sufficient condition is that (3.13) defines a positive definite matrix of frequency spectra.

To implement these conditions and obtain a valid covariance or variogram system, there are some particular methodologies. From them, the linear model of corregionalization will be used. The linear model of corregionalization generalizes (3.12) through

a decomposition of the whole covariance matrix

$$\underline{\underline{C}}(\vec{h}) = \sum_{k}^{K} \left(1 - \gamma_{(k)}(\vec{h})\right) \cdot \underline{\underline{C}}_{(k)},$$

where $\gamma_{(k)}(\vec{h})$ are K valid variogram models (possibly including a nugget effect), and $\underline{\underline{C}}_{(k)}$ form a set of K matrices. The model is automatically valid when the matrices are all positive definite. However, this condition is a sufficient but not necessary one. In fact, it is very restrictive: for instance, it cannot handle offset effects in cross-covariances [Wackernagel, 1998]. Instead, Yao and Journel [1998] suggest to validate a discrete version of the joint covariance model by obtaining a valid frequency spectrum (as described in proposition 3.2). This is a general validation approach, but it still lacks a straightforward implementation.

3.3 Linear prediction

3.3.1 General universal kriging

The main use of the RF formalism and the structural analysis functions is interpolation. As was said in section 3.1, a RF $\mathbf{Z}(\mathcal{D})$ as a whole cannot be observed, and instead one have always a regionalized sample, a collection of observations at different locations $\mathbf{z}_n = \mathbf{z}(\vec{x}_n), \vec{x}_n \in \mathcal{D}, n = 1, \ldots, N$. The goal will be the estimation of (some of the components of) the vector $\mathbf{Z}(\vec{x}_0)$, by using the observed sample. A measure of the error incurred is also sought. Kriging is a technique which provide these estimations through linear combinations of the observed data, and is regarded as best linear unbiased estimator, in the sense that its error has minimal variance among all linear combinations of observations. When the RF $\mathbf{Z} \in \mathbb{R}^C$ is a Gaussian one, it is known that kriging delivers also the conditional distribution of $\mathbf{Z}(\vec{x}_0)$ on the observed data. In the literature, the term *kriging* usually applies to univariate RFs, and *cokriging* is used when RFs have a vector as image. We will not do such a distinction, simply calling all these techniques by the generic name of kriging.

In general, all the coordinates of the vector $\mathbf{Z}$ shall not be observed at all locations, a situation called *non-collocated sampling*, or *undersampled case* [Journel and Huijbregts, 1978]. Contrarily, when at all sampled locations a whole vector is obtained, it is called a *collocated sampling* . In the first case, the most generic, one must consider each coordinate separately as an univariate RF, $Z_k(\mathcal{D})$. The sampled values observed for this RF, being real values, will be denoted by $\zeta_k(\vec{x}_n), n = 1, \ldots, N_k$. Assume that the mean $\mu_k(\vec{x})$ of these RFs is not known, but that it is known to be a linear combination of $(A + 1)$ known functions $\{f_0(\vec{x}), f_1(\vec{x}), \ldots, f_A(\vec{x})\}$ called *drift functions*. Usually, these function are polynomials or trigonometric functions, and $f_0(\vec{x}) \equiv 1$ is also a common assumption. In this situation, the best linear predictor (denoted by ζ_0^*) of the value of $Z_k(\vec{x}_0)$ at an unsampled location $\vec{x}_0 \in \mathcal{D}$, using the available information on

the drift and the observed values, is called *universal kriging* (UK) and is defined as a weighted *arithmetic* average

$$\zeta_0^* = \sum_{i=1}^{C} \sum_{n=1}^{N_i} \lambda_{ni} \cdot \zeta_i(\vec{x}_n). \tag{3.16}$$

The error variance attached to this estimator is called *kriging variance*, and its value is

$$\begin{aligned} \sigma_{UK}^2 &= \mathrm{E}\left[(z_0^* - Z_k(\vec{x}_0))^2\right] = C_{kk}(\vec{x}_0, \vec{x}_0) - 2\sum_{i=1}^{C} \sum_{n=1}^{N_i} \lambda_{ni} \cdot C_{ik}(\vec{x}_n, \vec{x}_0) + \\ &+ \sum_{i,j=1}^{C} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} \lambda_{ni}\lambda_{mj} C_{ij}(\vec{x}_n, \vec{x}_m), \end{aligned} \tag{3.17}$$

where $C_{ij}(\vec{x}_n, \vec{x}_m) = C_{ij}(\vec{x}_m - \vec{x}_n)$ is the covariance function between $Z_i(\vec{x}_n)$ and $Z_j(\vec{x}_m)$. The weight values λ_{ni} are obtained by minimizing the variance (3.17) subject to the so-called *universality conditions*, one for each function of the drift and each variable involved:

$$\sum_{n=1}^{N_i} \lambda_{ni} f_a(\vec{x}_n) = \delta_{ik} f_a(\vec{x}_0). \tag{3.18}$$

Using Lagrange coefficients ν_a, the weights can be computed through the system of equations

$$\sum_{j=1}^{C} \sum_{m=1}^{N_j} \lambda_{mj} C_{ij}(\vec{x}_n, \vec{x}_m) + \sum_{a=0}^{A} \nu_a f_a(\vec{x}_n) = C_{ik}(\vec{x}_n, \vec{x}_0)\,, \quad i = \{1, \ldots, C\}\,; n = \{1, \ldots, N_i\}$$

$$\sum_{n=1}^{N_i} \lambda_{ni} f_a(\vec{x}_n) = \delta_{ik} f_a(\vec{x}_0)\,, \quad i = \{1, 2, \ldots, C\}; a = \{0, 1, \ldots, A\}. \tag{3.19}$$

Note that in these equations, k is the index of the predicted variable, a the drift function index, i the predictand variable index, n the datum index, and 0 the index of the predicted location. For each i-th variable, the system has $N_i + A + 1$ equations and unknowns, one for each location i and one for each drift function a. Although fairly large, this system of equations can be solved through classical methods for linear systems. Three particular and simpler cases will be considered now: ordinary kriging, kriging of the drift and simple kriging.

In the first case, when only the constant drift function is considered $f_0(\vec{x}) \equiv 1$, the method here described amounts to simply considering the mean to be constant but unknown. This case is the most common, and it is called *ordinary, or general, kriging*. It can be performed using variograms and generalized cross-covariance instead of covariances. Recall that cross-variograms can replace generalized cross-covariance functions when these are symmetric.

3.3.2 Kriging of the drift

Now the goal is the estimation of the drift coefficients for a single variable Z_k. Since it will be used here only this variable, the index for the k-th variable is dropped in this section. Considering other variables would simply imply the extension of the covariance and drift matrices involved, in a similar fashion to what is done in the last section.

Recall that the drift $\mu(\vec{x}_n)$ has an expression

$$\mu^*(\vec{x}_n) = \sum_{a=1}^{A} \alpha_a f_a(\vec{x}_n), \quad \text{or} \quad \underline{\mu} = \underline{\underline{F}} \cdot \underline{\alpha}$$

which allows a decomposition of observed values $\zeta(\vec{x})$ in drift and residual values

$$\underline{\zeta} = \underline{\underline{F}} \cdot \underline{\alpha} + \underline{\upsilon}$$

We use here a matrix notation, where $\underline{\alpha} = (\alpha_a)$ is the vector of drift coefficients, $\underline{\zeta} = (\zeta(\vec{x}_n))$ are the observed values of Z at the sampled locations, $\underline{\upsilon} = (\upsilon(\vec{x}_n))$ are the residuals of Z after substracting the drift, $\underline{\underline{F}} = (F_{na})) = (f_a(\vec{x}_n))$ is the matrix of drift functions computed at each sampled location, and $\underline{\underline{\Sigma}} = (C(\vec{x}_n, \vec{x}_m))$ is the covariance matrix of the residuals at each sampled location, with $a = 0, 1, \ldots, A$ and $n, m = 1, 2, \ldots, N$.

Generalized least squares theory allows us to estimate the vector of coefficients $\underline{\alpha}$ through the matrix expression

$$\underline{\alpha}^* = \left(\underline{\underline{F}}^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \underline{\underline{F}}\right)^{-1} \cdot \underline{\underline{F}}^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \underline{\zeta}. \tag{3.20}$$

The variance-covariance error matrix associated to the estimator 3.20 is computed as

$$\mathrm{Cov}\left[\underline{\alpha}^*, \underline{\alpha}^*\right] = \left(\underline{\underline{F}}^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \underline{\underline{F}}\right)^{-1}. \tag{3.21}$$

Kriging of the drift gives the same results as universal kriging (in its univariate form) when the covariance between the data set and the kriged location has dropped to zero: in the system of equations of expression 3.19, the right-hand terms of the equations of the first kind are all zero. Outside the range, universal kriging results coincide with the drift [Chilès and Delfiner, 1999, p. 179]. It is interesting to note that the kriging variance of the drift itself at a given location can be easily computed using (3.21) by

$$\sigma^2_{KD} = \mathrm{Var}\left[m^*(\vec{x}_n)\right] = \mathrm{E}\left[\left(\mu^*(\vec{x}_n) - \mu(\vec{x}_n)\right)^2\right] = \underline{f}_n^t \cdot \left(\underline{\underline{F}}^t \cdot \underline{\underline{\Sigma}}^{-1} \cdot \underline{\underline{F}}\right)^{-1} \cdot \underline{f}_n, \tag{3.22}$$

where $\underline{f}_0 = (f_a(\vec{x}_0))$ is the vector of drift functions at location $\vec{x}_0$.

3.3.3 Simple kriging

For a single variable Z_k, which mean $\mu_k(\vec{x})$ is known everywhere in the domain $\vec{x} \in \mathcal{D}$, the technique used is called *simple kriging* (SK). Since k will be a constant, in this

section it is dropped from the notation: $Z = Z_k$ is the RF, $\mu(\cdot) = \mu_k(\cdot)$ its known mean, $\lambda_n = \lambda_{nk}$ the kriging weights, $\zeta(\cdot) = \zeta_k(\cdot)$ the regionalized sample, and $C(\cdot) = C_{kk}(\cdot)$ the covariance matrix. The predictor is

$$\zeta_0^* = \sum_{n=1}^{N} \lambda_n \zeta(\vec{x}_n) + \left(1 - \sum_{n=1}^{N} \lambda_n\right) \mu(\vec{x}_0). \tag{3.23}$$

The weights λ_n are obtained by solving the system of equations

$$\sum_{m=1}^{N} \lambda_m C(\vec{x}_n, \vec{x}_m) = C(\vec{x}_n, \vec{x}_0), \quad n = 1, 2, \ldots, N, \tag{3.24}$$

which minimizes the simple kriging variance

$$\sigma_{SK}^2 = \sum_{n,m=1}^{N} \lambda_n \lambda_m C(\vec{x}_n, \vec{x}_m) - 2 \sum_{n=1}^{N} \lambda_n C(\vec{x}_n, \vec{x}_0) + C(\vec{x}_0, \vec{x}_0). \tag{3.25}$$

Note that the extension of this technique to incorporate information from other variables can be done in a straightforward way by considering cross-covariances in the matrix $\underline{\underline{C}} = (C(\vec{x}_n, \vec{x}_m))$. A full study of this case is the subject of the next chapter.

3.3.4 Properties of kriging estimators

Simple kriging predictor (3.23) is the best linear unbiased predictor, once the mean is known. It satisfies the unbiasedness condition $\mathrm{E}\left[\zeta_0^* - Z(\vec{x}_0)\right] = 0$, and its error variance $\mathrm{Var}\left[\zeta_0^* - Z(\vec{x}_0)\right]$ is minimal by construction. In a Gaussian RF, simple kriging predictor can be interpreted as the result of a regression of the unknown variable at the unsampled location by using as predictands the variables at the sampled locations. The regression equation is (3.23), the joint distribution is a multivariate normal with parameters delivered by equation (3.1), the normal equations that must be solved to obtain the regression are (3.24), and the error variance attached to the regression is (3.25). As a standard result of regression theory, the conditional distribution of the predicted variable is

$$\left[Z(\vec{x}_0) | \zeta(\vec{x}_1), \zeta(\vec{x}_2), \ldots, \zeta(\vec{x}_N)\right] \sim N\left(\zeta_0^*, \sigma_{SK}^2\right). \tag{3.26}$$

This implies that

$$\mathrm{E}\left[\zeta_0^* | \zeta(\vec{x}_1), \zeta(\vec{x}_2), \ldots, \zeta(\vec{x}_N)\right] = \mathrm{E}\left[Z(\vec{x}_0) | \zeta(\vec{x}_1), \zeta(\vec{x}_2), \ldots, \zeta(\vec{x}_N)\right]. \tag{3.27}$$

So, conditionally to the observed data, the expectation of the predictor and the expectation of the unsampled variable are equal. Furthermore, ζ_0^* satisfies the *conditional unbiasedness* property

$$\mathrm{E}\left[Z(\vec{x}_0) | \zeta_0^*\right] = \zeta_0^*,$$

which is even of greater importance than the minimum variance condition. Indeed, for instance in water management, the decision to accept or reject a water volume for agricultural use may depend on the *estimated* conductivity, whereas its effects in the functions depend on the *true* conductivity. Conditional unbiasedness ensures that, on the average, we get what we expect [Chilès and Delfiner, 1999, p. 164].

All these properties of simple kriging (be it univariate or multivariate) disappear when the mean is unknown. Universal kriging is also by construction an unbiased linear predictor of minimum variance, but it is no longer conditionally unbiased, nor equation (3.27) holds. However, kriging tends to minimize the conditional bias in any case, even for non-Gaussian functions, since the kriging variance admits a decomposition

$$\mathrm{Var}\left[Z(\vec{x}_0) - \zeta_0^*\right] = \mathrm{E}\left[\mathrm{Var}\left[Z(\vec{x}_0)|\zeta_0^*\right]\right] + \mathrm{E}\left[\left(\mathrm{E}\left[Z(\vec{x}_0)|\zeta_0^*\right] - \zeta_0^*\right)^2\right],$$

where the first term on the right side is the expected conditional variance, and the second one is the variance of the conditional bias. Thus, one can be confident that the true conditional distribution is not far from an equivalent distribution to (3.26) with mean $\zeta^*_{0(UK)}$ and variance σ^2_{UK}. An assessment of this approximation is given by the analysis of the slope β of the regression of the true value $Z(\vec{x}_0)$ on its predictor ζ_0^*,

$$\beta = \frac{\mathrm{Cov}\left[Z(\vec{x}_0), \zeta_0^*\right]}{\mathrm{Var}\left[\zeta_0^*\right]} = 1 - \left(1 - \sum_{n=1}^{N} \lambda_{n(SK)}\right) \frac{\mathrm{Var}\left[m^*(\vec{x}_0)\right]}{\mathrm{Var}\left[\zeta_0^*\right]}. \tag{3.28}$$

Here $\lambda_{n(SK)}$ are weights of simple kriging (3.24), $\mathrm{Var}\left[\zeta_0^*\right]$ is the variance of the kriging method used, and $\mathrm{Var}\left[m^*(\vec{x}_0)\right]$ is the variance of the drift (3.22). The more similar β to one, the better the approximation of SK by UK is. This is achieved either by SK weights summing up to one, or by a small variance of the drift.

3.4 Bayesian Methods

3.4.1 Bayesian kriging

Omre [1987] introduced a model which essentially considers the drift functions of universal kriging to be a smooth RF, which first and second moments are known, a so-called *qualified guess*. Let us explain it in an univariate case: $Z(\vec{x})$ is a RF in a domain $\vec{x} \in \mathcal{D}$, and its qualified guess is $M(\vec{x})$, also a RF in the same domain. Let the moments of $M(\vec{x})$ be known, but not necessarily stationary,

$$\begin{aligned} \mathrm{E}\left[M(\vec{x})\right] &= \mu_M(\vec{x}) \\ \mathrm{Cov}\left[M(\vec{x}_n), M(\vec{x}_m)\right] &= C_M(\vec{x}_n, \vec{x}_m), \end{aligned}$$

and let the conditional moments of $Z(\vec{x})$ on $M(\vec{x})$ satisfy a generalized second-order stationarity condition

$$\begin{aligned} \mathrm{E}\left[Z(\vec{x})|M(\vec{x})\right] &= a_0 + M(\vec{x}) \\ \mathrm{Cov}\left[Z(\vec{x}_n), Z(\vec{x}_m)|M(\vec{x})\right] &= C_{Z|M}(\vec{x}_n - \vec{x}_m). \end{aligned}$$

Due to standard relationships between conditional and non-conditional moments of random variables, the non-conditional versions of the moments of $Z(\vec{x})$ are

$$\begin{aligned} \mathrm{E}\left[Z(\vec{x})\right] &= a_0 + \mu_M(\vec{x}) \\ \mathrm{Cov}\left[Z(\vec{x}_n), Z(\vec{x}_m)\right] &= C_{Z|M}(\vec{x}_n - \vec{x}_m) + C_M(\vec{x}_n, \vec{x}_m). \end{aligned} \tag{3.29}$$

Notice that this model amounts to the classical RF with an unknown but constant drift. Thus, it can be treated by ordinary kriging (page 59) with a composite covariance given by equation (3.29).

Finally, the model offers estimates of the RF $Z(\vec{x}_0)$ at unsampled locations, as well as a kriging variance. Under the assumption that both $M(\vec{x})$ and $Z(\vec{x})$ form a jointly bi-variate Gaussian function, these two measures can be interpreted as the mean and the variance of a normal distribution at the unsampled location $\vec{x}_0$. The validity of this interpretation is nevertheless subject to the limitations exposed in section 3.3.4 regarding the conditional expectation properties of universal kriging. No further assessment of uncertainty affecting the estimates was derived by Omre [1987] from this bayesian framework.

Further steps in the bayesian treatment of spatial problems were given by Le and Zidek [1992] and Handcock and Stein [1993], who introduced different bayesian models for estimation of parametric covariances. Diggle et al. [1998] account for these and other bayesian *improvements* of the estimation process, specially regarding its uncertainty. In this line, Chilès and Delfiner [1999, p. 190] note that the bayesian modelling of the drift tends to reduce the uncertainty of the final estimates, while a bayesian modelling of covariance parameters tends to increase it.

3.4.2 Model-based geostatistics

Diggle et al. [1998] introduce another model, which has the same relationship with kriging as generalized linear models [Nelder and Wedderburn, 1972] have with linear regression. In the scope of their model, the available data (a sample $y_1, y_2, \ldots, y_N$) are assumed to be generated by a model like

$$Y(\vec{x}) = \mu + S(\vec{x}) + Z,$$

with μ a constant mean effect, S a stationary Gaussian RF with $\mathrm{E}\left[S(\vec{x})\right] = 0$ and $\mathrm{Cov}\left[S(\vec{x}_n), S(\vec{x}_m)\right] = \sigma^2\rho(\vec{x}_m - \vec{x}_m)$, and $Z \sim \mathcal{N}(0, \tau^2)$ a white noise, independent of location. This model has the particular property that, conditional on the values of $S(\vec{x})$, the $Y(\vec{x})$ are mutually independent variables with distribution $Y|S(\vec{x}) \sim \mathcal{N}(\mu + S(\vec{x}), \tau^2)$.

The next step is assuming the existence of a series of explanatory variables for $S(\vec{x})$, and replacing the normal assumption by a generalized linear model, which says that for a known *link function* $h(\cdot)$,

$$h\left(\mathrm{E}\left[Y|S(\vec{x})\right]\right) = \sum_{a=1}^{A} f_a(\vec{x})\beta_a + S(\vec{x}).$$

The explanatory variables $f_a(\vec{x})$ play the role of the drift functions of universal kriging, and the β_a parameters of the drift coefficients.

The model is specified through the marginal distribution of $S(\vec{x})$ and the conditional distribution of $Y(\vec{x})$ on $S(\vec{x})$. In a bayesian framework, inference on the various parameters of the model $(\mu, \sigma, \tau, \beta_a)$ or any parameter of the correlation structure $\rho(\vec{x}_n - \vec{x}_m)$ would require the use of the likelihood of the marginal distribution of the observable $Y(\vec{x})$, which is not directly available. Then the authors use extensive computation methods (namely Markov Chain Monte Carlo, or MCMC methods) to:

1. estimate the posterior distribution of the correlation parameters from the marginal distribution of $S(\vec{x})$, which is known,
2. estimate the joint distribution of $S(\vec{x})$ and $Y(\vec{x})$,
3. estimate the posterior distribution of $S(\vec{x})$ conditional on the data, the estimated posterior distribution of correlation parameters and the prior distribution of regression parameters,
4. estimate with it the posterior distribution of the regression parameters β_a,
5. estimate with all of them the joint posterior distribution of $S(\vec{x})$ and $Y(\vec{x})$ including the locations to be predicted.

The objective in each one of these steps is estimated by simulating a large sample of the known distributions, using standard simulation techniques—Metropolis-Hastings algorithms —and the estimate is obtained as an average of the simulations, following the standard Monte Carlo procedure. The final result is a set of posterior distributions for each one of the parameters of the model, and the predictive distribution at each one of the desired locations.

These kind of models are highly elastic, since they can handle many different probability distributions (not only Gaussian-related ones) and several parameters with nonlinear relations (like the link function $h(\cdot)$) with the data. Clifford [1998] poses nevertheless some criticism to MCMC techniques, regarding the objectivity and reproductivity of obtained results. Other model-based bayesian estimation techniques can be found on the literature applied to spatial hazard problems, *e.g.*[Besag et al., 1991].

3.4.3 Bayesian/maximum entropy geostatistics

The primary interest of bayesian/maximum entropy methods, or BME [Christakos, 1990], is the estimation of the distribution of a RF $Z(\vec{x}_0)$ at unsampled locations given the observed sample $z_1, z_2, \ldots, z_N$, where for short $z_n = z(\vec{x}_n)$, and some generic objective constraints, *e.g.*fixed means, covariances, plausible intervals, quantiles, or any other information on the values of Z or its probability.

Consider the joint distribution $f(Z_0, Z_1, Z_2, \ldots, Z_N)$ of $\{Z(\vec{x}_0), Z(\vec{x}_1), \ldots, Z(\vec{x}_N)\}$ to be known. Then, given the observed sample, the RF has a distribution at the

unsampled location equal to

$$f(Z_0|z_1, z_2, \ldots, z_N) = \frac{f(Z_0, z_1, z_2, \ldots, z_N)}{f(z_1, z_2, \ldots, z_N)} \propto f(Z_0, z_1, z_2, \ldots, z_N)$$

due to the definition of conditional probability. Note that uppercase characters indicate a free-varying random variable, while lowercase characters are observed samples, *i.e.*fixed numbers. In words, if the joint distribution was known, the BME would deliver simply the conditional distribution at the unsampled location given the observed sample.

Now, the BME approach does not a priori assume any model for this joint distribution, as does the model-based technique of the last section and all Gaussian-related kriging techniques. Instead, it takes as joint distribution for the $\{Z_n\}$ the *less informative* existing distribution among those which satisfy a series of previously-known constraints. From the existing measures of information, Christakos [1990] chooses to use Shannon [1948] Entropy. Therefore, the joint distribution will be provided by Boltzmann's Theorem , which ensures that the maximum-entropy density is the exponential of a linear combination of the constraints [see, *e.g.*Leonard and Hsu, 1999, p. 122, for a complete account]. As a result, the distribution obtained with BME estimation is always from an exponential family, which gives the method some good analytical and numerical properties, *e.g.*to use MCMC methods, when no closed analytical form is available.

BME methods are extremely flexible, since they can incorporate any kind of objective information in the estimation procedure. As particular cases, if the available information is the mean and the covariance structure of the RF, then the joint distribution is a multivariate normal one, and the BME predictor coincides with simple kriging delivering the conditional Gaussian expectation. If the available information is given by the assumptions explained in the last section, then BME model coincides with the generalized linear geostatistical model. Finally, an example on the handling of categorical variables through BME methods is addressed in section 7.1. Other examples can be found in the monograph on the subject by Christakos [2000].

On the side of the flaws, BME method lacks simplicity. Its implementation usually is analytically untractable—in fact, the interesting cases of BME are those which do no evolve into closed analytical forms—, and most usually it relies on extensive computing techniques.

3.5 Change-of-support problems

Consider in this section only an univariate RF $Z(\vec{x})$ defined on the whole real line $\mathbb{R}$: Z will represent both the element of $\mathbb{R}$ as a vector and its coordinate in the canonical basis of $\mathbb{R}$, as was defined in section 2.5.1. Recall that this chapter began with the assumption that the RF $Z(\vec{x}) \in \mathbb{R}$ was defined on a *point-support* $\vec{x} \in \mathcal{D} \in R^p$. This is obviously an unrealistic assumption, since any physical property must be measured in a

given amount of time-space or matter, what it is called a *block-support*. Let $Z_v(\vec{x}) \in \mathbb{R}$ be then another RF defined on block-supports of volume v, which centers are located at $\vec{x} \in \mathcal{D}$. Now geostatistics can offer an answer to the following questions:

- which is the relationship between both RFs, on point and block-support?
- how can we *predict* the RF on block support by using measurements regarded as measured on point-support?
- which is the uncertainty linked to such a prediction? how can we compute the distribution of the RF on the block-support?

Obviously, this difference between point and block-support admits more than one single level: it is also possible to work with a series of nested volumes, like $\vec{x} \in v \subset V \subset W \subset \mathcal{D}$, where even the domain is finally considered as a block. This situation would arise when several possible risk management policies were linked to different volumes, *e.g.*water control in a waste-water-treating plant which needs to monitor that the average of Ammonia of the effluents in half an hour do never exceed a certain threshold, and at the same time that when it exceeds another threshold during more than ten minutes an alarm should be triggered.

3.5.1 Relationship between point and block-support RFs

If the point-support RF is known, then the values of any block-support RF can be computed.

Definition 3.15 (Sampling function) *A* sampling function *is a function* $p(\vec{x}) \in \mathbb{R}_+$*, with positive real images, defined on the whole domain* $\vec{x} \in \mathcal{D}$*, such that* $\int_{\mathcal{D}} p(\vec{x})d\vec{x} = 1$*.*

A sampling function represents an averaging process: it gives a weight to each point in the domain, so that all weights are positive and sum up to one. For instance, if

$$p_v(\vec{x}) = \begin{cases} \frac{1}{v}, & \vec{x} \in v \\ 0, & \text{otherwise,} \end{cases} \tag{3.30}$$

then $Z_v(\vec{x})$ is simply the arithmetic average of the point-support RF inside the block.

Definition 3.16 (Regularized RF) *Let* $Z(\vec{x})$ *be a point-support RF. Then the block-support RF can be computed as the convolution*

$$Z_v(\vec{x}) = \int p_v(\vec{h})Z(\vec{x} + \vec{h})d\vec{h}, \tag{3.31}$$

where the sampling function $p(\vec{h})$ *represents the averaging process.*

Definition 3.17 (Regularized structural functions) *Let $Z(\vec{x})$ be a point-support RF characterized by a covariance function $C(\vec{h})$. Then the block-support RF of definition 3.16 is characterized by the covariance function*

$$C_v(\vec{h}) = \int P_v(\vec{x}) C(\vec{h} + \vec{x}) d\vec{x}, \tag{3.32}$$

where P_v is the auto-convolution of p_v by itself

$$P_v(\vec{x}) = \int_{R^p} p_v(\vec{x}) p_v(\vec{x} + \vec{h}) d\vec{h}.$$

It is also possible to compute the variogram of the block-support RF by using expression (3.8) with block covariance (3.32). These covariance and variogram functions linked to the block-support RF are called regularized.

The regularized and point-support versions of these structural functions keep the following relationships

1. the shape of the functions (either the variogram or the covariance) is more or less the same,

2. however, around the origin, the regularized function is more likely to exhibit a parabolic behavior, which implies that the regularized RF is more *regular* than its point version,

3. and for second-order stationary RFs, the sill

$$C_v(0) = \mathrm{Var}\left[Z_v\right] = \int \int p(\vec{h}_1) C(\vec{h}_1 - \vec{h}_2) p(\vec{h}_2) d\vec{h}_1 d\vec{h}_2 < C(0) \tag{3.33}$$

of the regularized version is smaller or equal to the original sill $C(0) = \mathrm{Var}\left[Z(\vec{x})\right]$. The difference between the point-support and the regularized sills coincides with the *dispersion variance* of $\vec{x}$ in v.

Definition 3.18 (Dispersion variance) *The* dispersion variance *of a small block v partitioning a bigger one V in an intrinsic RF is defined as*

$$\sigma^2(v|V) = \int \int p_V(\vec{h}_1) \gamma(\vec{h}_1 - \vec{h}_2) p_V(\vec{h}_2) d\vec{h}_1 d\vec{h}_2 - \int \int p_v(\vec{h}_1) \gamma(\vec{h}_1 - \vec{h}_2) p_v(\vec{h}_2) d\vec{h}_1 d\vec{h}_2,$$

where $\gamma(\cdot)$ represents the point-support variogram. The dispersion variance is interpreted as the error variance affecting the value of Z_V in the bigger support when estimated using the value Z_v in the smaller support, or equivalently as the variance of Z_v in V.

Using the weighting function of equation (3.30), the dispersion variance becomes

$$\begin{aligned}\sigma^2(v|V) &= \frac{1}{V^2}\int_V\int_V \gamma(\vec{h}_1-\vec{h}_2)d\vec{h}_1d\vec{h}_2 - \frac{1}{v^2}\int_v\int_v \gamma(\vec{h}_1-\vec{h}_2)d\vec{h}_1d\vec{h}_2 = \\ &= \frac{1}{v^2}\int_v\int_v C(\vec{h}_1-\vec{h}_2)d\vec{h}_1d\vec{h}_2 - \frac{1}{V^2}\int_V\int_V C(\vec{h}_1-\vec{h}_2)d\vec{h}_1d\vec{h}_2,\end{aligned}$$

where the second equality only holds for second-order RFs.

Property 3.3 (Krige's relationship) *Dispersion variances satisfy the following additivity property for nested blocks,*

$$\sigma^2(v|W) = \sigma^2(v|V) + \sigma^2(V|W).$$

Note that, taking $v = \vec{x}$ and $W = \mathcal{D}$, we find: a) $\sigma^2(\vec{x}|\mathcal{D})$ is the variance of the point-support RF in the whole domain, its total variance, or the sill of its variogram; b) $\sigma^2(V|\mathcal{D})$ is equivalently the sill of the regularized variogram; and, c) the difference is the dispersion variance of the point-support $\vec{x}$ in the volume V.

3.5.2 Universal block kriging

A classical change-of-support problem is the estimation of an average (3.31) of a block v at $\vec{x}_0$ through a linear combination of point-support observations $\zeta_n = \zeta(\vec{x}_n), \vec{x}_n \in \mathcal{D}, n = 1, \ldots, N$. In general, this block v does not coincide with the regularization block explained in the last section, but will be bigger. Following step-by-step the case of point kriging (section 3.3) but for a single variable, the following elements are needed:

- the average value of the drift functions $f_a(\vec{x}_0 \in v)$ inside the block

$$f_a(v) = f_a(\vec{x}_0 \in v) = \int f_a(\vec{x}_0+\vec{h})p_v(\vec{h})d\vec{h},$$

- the covariance between each sample $\vec{x}_n$ and the block

$$C(\vec{x}_n, v) = C(\vec{x}_n, \vec{x}_0 \in v) = \int C(\vec{x}_n, \vec{x}_0+\vec{h})p_v(\vec{h})d\vec{h},$$

- and the variance of the block (3.33), here denoted by $C(v, v)$.

Definition 3.19 (Universal block kriging predictor) *With these expressions, the univariate universal block kriging predictor is*

$$\zeta_0^* = \sum_{n=1}^{N}\lambda_n\zeta_n,$$

where the weights are obtained by minimizing the system

$$\sum_{m=1}^{N} \lambda_m C(\vec{x}_n, \vec{x}_m) + \sum_{a=0}^{A} \nu_a f_a(\vec{x}_m) = C(\vec{x}_n, v) \quad , \quad n = \{1, 2, \ldots, N\},$$

$$\sum_{n=1}^{N} \lambda_m f_a(\vec{x}_m) = f_a(v) \quad , \quad a = \{0, 1, \ldots, A\}.$$

Note that this system is exactly equivalent to (3.19) but with all the expressions involving $\vec{x}_0$ replaced by their block counterparts defined above. Regarding the kriging variance, it corresponds to

$$\sigma_K^2 = \sum_{n,m=1}^{N} \lambda_n \lambda_m C(\vec{x}_n, \vec{x}_m) - 2 \sum_{n=1}^{N} \lambda_n C(\vec{x}_n, v) + C(v, v).$$

Equivalently, a block simple kriging predictor can be defined by adapting expression (3.24). Finally, it is worth summarizing the main properties of these simple and universal block kriging predictors.

Property 3.4 (Optimality of block kriging predictors) *The simple kriging (SK) and universal kriging (UK) predictors satisfy for block RFs the following properties:*

1. *the SK predictor is unbiased and of minimal variance among the predictors expressed as linear combinations;*

2. *in a Gaussian RF, the true block average value is also normally distributed, with the SK estimator and its kriging variance as the parameters of the conditional distribution, like in expression (3.26);*

3. *the UK predictor is also a best linear unbiased predictor, which takes into account the shape of the drift; however, it provides no longer the conditional distribution of the true block average value, neither in the Gaussian function case;*

4. *again, the departure of the block UK predictor from the distribution of the true block average will depend on the relation (3.28), like in point-support RFs.*

Block kriging is particularly well-suited for Gaussian RFs, because if a point-support RF is a Gaussian one, then any of the regularized block-support RFs one could define on it is also a Gaussian one. This justifies the fact that block simple kriging estimates the conditional distribution.

3.5.3 Global change-of-support

In this section, we look for a way to describe the variability of the mean value of the RF $Z(\vec{x})$ on blocks of size v partitioning the whole domain $\mathcal{D}$. This is achieved by

giving the probability distribution function for the value of the mean in a block taken at random from all the blocks in $\mathcal{D}$. However, some assumptions must be done on the properties of the RFs. These assumptions are called a *global change-of-support*.

In the last section, we introduced a known way to compute the mean of a RF for a block-support *conditional* to the data around and inside it: the kriging predictor. In the case of a Gaussian RF, this predictor and its variance yield also the distribution of the mean of the RF inside the block. Thus, block kriging with all the available data constitutes a global change-of-support model. As we will see, this coincides with the so-called *multi-Gaussian model* [Verly, 1983] for a very specific case.

Other change-of-support models (either conditional to the data set, or without conditioning) have been developed for cases where a Gaussian assumption was not allowed. These models relate the distribution of a point-support ($\vec{x}$) RFs with that of a block-support (V) RF, or the distributions associated with two different block-supports ($v \subset V$). The main condition expected for a change-of-support model is that the distribution of bigger blocks should be less *selective* than that of smaller blocks.

To understand the concept of selectivity, it is useful to recall its origin. When mining a mineral deposit, the rock is cut in blocks, which are sent either to the mill (as *ore*) or considered as *waste* using an estimated average of its content in metal (the so-called *grade*). This implies that a certain amount of metal is not processed, because it was in a block considered as waste. To avoid that, we should exploit the deposit in blocks of the smallest volume possible. In other words, the small blocks are considered more *selective* than the bigger ones, because if we could use them we would better select ore from waste blocks.

Chilès and Delfiner [1999] present some of the classical change-of-support models, as well as its commonly recognized limitations:

Definition 3.20 (Affine correction) *Let $Z(\vec{x}), Z_v(\vec{x})$ be a pair of point- and block-support Gaussian RFs; then they satisfy*

$$\frac{Z(\vec{x}) - \mu}{\sigma} \sim \frac{Z_v(\vec{x}) - \mu}{\sigma_v} \sim \mathcal{N}(0, 1). \tag{3.34}$$

The affine correction assumes this identity of the distributions even in the case of a non-Gaussian RF.

In practical cases, the distribution of the block-support is derived from the experimental cumulative histogram of point-supports, whereas means and variances are obtained from a classical variography analysis.

Definition 3.21 (Discrete Gaussian model) *Let $\phi(\cdot)$ be a transformation, such that the transformed point and block-support RFs form a bi-normally distributed pair, with a regression coefficient r; the distribution $\phi_v(z_v)$ of the block value z_v is computed by using*

$$\phi_v(z_v) = \mathrm{E}\left[\phi\left(r \cdot z_v + u \cdot \sqrt{1 - r^2}\right)\right], \tag{3.35}$$

where u is an auxiliary standard normal variable.

In applications, one must estimate the correlation coefficient. This is selected so that the variance of the block-support satisfies condition (3.33).

Definition 3.22 (Multi-Gaussian model) *Assume that $\phi(\cdot)$ transforms the point-support RF $Z(\vec{x})$ into a Gaussian RF $\zeta(\vec{x}) = \phi\left(Z(\vec{x})\right)$. Then the transformation of the RF at any set of locations will follow a multi-Gaussian distribution. Thus, block kriging will yield the distribution of the block-support RF. This distribution can be back-transformed through $\phi^{-1}(\cdot)$ to obtain the distribution of the block-support $Z(\vec{x})$ [Verly, 1983].*

These methods rely on different hypothesis, but when assuming joint Gaussianity of the RF they coincide: in this case, the transformations involved in the last definitions are the identity $\phi(z) = z$, and the discrete Gaussian model coincides with the multi-Gaussian and with the affine correction.

Note that if they are not Gaussian, all this models are approximations, which should be carefully used to avoid inconsistent results. The nowadays usefulness of these change-of-support models is decreasing, as computers are more powerful and large simulations are possible. These models were defined to compensate the limitations of computers with approximate analytical expressions drawn from theoretical assumptions. Their current interest may be instead focused in understanding the processes underlying the RF.

One of the usual ways to express the final estimated distributions is through the *selectivity curves*, which are almost always applied to strictly positive variables. Thus, they are further discussed in section 5.4.

3.6 Case study: conductivity

This section illustrates the methods of this chapter using the example already presented in section 2.5.1. This data set is a series of measurements of water conductivity obtained automatically by an online control station at different time moments. Figure 3.4 shows its time evolution during the years 2002-2003, as well as the evolution of water temperature. Figure 3.5 shows two details of this evolution, during July 2002 and July 2003. The presence of a drift in both series, specially in temperature, is self-evident, as is its connection with daily and yearly periods. To confirm this, the frequency spectrum of water temperature was studied: it is displayed in figure 3.6. Computation was done using function `fft` from the statistical software `R` [R Development Core Team, 2004].

This suggest the presence of two main contributions, the 24-hour period and the 1-year period drifts, as well as other complementary wave drifts of approximate periods of 2.5 days, 10 days, 25 days (a month), 42 days, 100 days (a season) and a year. The degree to which this simplification captures the variability of water temperature may be visually assessed by the scatter-plot of figure 3.7 Given the dynamic link assumed

to exist between solar radiation, water temperature and conductivity (section 1.4.1), the drift functions used in the kriging techniques are

$$f_0(t) = 1,\ f_{2i-1}(t) = \cos\left(\frac{2\pi t}{\tau_i}\right),\ f_{2i}(t) = \sin\left(\frac{2\pi t}{\tau_i}\right), \tag{3.36}$$

with the periods included in table 3.1. Note that not all periods from figure 3.6 were used, due to reasons explained below.

Table 3.1: Periods τ (in days) of the trigonometric drift functions considered in equation (3.36), extracted from figure 3.6.

i	1	2	3	4
τ_i	1	2.5	10	25

Here, the spatial dependence vector $\vec{x}$ has been replaced by a scalar time dependence t, which simplifies the notation. As an example, a separate study of the two months of July 2002 and July 2003 is conducted. This is the reason why periods higher than the month have not been considered in this analysis. Classical regression of conductivity data set during each one of these two months against the drift functions, for instance using expression (3.20) with uncorrelated residuals, gives a set of coefficients summarized in table 3.2. However, the assumption of uncorrelated residuals is verified to be false after looking at their estimated auto-covariance function (3.4), displayed in figure 3.8. This shows also the covariance function of the original conductivity data sets. Comparing them, one can visually assess the degree to which the drift has been successfully removed by this regression fitting: specially visible in July 2002, the 24-hour drift period has been mostly removed from the covariance, once the drift accounts for it.

Note that self-correlation of residuals makes regression an invalid technique. In particular, residuals have no longer zero mean in all the sampled period, nor are they homoscedastic. However, the regression functions are trigonometric functions of time, and the sampled time is relatively long, with a high and quite regular sampling density. These reasons allow us to assume the regression residuals to be practically homoscedastic and have zero mean. Thus, one has ground to interpret them as a second-order stationary RF, denoted as $Z(t)$. Its covariance is displayed in figure 3.8.

We apply afterwards universal kriging to the original data set, using the covariance function of the residuals as a structural tool. The final drift estimates are listed in table 3.2, while figure 3.9 shows also the final kriging auto-covariance of the residuals. Its experimental auto-covariance is represented in figure 3.9. The fitted covariance model is

$$C(h) = (8 \cdot \mathrm{Sph}(h|a = 0.25) + 6 \cdot \mathrm{Exp}(h|a = 2.5) + 3 \cdot \mathrm{Hole}(h|a_t = 4, a_d = 8)) \cdot 10^3, \tag{3.37}$$

with all ranges (a) and time (t) expressed in days. Recall that these models are defined in equations (3.9-3.11).

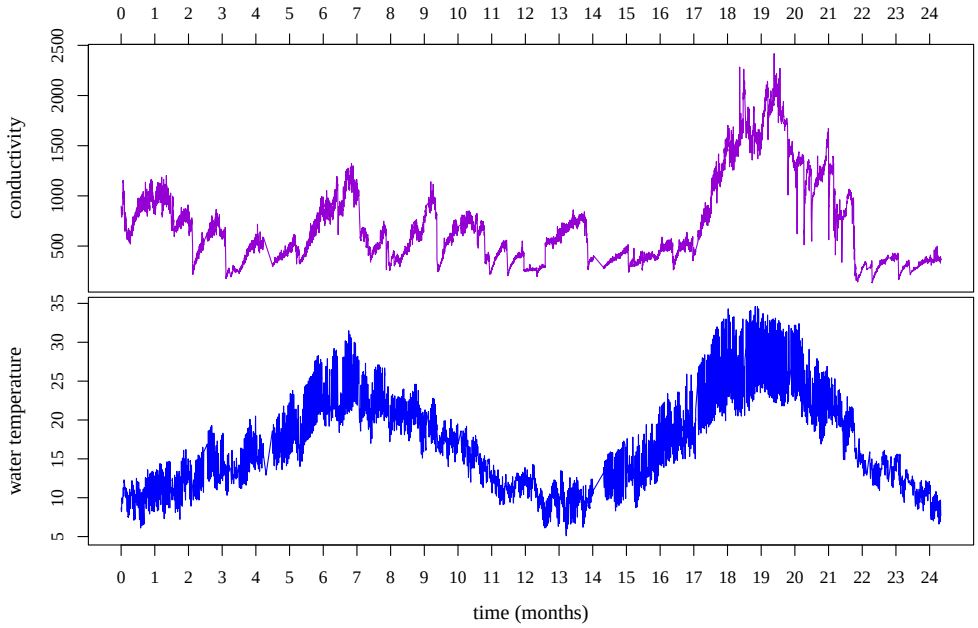

Figure 3.4: Time evolution of conductivity in $\mu S/cm$ (top), and water temperature in ^oC (bottom), during the years 2002-2003 at the Gualba station.

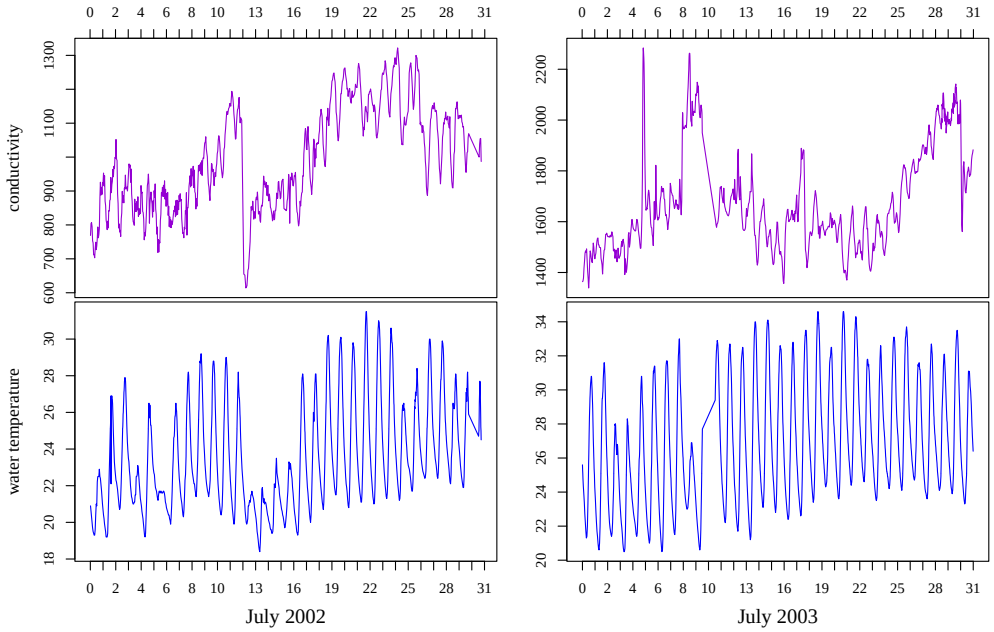

Figure 3.5: Time evolution of conductivity in $\mu S/cm$ (top), and water temperature in ^oC (bottom), during the months of July 2002 (left) and July 2003 (right) at the Gualba station. Note the different vertical scales.

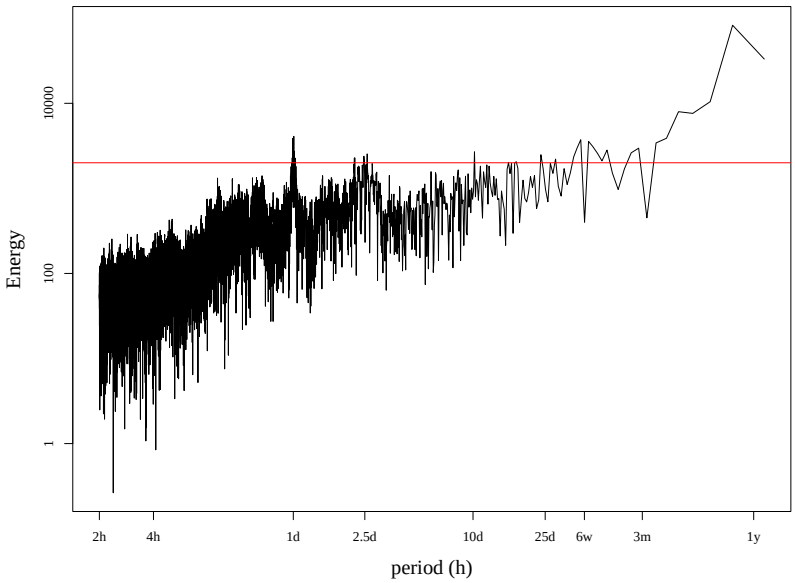

Figure 3.6: Frequency spectrum of water temperature, at Gualba station. Some clusters of high energy in high periods are detected, from which some representative periods were selected: those of 1, 2.5, 10, 25, 42, 100 and 365 days.

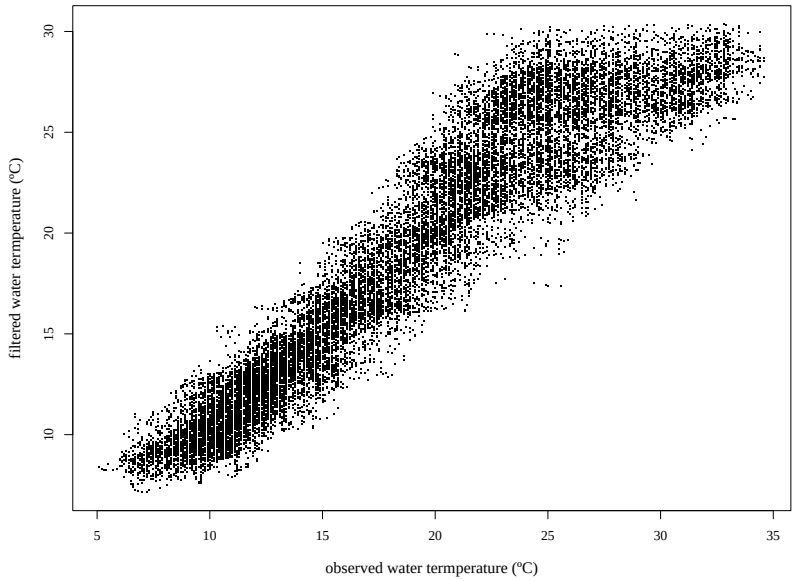

Figure 3.7: Scatter plot of observed water temperature against its regression prediction using the functions of equation (3.36).

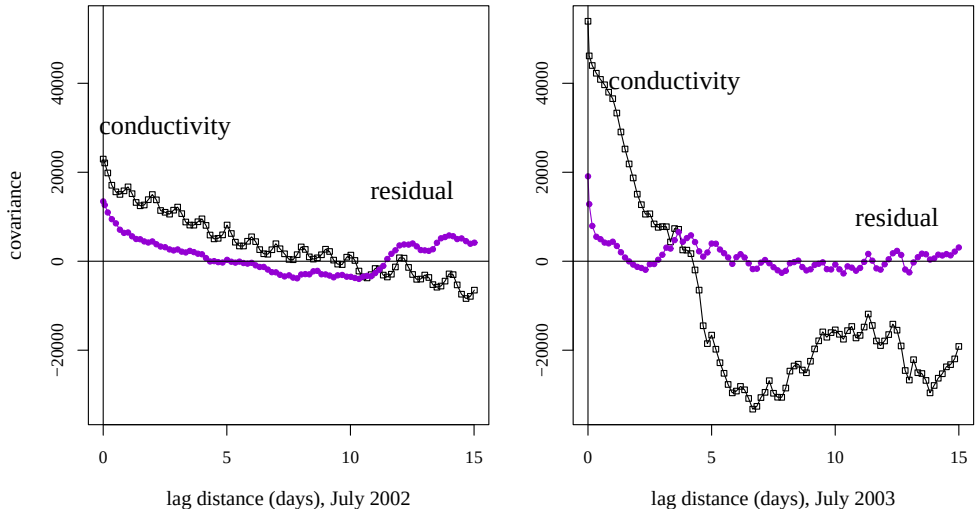

Figure 3.8: Covariance function of conductivity original data set (black, squares), and residuals of regression (violet, dots), for July 2002 (left) and July 2003 (right), at the Gualba station. The means for these covariances has been fixed as the constant value of the drift a_0 and zero respectively. Although the variability of both original series was strongly different, their de-trended series present a more similar behavior. Note that regression has removed an important part of the 24h-period.

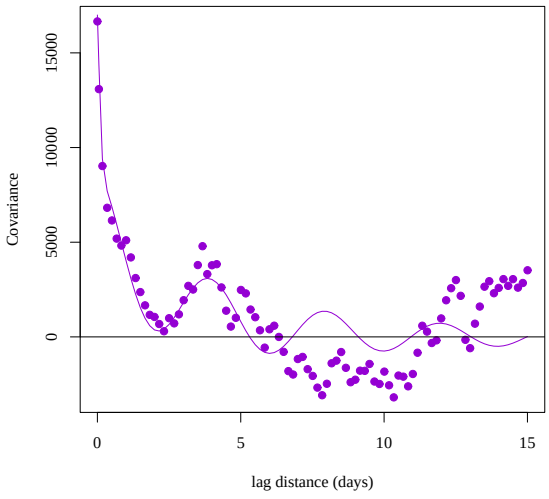

Figure 3.9: Covariance function of conductivity residuals.

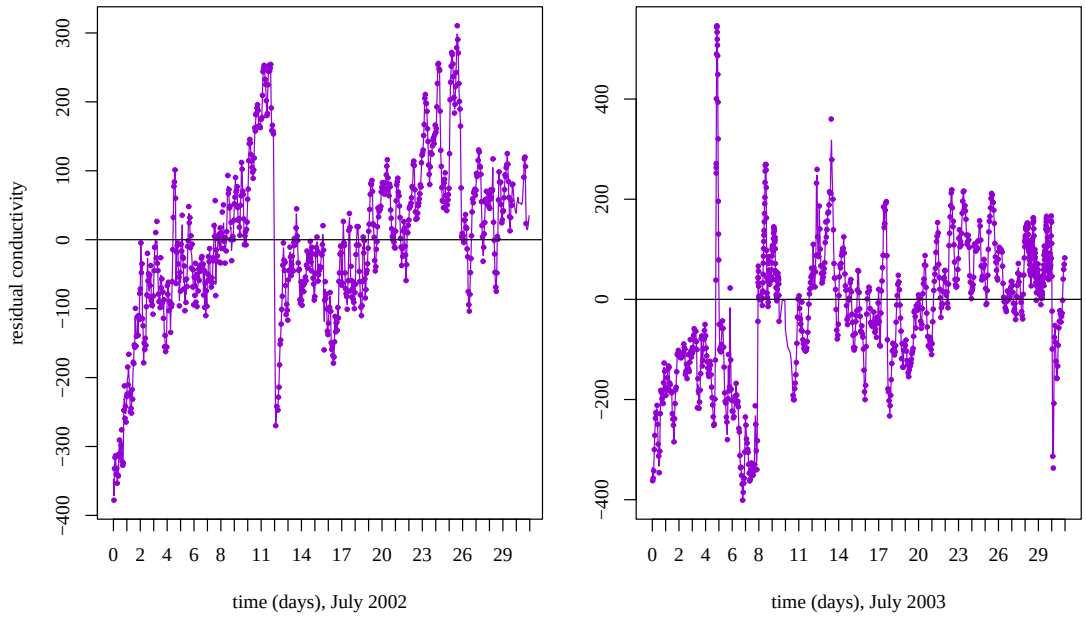

Figure 3.10: Time evolution of residual conductivity: residual data set (dots) and estimates (continuous line) for both July 2002 and 2003 series.

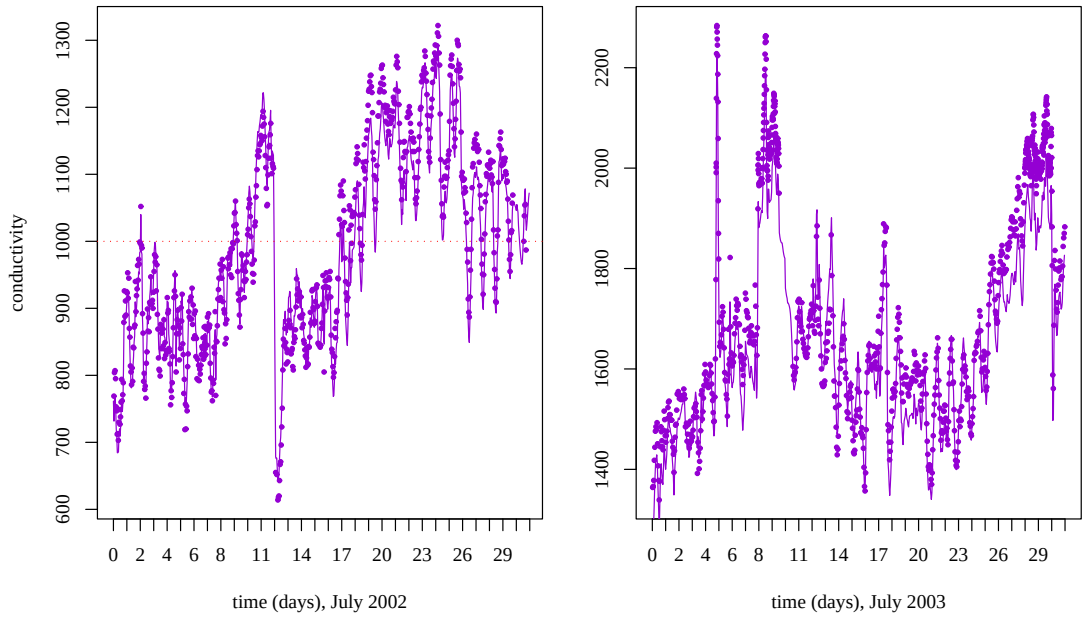

Figure 3.11: Time evolution of conductivity: data set (dots) and final estimates (continuous line).

Table 3.2: Fitted drift coefficients for classical regression (left) and kriging of the drift (right), expressed both as coefficients for sine and cosine wave functions, and as amplitude ($m_{2i} = \sqrt{a_{2i-1}^2 + a_{2i}^2}$) and phase ($\alpha_{2i} = \frac{\tau_i}{2\pi} \arctan \frac{a_{2i-1}}{a_{2i}}$, in days) of a cosine. The upper table contains the coefficients fitted to July 2002 series, and the lower table those for July 2003.

	Standard regression				Kriging of the drift			
i	a_{2i-1}	a_{2i}	m_{2i}	α_{2i}	a_{2i-1}	a_{2i}	m_{2i}	α_{2i}
0	0	994.24	994.24	0	0	982.73	982.73	0
1	52.14	-20.33	55.97	-0.19	50.96	-17.59	53.91	-0.20
2	20.01	-11.66	23.16	-0.42	15.90	-20.84	26.21	-0.26
3	48.50	14.70	50.68	2.03	31.37	26.77	41.24	1.38
4	97.07	77.04	123.93	3.58	67.02	55.24	86.85	3.51
0	0	1688.01	1688.01	0	0	1656.97	1656.97	0
1	-11.41	7.43	13.62	-0.16	-19.74	22.75	30.12	-0.11
2	19.63	32.13	37.65	0.22	18.38	20.31	27.39	0.29
3	-149.18	62.59	161.78	-1.87	-95.67	45.28	105.85	-1.80
4	120.03	117.89	168.24	3.16	94.26	131.02	161.41	2.48

Simple kriging, as explained in section 3.3.3, is the best option to interpolate a RF with known mean. To ensure that residuals and drifts coefficients have been satisfactorily estimated, simple kriging is conducted on the regression residuals using the experimental covariance (3.37). Figure 3.10 shows the results of this simple kriging, jointly with the observed residual values, obtained with both `GSLIB` programs [Deutsch and Journel, 1992], and `predict.gstat` from package `gstat` [Pebesma and Wesseling, 1998] for `R`.

Therefore, the parameters of this global model have been estimated satisfactorily, and both the trigonometric drift and the auto-correlation of the residuals are taken into account. Figure 3.11 represents this interpolation, jointly with the original conductivity data set.

By assuming expression (3.26) for these last kriging predictions, one can compute the hazard probability of exceeding the thresholds of $1000\mu S/cm$ and $2500\mu S/cm$ at each moment (see table 1.1). These probabilities may be taken as indicators of the water quality, the lower the probability the higher the quality. Figure 3.12 clearly shows the strong influence of the drift on the water quality, specially of the 60hour period. In spite of these fluctuations, conductivity measures were almost surely between 1000 and $2500\mu S/cm$ during July 2003, whereas in 2002 they were approximately half of the time above $1000\mu S/cm$, and the other half below this threshold.

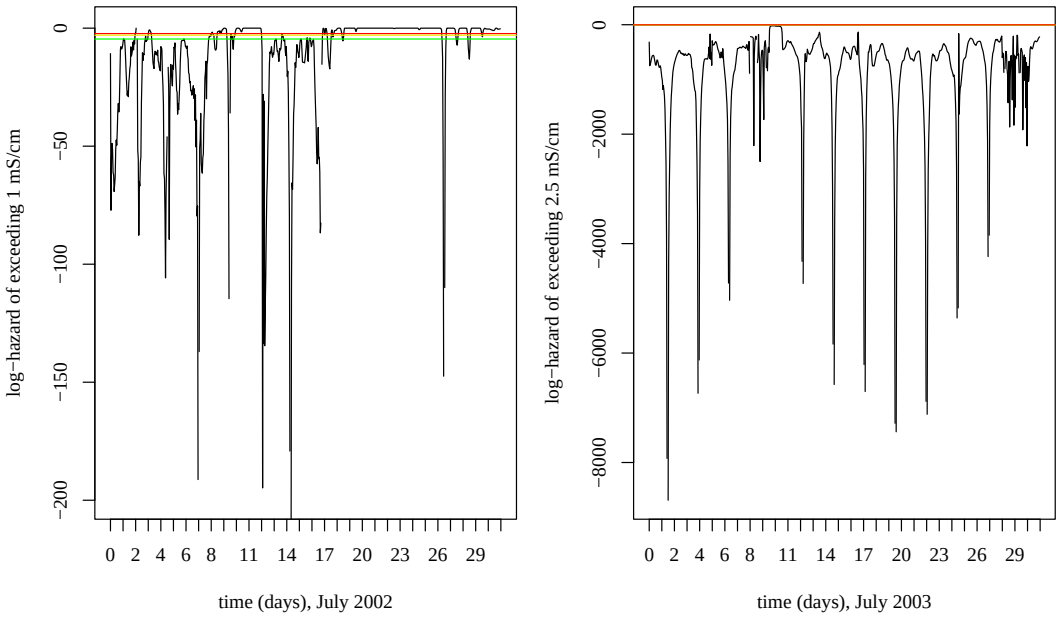

Figure 3.12: Time evolution of the hazard of exceeding $1000\mu S/cm$ of conductivity in the year 2002, and of $2500\mu S/cm$ in the year 2003. Red line marks the level of 0.10, yellow the 0.05 and green the 0.01 of probability.

3.7 Remarks

In this chapter some techniques for dependent data were summarized. Most of them were developed primarily by Matheron [1965], or they were based on this seminal work on geostatistics of RFs. All these prediction techniques yield particularly well-suited results when they work on real Gaussian RFs. Actually, the simple kriging predictor and its kriging variance were the estimates of the mean and the variance of the Gaussian distribution at unsampled locations, conditional to the observed data set. When the RF was not a Gaussian one, some techniques are useful, which essentially tried to transform the function to another one which was assumed to be Gaussian. This issue will be further developed in chapter 5.

The Gaussian assumption allows the computation of hazard probabilities of exceeding certain thresholds. Here, a case study on a time-dependent conductivity data set yielded a series of hazard estimates, which showed a strong oscillation. This hazard probability could be taken as a water quality index, based only on conductivity. However, in both studied cases these variations were not able to change the state of the river: during July 2002 conductivity was always below the quality threshold of $2500\mu S/cm$, but occasionally it was above the $1000\mu S/cm$ threshold. Contrarily, during July 2003 (a very dry year) conductivity measurements were always above $1000\mu S/cm$, but seldom above $2500\mu S/cm$ too. In management terms, water in this river during July

2002 might enter categories 1 (quality-demanding uses) to 3 (restricted uses), whereas in July 2003 it was clearly inside category 4 (minimal uses). Another issue is the clear 2.5 day-period oscillation detected in both series. This oscillation might be connected to the solar energy supply on the river, but this will be further discussed in chapter 8.

3.8 Addendum: validity of change-of-support models

Here we present the conditions that render a valid change-of-support model, according to Chilès and Delfiner [1999]. This material complements the exposition of section 3.5.3, and is again used in section 5.4, devoted to discuss the case of lognormal variables.

Property 3.5 (Conditions on a valid change-of-support model) *A valid change-of-support model linking a point-support RF with a block-support RF must satisfy the following conditions:*

1. *they have the same mean,*

$$\mathrm{E}\left[Z(\vec{x})\right] = \mathrm{E}\left[Z_v(\vec{x})\right] = \mathrm{E}\left[Z_V(\vec{x})\right] = m$$

2. *the variance of the block RF is linked through (3.33) to the covariance function of the point random-function,*

3. *the distribution of the bigger blocks is less selective than that of the smaller ones, which is in turn less selective that the distribution of the point-support RF (Cartier's relation).*

According to Chilès and Delfiner [1999], the concept of selectivity can be mathematically formulated in the following form.

Definition 3.23 (Selectivity) *A cumulative distribution $F_1(z)$ is more* selective *than another $F_2(z)$ if there exists a bivariate distribution $F_{Z_1Z_2}(z_1, z_2)$ with marginals F_1 and F_2 such that*

$$\mathrm{E}\left[Z_1|Z_2\right] = Z_2, \tag{3.38}$$

which is equivalent to say that the regression line of Z_1 on Z_2 is the identity line.

Following [Emery, 2004], *Cartier's relation* (property 3.5.3) is by far the most restrictive condition on the change-of-support models, and it actually entails *per se* all the three conditions of property 3.5.

Chapter 4

Geostatistics in an arbitrary Euclidean space

This chapter is the central part of this work. In the last one, we summarized the general geostatistical tools and techniques, essentially developed for Gaussian, thus real random functions (RF). In this chapter we generalize them by interpreting these classical definitions *as if they were applied to* the coordinates of a vectorial RF, what is called *principle of working on coordinates*[Pawlowsky-Glahn, 2003]. Then, we re-define all geostatistical concepts, estimators and its properties in terms of vectors and operators in an Euclidean space. Finally, we confirm that they are consistent with the principle of working on coordinates. In this way, we ensure that choosing a coordinate system does not affect our results.

4.1 Notation

Recall of Definition 2.6 (Euclidean space) *A set $\mathbb{E}$ is a D-dimensional Euclidean space if it is a D-dimensional real vector space equipped with a suitable scalar product. It is usually denoted by $\{\mathbb{E}, \oplus, \odot, \langle \cdot, \cdot \rangle_{\mathbb{E}}\}$.*

The following notation will be used. The vectors of the space $\mathbf{a} \in \mathbb{E}$ will be denoted by lowercase boldface Latin characters, and the scalars by lowercase Greek characters; the Abelian group operation, or sum, will be $\oplus$, and the external product $\odot$; the neutral element with respect to this sum is denoted with $\mathbf{n}$, and the inverse element of $\mathbf{a}$ by $\ominus\mathbf{a}$, where $\ominus$ is the inverse operation to sum, thus satisfying $\mathbf{a} \ominus \mathbf{a} = \mathbf{n}$; the scalar product will be $\langle \cdot, \cdot \rangle_{\mathbb{E}}$, and its associated norm and distance $\| \cdot \|_{\mathbb{E}}$ and $d(\cdot, \cdot)$, respectively. A basis of this space will be denoted by $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_D\}$, and the coordinates of any vector $\mathbf{a}$ with respect to it by using the equivalent lowercase Greek character $\underline{\alpha}$, where the underline indicates that it is a vector of $\mathbb{R}^D$. Slightly *forcing* the notation, we use also an expression like $\mathbf{a} = \underline{\alpha} \odot \mathbf{E}$ to say that we apply the coordinates to the basis and recover the original vector. Note also that throughout the next chapters, the neutral element of this Euclidean structure (replacing the zero vector $\underline{0}$) is denoted by $\mathbf{n}$.

Any linear application $T(\cdot) : \mathbb{E} \to \mathbb{F}$ acting on a vector $\mathbf{a}$ will be written as $T\mathbf{a} = T(\mathbf{a})$. Note that this application may have different image and origin spaces. Given basis $\mathbf{E}$ of $\mathbb{E}$ and $\mathbf{F}$ of $\mathbb{F}$, these applications are fully characterized by a matrix $\underline{\underline{T}}$, where the j-th column contains the coordinates of the image of $T(\mathbf{e}_j)$ in the basis $\mathbf{F}$.

Finally, if $\mathbf{E}$ and $\mathbf{F}$ are two basis of $\mathbb{E}$, there exists a matrix $\underline{\underline{\varphi}}$ containing in each column the coordinates of an element of $\mathbf{F}$ with respect to the basis $\mathbf{E}$. The inverse $\underline{\underline{\varphi}}^{-1}$ contains by columns the coordinates of an element of $\mathbf{E}$ with respect to the basis $\mathbf{F}$. It these two basis are orthonormal ones, then the matrix satisfies $\underline{\underline{\varphi}}^{-1} = \underline{\underline{\varphi}}^{t}$ and it contains in the columns the coordinates of each element of $\mathbf{E}$ with respect to the basis $\mathbf{F}$, and in the rows the coordinates of the elements of the basis $\mathbf{E}$ with respect to the basis $\mathbf{F}$.

4.2 Random function

Let $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$ be a point (or the center of a block v) in a domain $\mathcal{D}$ of the space-time real space, with $p \in \{1, 2, 3, 4\}$. Let $\mathbf{Z}(\vec{x}) \in \mathbb{E}$ be a vector-valued RF, and $\underline{Z}(\vec{x}) \in \mathbb{R}^D$ be the coordinates of $\mathbf{Z}$ with respect to a given basis $\mathbf{E}$. Let $\mathbf{z}(\vec{x}_1), \mathbf{z}(\vec{x}_2), \ldots, \mathbf{z}(\vec{x}_N)$ be an observed sample of this vector-valued RF, and $\underline{\zeta}(\vec{x}_1), \underline{\zeta}(\vec{x}_2), \ldots, \underline{\zeta}(\vec{x}_N)$ the coordinates of this sample. The goal of this chapter will be the prediction of the RF $\mathbf{Z}(\vec{x}_0)$ at an unsampled location $\vec{x}_0$, and of its error variance-covariance matrix.

Definition 4.1 (Stationarity in $\mathbb{E}$) *Let $\boldsymbol{Z}(\vec{x})$ be a RF with domain $\mathcal{D} \subset \mathbb{R}^p$ and image $\mathbb{E}$. Then it is called*

1. strongly stationary*, when for any set of $B_n \subset \mathbb{E}$ and for all set of locations $\{\vec{x}_n\} \in \mathcal{D}$, the following probability is invariant by translation $\vec{h}$:*

$$\Pr\left[(\boldsymbol{Z}_1(\vec{x}+\vec{h}) \in B_1) \cap (\boldsymbol{Z}_2(\vec{x}+\vec{h}) \in B_2) \cap \ldots \cap (\boldsymbol{Z}_N(\vec{x}+\vec{h}) \in B_N)\right] =$$

$$= \Pr\left[(\boldsymbol{Z}_1(\vec{x}) \in B_1) \cap (\boldsymbol{Z}_2(\vec{x}) \in B_2) \cap \ldots \cap (\boldsymbol{Z}_N(\vec{x}) \in B_N)\right];$$

2. second-order stationary*, when for any pair of locations $\vec{x}_n, \vec{x}_m \in \mathcal{D}$, the mean vector and the covariance operator are translation-invariant, or*

$$\mathrm{E}_{\mathbb{E}}\left[\boldsymbol{Z}(\vec{x}_n)\right] = \boldsymbol{\mu} \quad \text{and} \quad \mathrm{Cov}\left[\boldsymbol{Z}(\vec{x}_n), \boldsymbol{Z}(\vec{x}_m)\right] = C(\cdot; \vec{x}_m - \vec{x}_n);$$

3. intrinsic*, when for any pair $\vec{x}_n, \vec{x}_m \mathcal{D}$, the increments $(\boldsymbol{Z}(\vec{x}_m) \ominus \boldsymbol{Z}(\vec{x}_n))$ have neutral vector mean and stationary variance operator:*

$$\mathrm{E}_{\mathbb{E}}\left[\boldsymbol{Z}(\vec{x}_m) \ominus \boldsymbol{Z}(\vec{x}_n)\right] = \boldsymbol{n} \quad \text{and} \quad \mathrm{Var}_{\mathbb{E}}\left[\boldsymbol{Z}(\vec{x}_m) \ominus \boldsymbol{Z}(\vec{x}_n)\right] = \gamma(\cdot; \vec{x}_m - \vec{x}_n);$$

The covariance and variance operators are shown in this case with an argument and a parameter, *e.g.*$C(\cdot; \vec{x}_m - \vec{x}_n)$. Any possible vector on $\mathbb{E}$ is the argument instead of the dot, whereas the lag distance plays the role of a parameter.

Property 4.1 *A vector-valued RF* $\boldsymbol{Z}(\vec{x}) \in \mathbb{E}$ *is stationary (strong, second-order, or intrinsic sense) on* $\mathbb{E}$ *if its coordinates* $\underline{Z}(\vec{x}) \in \mathbb{R}^D$ *form a stationary RF (strong, second-order, or intrinsic sense).*

Proof immediately follows from the identification between probability laws for vectors and its coordinates (for strong stationarity), and the identification of mean vector and variance/covariance operators with their coordinate counterparts (for second-order or intrinsic stationarity). Furthermore, this identification is the conceptual proof of the following proposition.

Proposition 4.1 *A second-order stationary RF on* $\mathbb{E}$ *with respect to a basis is also second-order stationary with respect to any other basis of* $\mathbb{E}$*. An intrinsic RF* $\mathbb{E}$ *with respect to a basis satisfies also the intrinsic hypothesis with respect to any other basis of* $\mathbb{E}$*.*

Definition 4.2 (Gaussian RF on $\mathbb{E}$) *A vector-valued RF* $\boldsymbol{Z}(\vec{x}) \in \mathbb{E}$ *is said to be a Gaussian RF on* $\mathbb{E}$*, if for any testing vector* $\boldsymbol{z}_0 \in \mathbb{E}$*, the projection* $\langle \boldsymbol{z}_0, \boldsymbol{Z}(\vec{x}) \rangle_{\mathbb{E}}$ *form a real Gaussian RF.*

4.3 Structural analysis

This section focuses on the characterization of the structural functions according to definition 4.1: the operators of covariance $C(\cdot; \vec{x}_m - \vec{x}_n)$ and variogram $\gamma(\cdot; \vec{x}_m - \vec{x}_n)$. From now on, the dot in these expressions is dropped, so that,*e.g.* $C(\vec{x}_m - \vec{x}_n) \equiv C(\cdot; \vec{x}_m - \vec{x}_n)$. Here we show that for any given basis, the basic properties of the covariance and variogram expressed in coordinates are kept. In particular, the validity of a system (*i.e.*, its positive-definite character), the symmetry of a system and the global range of a system are treated.

Property 4.1 states that given the stationarity of the RF with respect to a basis, it is stationary with respect to any other basis of the space. Not only so, but also there is a linear relationship between the expectations or the covariance functions or variograms in the two basis,

$$\begin{aligned} \mathrm{E}\left[\underline{Z}_{\mathbf{E}}\right] &= \underline{\underline{\varphi}} \cdot \mathrm{E}\left[\underline{Z}_{\mathbf{F}}\right], \\ \underline{\underline{C}}_{\mathbf{E}}(\vec{h}) &= \underline{\underline{\varphi}} \cdot \underline{\underline{C}}_{\mathbf{F}}(\vec{h}) \cdot \underline{\underline{\varphi}}^t, \\ \underline{\underline{\gamma}}_{\mathbf{E}}(\vec{h}) &= \underline{\underline{\varphi}} \cdot \underline{\underline{\gamma}}_{\mathbf{F}}(\vec{h}) \cdot \underline{\underline{\varphi}}^t. \end{aligned} \tag{4.1}$$

It is worth noting that these properties are not restricted to orthonormal bases, and the following results will be valid for them.

Note that in expressions (4.1) and (4.1), the matrices $\underline{\underline{C}}$ and $\underline{\underline{\gamma}}$ contain respectively auto-covariance functions (definition 3.5) and direct variograms (definition 3.6) in the diagonal, whereas the off-diagonal terms are cross-covariance functions (definition 3.13)

and cross-variograms (definition 3.14), all of them defined on the coordinates with respect to a basis. In this context the matrix of coefficients of codispersion (3.15) at a lag distance $\vec{h}$ might be interpreted as the correlation matrix of the increments of a RF like $\mathbf{Z}(\vec{x}+\vec{h}) \ominus \mathbf{Z}(\vec{x})$.

Proposition 4.2 *If $\underline{\underline{C}}(\vec{h})$ forms a valid covariance system (*i.e.*with positive definite spectral densities, according to proposition 3.2) then $\underline{\underline{K}}(\vec{h}) = [\underline{\underline{\varphi}} \cdot \underline{\underline{C}}(\vec{h}) \cdot \underline{\underline{\varphi}}^t]$ is also valid.*

Proof: Consider an integral applied to a matrix of functions as the matrix of the integrals of each component of the matrix, $\int \underline{\underline{C}}(\vec{h}) d\vec{h} = \int \left(C_{ij}(\vec{h}) \right) d\vec{h} = \left(\int C_{ij}(\vec{h}) d\vec{h} \right)$. Given the linearity of the integral operator, the Fourier Transform operator—denoted by $\mathcal{F}(\cdot)$—is also a linear one. Thus, $\mathcal{F}\left(\underline{\underline{K}}\right) = \mathcal{F}\left(\underline{\underline{\varphi}} \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}}^t\right) = \underline{\underline{\varphi}} \cdot \mathcal{F}\left(\underline{\underline{C}}\right) \cdot \underline{\underline{\varphi}}^t$. Then $\mathcal{F}\left(\underline{\underline{K}}\right)$ is a positive definite matrix since, for any complex vector $\underline{\lambda} \in \mathbb{C}^D$,

$$\begin{aligned} \underline{\lambda} \cdot \mathcal{F}\left(\underline{\underline{K}}\right) \cdot \bar{\underline{\lambda}} &= \underline{\lambda} \cdot \left(\underline{\underline{\varphi}} \cdot \mathcal{F}\left(\underline{\underline{C}}\right) \cdot \underline{\underline{\varphi}}^t\right) \cdot \bar{\underline{\lambda}} = \left(\underline{\lambda} \cdot \underline{\underline{\varphi}}\right) \cdot \mathcal{F}\left(\underline{\underline{C}}\right) \cdot \left(\underline{\underline{\varphi}}^t \cdot \bar{\underline{\lambda}}\right) = \\ &= \underline{\mu} \cdot \mathcal{F}\left(\underline{\underline{C}}\right) \cdot \bar{\underline{\mu}} \geq 0, \end{aligned}$$

as $\overline{\underline{\lambda} \cdot \underline{\underline{\varphi}}} = \underline{\underline{\varphi}}^t \cdot \bar{\underline{\lambda}} = \bar{\underline{\mu}}$, due to standard properties of conjugation and transposition of complex matrices. □

Proposition 4.3 *If $\underline{\underline{C}}(\vec{h})$ is symmetric for a given basis, then it is symmetric for any other basis of $\mathbb{E}$.*

This is straightforward to show, attending to the fact that $\left(\underline{\underline{A}} \cdot \underline{\underline{B}} \cdot \underline{\underline{C}}\right)^t = \underline{\underline{C}}^t \cdot \underline{\underline{B}}^t \cdot \underline{\underline{A}}^t$.

Proposition 4.4 *If $\underline{\underline{C}}(\vec{h})$ is zero for a given basis at a given lag $\vec{h}$, then it is zero for any other basis of $\mathbb{E}$ at that lag. Consequently, beyond the global range of all the covariances in a given basis, the covariance is zero for any basis.*

It is interesting to note that beyond that global range, the covariance endomorphism is the null one: $C(\mathbf{z}; \vec{h}) = \mathbf{n}$ for all $\mathbf{z} \in \mathbb{E}$.

4.4 Linear prediction

In this section we closely follow the exposition of kriging predictor and its properties given by Pawlowsky-Glahn and Olea [2004, p. 69-76], who in turn use a matrix notation due to Myers [1982]. Here we generalize these expressions to deal with a RF valued on any Euclidean space. In this section we present some results which hold for vectors or endomorphisms of the space, thus are basis-independent, and some which rely upon the orthogonality of a basis. A special attention should be paid to this detail. We use the regression concepts introduced in section 2.3.4, given the connection between kriging and regression stated in section 3.3.4.

4.4.1 The general case of kriging

At an unsampled location $\vec{x}_0$, the unknown $\mathbf{Z}(\vec{x}_0)$ is estimated from the available sample by means of a sum of affine *linear* functions of the observed data,

$$\mathbf{z}^*(\vec{x}_0) = \mathbf{c} \oplus \bigoplus_{n=1}^{N} \lambda_n \mathbf{z}(\vec{x}_n), \tag{4.2}$$

where $\mathbf{c} \in \mathbb{E}$ is a vector of constants and, for each $n = 1, 2, \ldots, N$, λ_n is an endomorphism. Given a basis of $\mathbb{E}$, we may express λ_n as a matrix $\underline{\underline{\lambda}}_n$ Element λ_{ij}^n of $\underline{\underline{\lambda}}_n$ measures the influence of the j-th coordinate of observation $\mathbf{z}(\vec{x}_n)$ on the i-th coordinate of $\mathbf{z}^*(\vec{x}_0)$.

Let $\mathbf{z} = (\mathbf{z}(\vec{x}_1), \mathbf{z}(\vec{x}_2), \ldots, \mathbf{z}(\vec{x}_N))$ be the vector containing all the available observations of the RF $\mathbf{Z}(\vec{x})$. It is clear that its sample space is $\mathbb{F} = \mathbb{E}^N$, and that $\mathbf{z}$ has $D \cdot N$ components We can define on $\mathbb{F}$ an Euclidean structure inherited from that of $\mathbb{E}$, and we can also use a N-tuple replication of a basis of $\mathbb{E}$ as a basis for $\mathbb{F}$. Equation (4.2) can be written as

$$\mathbf{z}^*(\vec{x}_0) = \mathbf{c} \oplus \Lambda \mathbf{z}, \tag{4.3}$$

where $\Lambda : \mathbb{F} \to \mathbb{E}$ is a linear transformation. Given a basis of $\mathbb{E}$ and its associated basis in $\mathbb{F}$, this linear transformation is expressed as a $D \times (D \cdot N)$ matrix $\underline{\underline{\Lambda}} = (\underline{\underline{\lambda}}_1 \, \underline{\underline{\lambda}}_2 \, \cdots \, \underline{\underline{\lambda}}_N)$, where each $\underline{\underline{\lambda}}_n$ is the matrix of each endomorphism λ_n. Equation (4.3) expresses the kriging predictor as a single affine linear function.

Property 4.2 *For the true but unknown $\boldsymbol{Z}(\vec{x}_0)$ and its predictor $\boldsymbol{z}(\vec{x}_0)$ defined in equation (4.2), the following properties hold:*

1. *the vector of expected values of $\boldsymbol{z}^*(\vec{x}_0)$ is $\mathrm{E}_{\mathbb{E}}[\boldsymbol{z}^*(\vec{x}_0)] = \boldsymbol{c} \oplus \bigoplus_{n=1}^{N} \lambda_n \boldsymbol{m}$, where $\boldsymbol{m} = \mathrm{E}_{\mathbb{E}}[\boldsymbol{Z}(\vec{x})]$;*

2. *the variance operator of $\boldsymbol{z}^*(\vec{x}_0)$ is*

$$\mathrm{Var}_{\mathbb{E}}[\boldsymbol{z}^*(\vec{x}_0)] = \bigoplus_{n=1}^{N} \bigoplus_{m=1}^{N} \lambda_n C(\vec{x}_m - \vec{x}_n) \lambda_m^t,$$

   where λ^t represents the adjoint operator of λ (definition 2.11);

3. *the vector of expected prediction errors is $\mathrm{E}_{\mathbb{E}}[\boldsymbol{z}^*(\vec{x}_0) \ominus \boldsymbol{Z}(\vec{x}_0)] = \boldsymbol{c} \ominus \boldsymbol{m} \oplus \bigoplus_{n=1}^{N} \lambda_n \boldsymbol{m}$;*

4. *the variance operator of prediction errors is*

$$\begin{aligned} \Sigma_{K,\mathbb{E}} = \mathrm{Var}_{\mathbb{E}}[\boldsymbol{z}^*(\vec{x}_0) \ominus \boldsymbol{Z}(\vec{x}_0)] \;=\; & \bigoplus_{n=1}^{N} \bigoplus_{m=1}^{N} \lambda_n C(\vec{x}_m - \vec{x}_n) \lambda_m^t \oplus C(\vec{0}) \oplus \\ & \ominus \bigoplus_{n=1}^{N} \left(\lambda_n C(\vec{x}_0 - \vec{x}_n) \oplus C^t(\vec{x}_0 - \vec{x}_n) \lambda_n^t \right); \end{aligned}$$

5. *the predictor $\boldsymbol{z}^*(\vec{x}_0)$ is unbiased if and only if $\boldsymbol{c} = \boldsymbol{m} \ominus \bigoplus_{n=1}^{N} \lambda_n \boldsymbol{m}$.*

Sketch of a proof: The proof of item 1 follows from the linearity of the expectation $\mathrm{E}_{\mathbb{E}}[\cdot]$ and of the predictor (4.2) with respect to the operations of the space, $\oplus$ and the application of endomorphisms, as well as from the second order stationarity of the vector RF. Proof of item 2 is derived using expression (4.3) and the fact that $\mathrm{Var}_{\mathbb{E}}[\mathbf{a}\oplus\Lambda\mathbf{Z}] = \Lambda\mathrm{Var}_{\mathbb{E}}[\mathbf{Z}]\,\Lambda^t$ for any constant vector $\mathbf{a}$, any linear operator Λ and any random vector $\mathbf{Z}$ [Eaton, 1983, p. 76]. This item also needs to take into account the relationship between the variance of the observed data vector and the covariance function: $\mathrm{Var}_{\mathbb{E}}[\mathbf{z}]$ may be understood as an array of endomorphisms, where cell (n,m) contains the operator $\mathrm{Cov}_{\mathbb{E}}[\mathbf{z}(\vec{x}_n),\mathbf{z}(\vec{x}_m)]$, which by definition is equal to $C(\vec{x}_m-\vec{x}_n)$. Proof of item 3 directly follows from the first property. The proof of item 4 is derived using equation (4.3), writing

$$\mathrm{Var}_{\mathbb{E}}[\mathbf{z}^*(\vec{x}_0)\ominus\mathbf{Z}(\vec{x}_0)] =$$

$$= \mathrm{Var}_{\mathbb{E}}[\Lambda\mathbf{z}]\oplus\mathrm{Var}_{\mathbb{E}}[\mathbf{Z}(\vec{x}_0)]\ominus\mathrm{Cov}_{\mathbb{E}}[\Lambda\mathbf{z},\mathbf{Z}(\vec{x}_0)]\ominus\left(\mathrm{Cov}_{\mathbb{E}}[\Lambda\mathbf{z},\mathbf{Z}(\vec{x}_0)]\right)^t$$

and considering the fact that $\mathrm{Cov}_{\mathbb{E}}[\Lambda\mathbf{z},\mathbf{Z}(\vec{x}_0)] = \bigoplus_{n=1}^{N}\lambda_n C(\vec{x}_0-\vec{x}_n)$, again due to the definition $C(\vec{x}_0-\vec{x}_n)\equiv\mathrm{Cov}_{\mathbb{E}}[\mathbf{z}(\vec{x}_n),\mathbf{Z}(\vec{x}_0)]$. Item 5 is the direct application of the definition of unbiasedness and item 3. Note that all covariances in these proofs are defined as operators, and consequently any representation in any coordinate system will be valid. □

As can be seen from item 4 of property 4.2, a covariance operator of prediction errors exists. The classical geostatistical optimization criterion is to minimize a scalar measure of dispersion, but not an operator-valued one. However, if we switch to the coordinate context, we may use Myers [1982] approach: he defined the *prediction variance* for $\mathbf{Z}(\vec{x}_0)$, *i.e.*the value to minimize in the kriging procedure, as the trace of the variance matrix,

$$\sigma^2_{K,\mathbb{E}} = \mathrm{Tr}\left[\mathrm{Var}_{\mathbb{E}}[\mathbf{z}^*(\vec{x}_0)\ominus\mathbf{Z}(\vec{x}_0)]\right]. \tag{4.4}$$

Note that this definition is also well-suited for operators, because the trace of the matrix of an operator does not change with changes of basis. Therefore, $\sigma^2_{K,\mathbb{E}}$ is a property of the variance operator, and is basis-independent.

From equation (4.4), item 4 in property 4.2, and the fact that the adjoint operator has the same associated trace as the original one, the following equation results

$$\sigma^2_{K,\mathbb{E}} = \sum_{n=1}^{N}\sum_{m=1}^{N}\mathrm{Tr}\left[\lambda_n C(\vec{x}_m-\vec{x}_n)\lambda_m^t\right] + \mathrm{Tr}\left[C(\vec{0})\right] - 2\sum_{n=1}^{N}\mathrm{Tr}\left[\lambda_n C(\vec{x}_0-\vec{x}_n)\right].$$

Note that traces are real numbers, and they are operated with classical sum and product.

Property 4.3 *The kriging predictor minimizes the expectation of the squared distance in* $\mathbb{E}$ *between the true value* $\boldsymbol{Z}(\vec{x}_0)$ *and its prediction* $\boldsymbol{z}^*(\vec{x}_0)$.

Proof: Given the identification of the distance in $\mathbb{E}$ between two vectors and the distance in $\mathbb{R}^D$ between its coordinates in an orthonormal basis, it holds that

$$\begin{aligned} \mathrm{E}\left[d^2(\mathbf{z}^*(\vec{x}_0), \mathbf{Z}(\vec{x}_0))\right] &= \mathrm{E}\left[\sum_{i=1}^{D} \left(\zeta^*(\vec{x}_0)_i - Z_i(\vec{x}_0)\right)^2\right] = \sum_{i=1}^{D} \mathrm{E}\left[\left(\zeta_i^*(\vec{x}_0) - Z_i(\vec{x}_0)\right)^2\right] \quad (4.5) \\ &= \sum_{i=1}^{D} \mathrm{Var}\left[\zeta^*(\vec{x}_0)_i - Z_i(\vec{x}_0)\right] = \mathrm{Tr}\left[\mathrm{Var}_{\mathbb{E}}\left[\mathbf{z}^*(\vec{x}_0) \ominus \mathbf{Z}(\vec{x}_0)\right]\right], \quad (4.6) \end{aligned}$$

Thus, the trace of the covariance matrix of equation (4.4) computed with orthonormal coefficients is equal to the expected squared distance between the prediction and its true value. Minimizing the first, we minimize the second. This proof is only valid for kriging systems built using orthonormal bases. However, since we know that both the distance and the trace of the variance operator are not-coordinate dependent, we may be sure that this is valid for any basis. Note the coincidence of this criterion with the minimal-norm criterion of regression (section 2.3.4). Furthermore, this property will be generalized using coordinate arguments in the next sections.

□

4.4.2 Simple kriging

Definition 4.3 (Simple kriging) *The* simple kriging *predictor of $\boldsymbol{Z}(\vec{x}_0)$ is the affine lineal transformation of kriging (eq. 4.2) subject to unbiasedness, which is achieved by*

$$\boldsymbol{c} = \left(I \ominus \bigoplus_{n=1}^{N} \lambda_n\right) \boldsymbol{m} = \boldsymbol{m} \ominus \bigoplus_{n=1}^{N} \lambda_n \boldsymbol{m}. \quad (4.7)$$

Therefore, simple kriging needs a known mean vector $\boldsymbol{m}$ of the RF.

Some particular comments on simple kriging (usually abbreviated as SK) follow.

- Note that this definition ensures the unbiasedness of the predictor $\mathbf{z}^*(\vec{x}_0)$ for any set of endomorphisms $\{\lambda_n, n = 1, 2, \ldots, N\}$.
- SK is also useful when the mean is not constant but known. In this case, a residual RF is computed as $\mathbf{Y}(\vec{x}) = \mathbf{Z}(\vec{x}) \ominus \mathbf{m}(\vec{x})$, and simple kriging is applied to it, taking $\mathbf{c} = \mathbf{n}$. Final predictions are then recovered by adding the mean to the simple kriging predictor $\mathbf{z}^*(\vec{x}_0) = \mathbf{m}(\vec{x}_0) \ominus \mathbf{y}_0^*$

To derive the endomorphisms $\{\lambda_n\}$, take the compact notation of the kriging predictor (4.3) and its minimization criterion expressed as minimal distance of error in equation (4.5). It is clear that the SK predictor and its error are exactly equivalent to regression predictor (2.10) and its error (2.11). Therefore, solution of SK is the same as obtained for regression (equations 2.12 and 2.13). In particular, the joint operator Λ is found to be

$$\Lambda = \left(\mathrm{Cov}_{\mathbb{E}}\left[\mathbf{Z}(\vec{x}_0), \mathbf{z}\right]\right) \left(\mathrm{Var}_{\mathbb{E}}\left[\mathbf{z}\right]\right)^{-1}, \quad (4.8)$$

and the constant value is equal to (4.7). Recall that $\mathbf{z}$ is the concatenated vector of all observed vectors.

This operator-valued solution ensures us the independence of results with respect to the chosen basis. However, it is difficult to work with it. Therefore, we switch here to the coordinate approach, and derive some interesting properties of the SK predictor and error variance structure. The following results assume a fixed basis of $\mathbb{E}$, not necessarily an orthonormal one.

Property 4.4 *If the vector of expected values $\boldsymbol{m}$ is known, then the prediction variance $\sigma^2_{K,\mathbb{E}}$ reaches a minimum when the set of matrices $\{\underline{\underline{\lambda}}_n, n = 1, 2, \ldots, N\}$ satisfies the system of equations*

$$\sum_{n=1}^{N} \underline{\underline{\lambda}}_n \cdot \underline{\underline{C}}(\vec{x}_m - \vec{x}_n) = \underline{\underline{C}}(\vec{x}_m - \vec{x}_0), \quad m = 1, 2, \ldots, N.$$

This system can be written in a compact form as $\underline{\underline{C}} \cdot \underline{\underline{\Lambda}}^t = \underline{\underline{c}}_0$, where $\underline{\underline{C}}$ contains $ND \times ND$ covariances among all the coordinates in all the sampled locations, and $\underline{\underline{c}}_0$ contains the $ND \times D$ covariances between all the coordinates in the sampled locations and those in the predicted location $\vec{x}_0$. With this notation, $\underline{\underline{\Lambda}}^t = \underline{\underline{C}}^{-1} \cdot \underline{\underline{c}}_0$.

Note the coincidence of this last expression for $\underline{\underline{\Lambda}}$ and equation (4.8) for an orthonormal basis.

Property 4.5 *The prediction covariance between the i-th and the j-th coordinate of the SK predictor is*

$$\sigma_{ij} = C_{ij}(\vec{0}) - \sum_{n=1}^{N} \sum_{k=1}^{D} \lambda_{kj,n} C_{ki}(\vec{x}_0 - \vec{x}_n). \tag{4.9}$$

The resulting matrix $\underline{\underline{\sigma}} = (\sigma_{ij})$ is the symmetric covariance matrix of kriging errors [Chilès and Delfiner, 1999, p. 311]. Recall that for an orthonormal basis, $\underline{\underline{\sigma}}$ is equal to the matrix of the endomorphism $\Sigma_{K,\mathbb{E}}$ obtained in property 4.2.4.

Property 4.6 *The prediction variance of the SK predictor is*

$$\sigma^2_{SK,\mathbb{E}} = \sum_{i=1}^{D} \sigma_{ii} = \mathrm{Tr}\left[\underline{\underline{C}}(\vec{0})\right] - \sum_{n=1}^{N} \mathrm{Tr}\left[\underline{\underline{\lambda}}_n \cdot \underline{\underline{C}}(\vec{x}_0 - \vec{x}_n)\right].$$

Proofs of these properties are omitted, since they involve only covariance functions, which are defined in the real coordinates. Equivalent proofs for real vector RFs can be found in Myers [1982], Chilès and Delfiner [1999, p. 311] or Pawlowsky-Glahn and Olea [2004, p. 72-75]. However, the next property is proven, without using operator-driven arguments, to show that the coordinate approach is also self-consistent.

Property 4.7 *The SK prediction does not depend on the basis chosen to represent the vectors of* $\mathbb{E}$*. The kriging error variance matrix, being always expressed in coordinates, depends on the basis, but satisfies the standard conditions of change of basis (4.1).*

Proof: In this proof we drop the double underlining of matrices, because all the elements involved are of this kind. Also, without loss of generality, we consider the mean $\mathbf{m}$ to be the neutral element of $\mathbb{E}$, so its coordinates are $\underline{0}$. Let φ be the $D \times D$ matrix of change of coordinates from basis $\mathbf{F}$ to $\mathbf{E}$, and denote the coordinates with respect to these basis of the spatial distribution $\{\mathbf{z}(\vec{x}_n), n = 1, 2, \ldots, N\}$ with the matrices ζ and ξ, both with $(ND \times 1)$ elements. Then $\zeta = \varphi_N \cdot \xi$, where φ_N represents a $ND \times ND$ block diagonal matrix, with N diagonal blocks equal to φ, and all the elements outside the diagonal equal to zero. Let $C(\vec{h})$ represent the variance structure of the representation ζ, and $K(\vec{h})$ the variance structure of ξ. Then equation (4.1) shows us that it holds $C(\vec{h}) = \varphi \cdot K(\vec{h}) \cdot \varphi^t$. Following property 4.4, let C and K represent respectively the variance matrices of the ζ and ξ coordinate vectors of the spatial distribution, thus $C = \varphi_N \cdot K \cdot \varphi_N^t$. The independent terms of the SK equation (in property 4.4) are denoted respectively by c_0 and k_0 and they satisfy the same expression $c_0 = \varphi_N \cdot k_0 \cdot \varphi^t$. Finally, let the predictions obtained be respectively $\zeta_0 = \Lambda_c \cdot \zeta$ and $\xi_0 = \Lambda_k \cdot \xi$. These weights can be computed according to property 4.4 as $\Lambda_c^t = C^{-1} \cdot c_0$ and $\Lambda_k^t = K^{-1} \cdot k_0$. Then

$$\begin{aligned} \Lambda_c^t &= C^{-1} \cdot c_0 = \left(\varphi_N \cdot K \cdot \varphi_N^t\right)^{-1} \cdot \left(\varphi_N \cdot k_0 \cdot \varphi^t\right) = \varphi_N^{-t} \cdot K^{-1} \cdot \varphi_N^{-1} \cdot \varphi_N \cdot k_0 \cdot \varphi^t = \\ &= \varphi_N^{-t} \cdot K^{-1} \cdot k_0 \cdot \varphi^t = \varphi_N^{-t} \cdot \Lambda_k^t \cdot \varphi^t. \end{aligned}$$

Replacing these weights in the predictor

$$\zeta_0 = \Lambda_c \cdot \zeta = \left(\varphi_N^{-t} \cdot \Lambda_k^t \cdot \varphi^t\right)^t \cdot \varphi_N \cdot \xi = \varphi \cdot \Lambda_k \cdot \varphi_N^{-1} \cdot \varphi_N \cdot \xi = \varphi \cdot \Lambda_k \cdot \xi = \varphi \cdot \xi_0,$$

which implies that the predictor itself satisfies the same change of coordinates relationship as the spatial distribution, and thus $\mathbf{z}^*(\vec{x}) = \zeta_0 \odot \mathbf{E} = \xi_0 \odot \mathbf{F}$.

Regarding the kriging variance, we can write $\Sigma^c = C(\vec{0}) - \Lambda_c \cdot c_0$ and $\Sigma^k = K(\vec{0}) - \Lambda_k \cdot k_0$, where Σ^c and Σ^k are the variance matrices of kriging errors expressed in the two coordinate systems. Then

$$\begin{aligned} \Sigma^c &= \varphi \cdot K(\vec{0}) \cdot \varphi^t - \left(\varphi_N^{-t} \cdot \Lambda_k^t \cdot \varphi^t\right)^t \cdot \varphi_N \cdot k_0 \cdot \varphi^t = \\ &= \varphi \cdot K(\vec{0}) \cdot \varphi^t - \varphi \cdot \Lambda_k \cdot \varphi_N^{-1} \cdot \varphi_N \cdot k_0 \cdot \varphi^t = \\ &= \varphi \cdot K(\vec{0}) \cdot \varphi^t - \varphi \cdot \Lambda_k \cdot k_0 \cdot \varphi^t = \varphi \cdot \Big(K(\vec{0}) - \Lambda_k \cdot k_0\Big) \cdot \varphi^t = \varphi \cdot \Sigma^k \cdot \varphi^t. \end{aligned}$$

Consequently, the kriging variance matrix obtained using a basis can be linearly transformed to the kriging variance matrix expressed in any other basis, in accordance with equation (4.1). □

Proposition 4.5 *If the vector RF $\boldsymbol{Z}(\vec{x})$ is a stationary Gaussian one on $\mathbb{E}$ according with definition 4.2, then SK vector predictor and its operator variance give the parameters of the normal distribution on $\mathbb{E}$ for $\boldsymbol{Z}(\vec{x}_0)$ conditional on the observed spatial distribution $\boldsymbol{z}(\vec{x}_1), \boldsymbol{z}(\vec{x}_2), \ldots, \boldsymbol{z}(\vec{x}_N)$,*

$$[\boldsymbol{Z}(\vec{x}_0)|\boldsymbol{z}(\vec{x}_1), \boldsymbol{z}(\vec{x}_2), \ldots, \boldsymbol{z}(\vec{x}_N)] \sim \mathcal{N}_{\mathbb{E}}\left(\boldsymbol{z}^*(\vec{x}_0), \Sigma_{K,\mathbb{E}}\right).$$

Given the identification between the normal distribution of $\mathbf{Z}(\vec{x})$ on $\mathbb{E}$ and that of its coordinates $\underline{Z}(\vec{x})$ on $\mathbb{R}^D$, this property is equivalent to the conditional expectation satisfied by SK in $\mathbb{R}$. For a proof of the identification between the SK predictor and variance with the parameters of the multivariate normal distribution in the $\mathbb{R}^D$ case see Pawlowsky-Glahn and Olea [2004].

4.4.3 Universal kriging

In the usual situation, the mean is not known, and it might even be considered not a constant. In this case, simple kriging is not applicable. Therefore, we cannot ensure unbiasedness of the kriging predictor (4.2) by satisfying the unbiasedness condition on the constant (4.7) and leaving the endomorphisms λ_n free. Instead, we take $\mathbf{c} = \mathbf{n}$ to filter the unknown mean from the predictor, and then look for restrictions on the endomorphisms. The resulting predictor is called *Universal Kriging*. Let us explain this step by step.

Definition 4.4 (Drift functions) *Assume the mean of a RF to be unknown, but a linear function of $(A+1)$ known real-valued functions $g_a(\vec{x}), a = 0, 1, \ldots, A$:*

$$\boldsymbol{m}(\vec{x}) = \bigoplus_{a=0}^{A} g_a(\vec{x}) \odot \boldsymbol{b}_a. \tag{4.10}$$

These functions are called drift functions.

Property 4.8 *Unbiasedness of the kriging linear predictor (4.2) is achieved by forcing*

$$g_a(\vec{x}_0) I = \bigoplus_{n=1}^{N} g_a(\vec{x}_n) \odot \lambda_n, \tag{4.11}$$

with $a = 0, 1, \ldots, A$. These $A+1$ vector equations are called universality conditions.

Proof: Taking the condition for unbiasedness from property 4.2.5, we obtain

$$\mathbf{n} = \mathbf{c} = \left(I \ominus \bigoplus_{n=1}^{N} \lambda_n\right) \mathbf{m} = T\mathbf{m},$$

with operator $T = I \ominus \bigoplus_{n=1}^{N} \lambda_n$. Now replacing the mean by its expression as a combination of the drift functions, it yields

$$\mathbf{n} = T\left(\bigoplus_{a=0}^{A} g_a(\vec{x}) \odot \mathbf{b}_a\right) = \bigoplus_{a=0}^{A} g_a(\vec{x}) \odot T\mathbf{b}_a$$

by linearity of the operator T. Since this equality with the neutral element has to be fulfilled for all possible $\mathbf{b}_a \in \mathbb{E}$, then each of the operators $T'_a = (g_a(\vec{x}) \odot T)$ must be the null operator

$$N = g_a(\vec{x}) \odot T = g_a(\vec{x}) \odot \left(I \ominus \bigoplus_{n=1}^{N} \lambda_n \right).$$

Recall that the null operator satisfies $N\mathbf{a} = \mathbf{n}$ for any $\mathbf{a} \in \mathbb{E}$, and it is the neutral element of the vector space of endomorphisms $L(\mathbb{E}, \mathbb{E})$. Given the linearity of the external product, we may rearrange this expression and obtain equation (4.11). □

Definition 4.5 (Universal kriging) *The* universal kriging *predictor is a linear transformation of the observations (eq. 4.2), with $\boldsymbol{c} = \boldsymbol{n}$ and endomorphisms satisfying the universality condition (4.11) for each drift function.*

Some comments regarding particular cases of universal kriging follow.

- If we take a single drift function considered constant, $g_0(\vec{x}) \equiv 1$, then the technique is called *general or ordinary kriging* (abbreviated OK). In this ordinary case, the variogram functions are enough to apply kriging, because they filter directly the unknown but constant mean. However, recall that variograms do not capture asymmetry features of the covariance structure. Also the universality conditions are in this ordinary case simply

$$I = \bigoplus_{n=1}^{N} \lambda_n.$$

- The term *universal kriging* (abbreviated UK) is usually reserved for the case when $g_i(\vec{x})$ are the polynomials of some degree of location $\vec{x}$, always including the constant as the first drift function.

- If we take as drift functions the constant and a known external variable, which depends also on location $g_1(\vec{x})$ (*e.g.*topography trying to predict a yearly temperature RF for instance), then the obtained technique is usually called *kriging with an external drift or trend kriging* (abbreviated TK).

Despite these variations, we will refer to all of them in the next developments under the term *universal kriging* (UK). Note that this model is connected to regression in an Euclidean space, as was explained in section 2.3.4.

As happened with simple kriging, UK system expressed in operators is particularly cumbersome. Therefore, we introduce again an arbitrary basis of the space and derive matrix formulae, so that the procedure is better understood.

If we express all endomorphisms and vectors with respect to a chosen basis of $\mathbb{E}$, then the prediction variance $\sigma^2_{K,\mathbb{E}}$ has to be modified to include (through Lagrange multipliers) the set of universality conditions expressed in coordinates:

$$\underline{\underline{0}} = \sum_{n=1}^{N} g_a(\vec{x}_n) \cdot \underline{\underline{\lambda}}_n - g_a(\vec{x}_0) \underline{\underline{I}}_D \quad a = 0, 1, \ldots, A.$$

Note that in total, there are $(A+1)D^2$ universality conditions, which weighted sum with their Lagrange multipliers can be expressed as:

$$\sum_{a=0}^{A}\sum_{i=1}^{D}\sum_{j=1}^{D}\nu_{a,ij}\left(\sum_{n=1}^{N}g_a(\vec{x}_n)\cdot\lambda_{n,ij}-g_a(\vec{x}_0)\delta_{ij}\right)=$$

$$=\mathrm{Tr}\left[\sum_{a=0}^{A}\underline{\underline{\nu}}_a\left(\sum_{n=1}^{N}g_a(\vec{x}_n)\cdot\underline{\underline{\lambda}}_n^t-g_a(\vec{x}_0)\cdot\underline{\underline{I}}_D\right)\right].$$

The final quantity to minimize is then

$$Q=\mathrm{Tr}\left[\sum_{n=1}^{N}\sum_{m=1}^{N}\underline{\underline{\lambda}}_n\cdot\underline{\underline{C}}_{nm}\cdot\underline{\underline{\lambda}}_m^t++\underline{\underline{C}}_{00}-2\sum_{n=1}^{N}\underline{\underline{\lambda}}_n\cdot\underline{\underline{C}}_{n0}+\sum_{a=0}^{A}\underline{\underline{\nu}}_a\left(g_a(\vec{x}_n)\cdot\underline{\underline{\lambda}}_n^t-g_a(\vec{x}_0)\cdot\underline{\underline{I}}_D\right)\right],$$

where $\underline{\underline{C}}_{nm}\equiv\underline{\underline{C}}(\vec{x}_m-\vec{x}_n)$. Note that the matrix inside Tr[·] has been simplified attending to the fact that the trace of a matrix coincides with that of its transpose. By derivation of Q as a function of $\lambda_{n,ij}$ and $\nu_{a,ij}$, equating the result to 0, and rearranging it in matrices, it can be shown that Q reaches a minimum when the set of matrices $\{\underline{\underline{\lambda}}_n, n=1,2,\ldots,N\}$ satisfy the system of equations

$$\sum_{n=1}^{N}\underline{\underline{\lambda}}_n\cdot\underline{\underline{C}}(\vec{x}_m-\vec{x}_n)+\sum_{a=0}^{A}g_a(\vec{x}_n)\cdot\underline{\underline{\nu}}_a=\underline{\underline{C}}(\vec{x}_m-\vec{x}_0)\quad m=1,2,\ldots,N. \tag{4.12}$$

$$\sum_{n=1}^{N}g_a(\vec{x}_n)\cdot\underline{\underline{\lambda}}_n=g_a(\vec{x}_0)\underline{\underline{I}}_D\quad a=0,1,\ldots,A. \tag{4.13}$$

This system can be written in a compact form as $\hat{\underline{\underline{C}}}\cdot\hat{\underline{\underline{\Lambda}}}^t=\hat{\underline{\underline{c}}}_0$, where

$$\hat{\underline{\underline{C}}}=\begin{pmatrix}&&&\underline{\underline{G}}_1^t\\&\underline{\underline{C}}&&\vdots\\&&&\underline{\underline{G}}_N^t\\\underline{\underline{G}}_1&\cdots&\underline{\underline{G}}_N&\underline{\underline{0}}\end{pmatrix}$$

contains the $ND\times ND$ covariance matrix $\underline{\underline{C}}$ of the simple kriging system, and the set of N block matrices $\underline{\underline{G}}_n^t=\left(g_0(\vec{x}_n)\underline{\underline{1}}_D,\ldots,g_A(\vec{x}_n)\underline{\underline{1}}_D\right)$, where each block contains D identical rows and columns; also,

$$\hat{\underline{\underline{\Lambda}}}^t=\begin{pmatrix}\underline{\underline{\Lambda}}^t\\\underline{\underline{\nu}}\end{pmatrix},$$

contains the matrix of $D\times ND$ weights Λ of equation (4.3), and the Lagrange multipliers $\nu_{a,ij}$ involved in each one of the $(A+1)D^2$ universality conditions, arranged in a column

block matrix

$$\underline{\underline{\nu}} = \begin{pmatrix} \underline{\underline{\nu}}_0 \\ \underline{\underline{\nu}}_1 \\ \vdots \\ \underline{\underline{\nu}}_A \end{pmatrix}$$

with $\underline{\underline{\nu}}_a = (\nu_{a,ij})$; finally,

$$\hat{\underline{\underline{c}}}_0 = \begin{pmatrix} \underline{\underline{c}}_0 \\ \underline{\underline{G}}^t \\ \underline{\underline{G}}_0 \end{pmatrix},$$

contains the covariance $\underline{\underline{c}}_0$ between all the coordinates in the sampled locations and those in the predicted location $\vec{x}_0$, as well as $\underline{\underline{G}}_0$, the independent terms of the universality conditions (4.11). With this notation, the solution is obtained with

$$\hat{\underline{\underline{\Lambda}}}^t = \hat{\underline{\underline{C}}}^{-1} \cdot \hat{\underline{\underline{c}}}_0.$$

Property 4.9 *The prediction covariance between the i-th and the j-th coordinate of the universal kriging predictor is [Chilès and Delfiner, 1999, p. 311]:*

$$\sigma_{ij} = C_{ij}(\vec{0}) - \sum_{n=1}^{N}\sum_{k=1}^{D} \lambda_{kj,n} C_{ki}(\vec{x}_0 - \vec{x}_n) - \sum_{a=0}^{A} \nu_{a,ij} g_a(\vec{x}_0).$$

Therefore, the prediction variance of the universal kriging predictor is

$$\sigma^2_{SK,\mathbb{E}} = \sum_{i=1}^{D} \sigma_{ii} = \mathrm{Tr}\left[\underline{\underline{C}}(\vec{0})\right] - \sum_{n=1}^{N} \mathrm{Tr}\left[\underline{\underline{\lambda}}_n \cdot \underline{\underline{C}}(\vec{x}_0 - \vec{x}_n)\right] - \sum_{a=0}^{A} g_a(\vec{x}_0)\mathrm{Tr}\left[\underline{\underline{\nu}}_a\right].$$

Property 4.10 *The universal kriging prediction does not depend on the basis chosen to represent the vectors of $\mathbb{E}$. The kriging error covariance matrix, being always expressed in coordinates, depends on the basis, but satisfies the standard conditions of change of basis (4.1).*

Sketch of a proof: The proof is exactly equivalent to that of property 4.7 by using the same notation and taking into account that

$$\begin{aligned} \hat{C} &= \varphi_{(N+1)} \cdot \hat{K} \cdot \varphi^t_{(N+1)} \\ \hat{c}_0 &= \varphi_{(N+1)} \cdot \hat{k}_0 \cdot \varphi^t \\ g_a(\vec{x}) \odot \mathbf{c}_a = g_a(\vec{x}) \underline{c}_a \odot \mathbf{E} &= g_a(\vec{x}) \odot (\varphi \underline{k}_a \odot \mathbf{F}) = (g(\vec{x}) \cdot \varphi) \odot \mathbf{k}. \end{aligned}$$

Note that $\hat{C}, \hat{K}, \hat{c}_0$ and $\hat{k}_0$ are real-valued covariance matrices, while $\mathbf{c}_a, \mathbf{k}_a \in \mathbb{E}$ are the vectors of constants which describe the mean of the function $\mathbf{m} = \bigoplus g_a(\vec{x}_0) \odot \mathbf{c}_a$ jointly with the drift functions $g_a(\vec{x})$. Finally, $\underline{c}_a, \underline{k}_a \in \mathbb{R}^D$ are the coordinates of these vectors of constants. □

Recall of Property 4.3 *The kriging predictor minimizes the expectation of the squared distance in* $\mathbb{E}$ *between the true value* $\boldsymbol{Z}(\vec{x}_0)$ *and its prediction* $\boldsymbol{z}^*(\vec{x}_0)$.

Proof: Given the independence of the simple and universal kriging predictors with respect to the basis (properties 4.7 and 4.10) and the fact that the distance in $\mathbb{E}$ does not depend on the basis chosen for the space, note that this property 4.3 is now valid for any basis, not necessarily an orthonormal one. □

4.5 Remarks

Here we presented a generalization of some of the most-frequently used geostatistical concepts and techniques to deal with dependent observations which sample space can be meaningfully structured as an Euclidean space. It this chapter, we have shown many results, which deserve a clear summary.

- Concepts related to real RFs easily translate into vector-valued RFs.
- The structural functions can be defined as endomorphisms depending on a parameter (the lag distance). This ensures that their properties (positive-definiteness, symmetry, ranges) are intrinsic to them, and not artifacts induced by the choice of a basis. However, we have proven this intrinsic character also by choosing two arbitrary basis and comparing the properties of the obtained structural functions.
- Kriging predictors can be built as affine linear transformations of the observed data. The kriging procedure looks for those endomorphisms which produce a smaller error variance operator. The concept of *size of an operator* is chosen to be the trace of any of its matrices in a basis, because the trace of a matrix is invariant by changes-of-basis. This coincides with Myers [1982] approach.
- The kriging error variance is defined as an endomorphism, which again does not depend on any basis of representation. However, we have proven again that the kriging error matrices obtained with two bases are related with the standard formulae of change-of-basis.

Summarizing, we have built the most usual geostatistical concept, tools and techniques directly using vectors and endomorphisms in an Euclidean space, without using any basis. Therefore, the choice of a basis will not affect our results. However, the development of all results in terms of vectors and endomorphisms is quite cumbersome, and difficult to deal with. Fortunately, in practical application we can choose any basis and work with the coordinates with respect to it, following the *principle of working on coordinates* [Pawlowsky-Glahn, 2003]. This chapter ensure that this choice will not affect the final results, thus it is a confirmation of the applicability of this *principle* in the geostatistical field.

This chapter does not contain any example, because other chapters of this thesis present particular cases, which may serve as illustration. The real space has been

already treated in chapter 3, and further examples will be given in chapter 5 (the positive real space), and in chapters 6 and 7 (the Simplex, for compositions and for discrete probability densities, respectively).

Chapter 5

Geostatistics in the positive real space

5.1 Lognormal kriging

Let $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$ be a point (or the center of a block v) in a domain $\mathcal{D}$ of the space-time real space, with $p \in \{1, 2, 3, 4\}$. Let $\mathbf{Z}(\vec{x}) \in \mathbb{R}_+$ be a positive random function (RF), where $\mathbb{R}_+$ denotes the positive part of the real line, $\mathbb{R}$. Let $\mathbf{z}(\vec{x}_1), \mathbf{z}(\vec{x}_2), \ldots, \mathbf{z}(\vec{x}_N)$ be an observed sample of this RF. The fact that these values are positive preclude in general to assume $\mathbf{Z}(\vec{x})$ to be a Gaussian RF, since this should be defined on the whole $\mathbb{R}$. Then, a logarithmic transformation may be applied to the data set (specially if it has a positively-skewed histogram) to obtain scores $\zeta(\vec{x}_n) = \log z(\vec{x}_n)$, which are assumed to form the sample of $Z(\vec{x})$, a Gaussian function. This implies that the original untransformed RF $\mathbf{Z}(\vec{x})$ is assumed to be a lognormal one. In this section the bold uppercase letter $\mathbf{Z}$ represents the positive RF, and its lowercase counterparts $\mathbf{z}(\vec{x}_n)$ are the observed sample. The logarithmically transformed samples are denoted by the Greek lowercase counterpart $\zeta(\vec{x}_n)$ and its real-valued RF by Z. Whenever any of these samples or RFs form a vector of real values, this will be indicated by a simple underlining, like $\underline{z}_n$ or $\underline{Z}$ respectively.

Assuming $Z \in \mathbb{R}$ to be a Gaussian function, either stationary or intrinsic, the theory and methods introduced in chapter 3 are applicable. Thus, using covariances, variograms and drift functions defined on the $\zeta(x_n)$ values, each kriging technique will provide the best linear unbiased estimate for $Z(\vec{x}_0)$ at an unsampled location $\vec{x}_0$, with the usual assumptions for each of these methods. Let then ζ^*_{SK} and $\sigma^2_{\zeta,SK}$ denote the simple kriging predictor (3.23) and its variance (3.25) for a real-valued RF, and equivalently, let ζ^*_{UK} and $\sigma^2_{\zeta,UK}$ be their universal kriging counterparts (3.16) and (3.17).

Property 5.1 (lognormal kriging predictor) *With the notations already introduced, the conditional expectation of $Z(\vec{x}_0)$ is*

$$\boldsymbol{z}^*_{SK} = \exp\left(\zeta^*_{SK} + \frac{1}{2}\sigma^2_{\zeta,SK}\right), \tag{5.1}$$

and its variance is

$$\sigma^2_{z,SK} = \left(z^*_{SK}\right)^2 \cdot \left(\exp\left(\sigma^2_{\zeta,SK}\right) - 1\right). \tag{5.2}$$

Sketch of a proof: These properties are directly deduced from the relationship between lognormal and normal expectations given by equation (2.22) and by the fact that SK predictor and variance give the conditional expectation (3.26). Due to these arguments, (5.1) is called the *lognormal kiging predictor*, and this is the reason why we used notations $\mathbf{z}^*_{SK}$ and $\sigma^2_{\mathbf{z},SK}$, explicitly marking their role as kriging predictors. □

Simple kriging is seldom applicable, since it needs the mean to be known. A possible solution consists in replacing the $\mathbf{z}^*_{SK}$ predictor by $\mathbf{z}^*_{UK}$. However, it is unknown to which extent expressions (5.1) and (5.2) keep their meaning of conditionally unbiased predictors of $\mathbf{Z}(\vec{x}_0)$ and its variance. Also, it is unclear whether it is better to replace in these equations the simple kriging variance $\sigma^2_{\zeta,SK}$ by its universal estimate $\sigma^2_{\zeta,UK}$—opinion of Cressie [1991], according to Clark and Harper [2000, p. 323]—or to leave it unchanged—as proposed by Journel and Huijbregts [1978]—. All these authors notice the extreme sensitivity of predictor (5.1), since any error in the kriging or variance estimations become exponentially magnified. The uncertain fitting of the sill is particularly important, as it has a direct effect on the kriging variance. Moreover, deviations from lognormality of the data may dramatically invalidate its use [Clark and Harper, 2000, David, 1977].

In addition, Journel and Huijbregts [1978, p. 572] show that equation (5.1) is usually locally biased: given $\bar{z}$, the estimated expected value of $\mathbf{Z}$, they propose to modify (5.1) by a multiplicative factor

$$K = \frac{\bar{z}}{\sum_{\vec{x}_0 \in \mathcal{D}} \mathbf{z}^*_{SK}(\vec{x}_0)}$$

This correction was proposed for block kriging and, according to Chilès and Delfiner [1999, p. 192], should be avoided in point kriging problems, since the obtained predictor is no longer an exact interpolator.

Likewise, Equation (5.1) was shown to be a bad estimator of the local conditional expectation by Roth [1998]. His example shows a case with final estimates surpassing the convex envelope of the observed data, being all kriging weights strictly positive. He blames the presence of the kriging variance for this behavior, and advocates for the estimation of the median, instead of the mean. In agreement with this proposal, Chilès and Delfiner [1999, p. 191] explain that good results are obtained working with quantiles, since they simply follow the logarithmic transformation. In other words, the direct relationship $m_{\underline{Z}} = \log m_{\mathbf{Z}}$ between the medians $m_{\underline{Z}}$ and $m_{\mathbf{Z}}$ allows an easy estimation of the median of $\mathbf{Z}$ as a central tendency measure. This approach offers a way to compute confidence intervals around the median, like in expression (2.17), which gives a 95% confidence interval

$$\mathbf{Z} \in \left(\exp\left(\mathbf{z}^*_{SK} - 1.96\sigma_{\zeta,SK}\right); \exp\left(\mathbf{z}^*_{SK} + 1.96\sigma_{\zeta,SK}\right)\right). \tag{5.3}$$

However, confidence intervals (5.3) are not considered to be optimal, as they do not have minimal Euclidean length in $\mathbb{R}$. According to Clark and Harper [2000], this issue was studied in the non-geostatistical case by Sichel [1971], who obtained a set of multiplicative coefficients intended to compute unbiased means and optimal confidence intervals for the lognormal distribution. These coefficients depend on the sample size and the logarithmic variance, but they do not capture explicitly spatial correlation, and their application must be restricted to low-variance data. Another option is to use expressions (5.1) and (5.2) to compute directly a *nominal* interval with them, based on Gaussian assumptions, although its physical meaning is dubious: as shown by Tolosana-Delgado and Pawlowsky-Glahn [2003], it might include a fairly high proportion of negative values.

If the interest is simulation, then (5.1) and (5.2) are seldom used [Deutsch and Journel, 1992, p. 76]. Instead, the kriging prediction (either $\mathbf{z}^*_{SK}$ or $\mathbf{z}^*_{UK}$) and its variance $\sigma^2_{\zeta,K}$ are used to simulate normal scores of ζ, that afterwards are transformed to simulations of $\mathbf{z}$ through $\mathbf{z} = \exp(\zeta)$. This procedure is consistent with the median estimates and confidence intervals like (5.3).

Block lognormal kriging is a far more complex issue, which involves many theoretical and practical considerations, being the most important one the so-called permanence of lognormality: having lognormal point values, any weighted arithmetic mean, *e.g.*any kriging predictor (5.1) or a spatial average like (3.31), will never be lognormal. This well-known result formally invalidates the rigorous application of basic change-of-support models to lognormal sets, although Rendu [1979] shows that the discrete Gaussian model gives fairly good empirical results. Journel [1980] presents an approximate but elegant solution based on point lognormal kriging, specially suited to describe detailed local variability through block distribution functions and selectivity curves (see section 5.4) instead of by a detailed kriged map. Marcotte and Groleau [1997] try a complex solution, assuming joint lognormality of the estimator and the real value, and obtain an estimator similar to (5.2), by replacing the kriging variance by a conditional variance which is usually smaller. Dowd [1982] develops an alternative requiring the numerical resolution of a system of integral equations, without assuming permanence of lognormality.

5.2 Positive real line space structure

The problems of lognormal kriging might be related to the fact that it does not take full profit of an own Euclidean structure for the positive real line $\mathbb{R}_+$. This structure is particularly suited to capture multiplicative processes, and compares individuals in a logarithmic scale. It has been detailed throughout chapter 2, and specially in section 2.5.2. In particular, given a basis of $\mathbb{R}_+$ like $\mathbf{e} = (a) \neq (1)$, the coefficient of any element $\mathbf{z}$ of the space is

$$\zeta = \log_a \mathbf{z}.$$

Using these coefficients, a Lebesgue measure (2.3) can be defined in the Euclidean structure of $\mathbb{R}_+$, which allows to introduce the normal distribution on $\mathbb{R}_+$ as an alternative to the lognormal distribution [Mateu-Figueras et al., 2002]. Both this normal distribution and the Lebesgue measure on $\mathbb{R}_+$ are represented in figure 2.4. At the sight of these measure considerations, it seems clear that lognormal kriging follows a two-step estimation procedure,

1. first, it estimates the logarithm of the RF at the unsampled location, by minimizing the natural distance between the true value of $\zeta(\vec{x}_0)$ and its estimate $\zeta^*_{\cdot K}$ on $\mathbb{R}_+$;

2. second, it estimates the RF at the unsampled location $\mathbf{z}(\vec{x}_0)$ as the mean of a lognormal distribution; recall that this distribution is built with reference to the Lebesgue measure on $\mathbb{R}$.

It might be argued that such a procedure mixes a logarithmic distance and a Lebesgue measure which are not fully compatible. This lack of compatibility may be the source of lognormal kriging problems.

5.3 Kriging in the positive real space

Taking into account that the positive real line $\mathbb{R}_+$ may be given an Euclidean space structure, we can apply the general kriging techniques of chapter 4 to any RF with a positive scale. In this case, the space is one-dimensional, and the coordinate is computed by taking the logarithm, thus giving as the kriging predictor

$$\zeta^*_0 = \log_a \mathbf{z}^*_0 = \log_a \mathbf{c} + \sum_{n=1}^{N} \lambda_n \cdot \log_a \mathbf{z}(\vec{x}_n) = \log_a(c) + \sum_{n=1}^{N} \lambda_n \cdot \zeta(\vec{x}_n),$$

where unbiasedness is ensured setting $\log_a \mathbf{c} = 0$ in Universal Kriging, and $\log_a \mathbf{c} = \mu(1 - \sum \lambda_n)$ in Simple Kriging (property 4.2.5, and definitions 4.3 and 4.5). Its kriging variance is

$$\sigma^2_{K,\mathbb{R}_+} = \mathrm{Var}\left[\log_a \mathbf{z}^*_0 - \log_a \mathbf{Z}(\vec{x}_0)\right],$$

These values are equivalent to respectively ζ^*_{SK} and $\sigma^2_{\zeta,SK}$ or their universal kriging counterparts, introduced in section 5.1, and used in the first step of lognormal kriging, as has been explained in section 5.2. Thus, the kriging system to be solved with kriging in $\mathbb{R}_+$ is exactly equivalent to that of lognormal kriging. However, the best unbiased linear predictor of $\mathbf{Z}(\vec{x})$ is

$$\mathbf{z}^*(\vec{x}_0) = \zeta^*_0 \odot \mathbf{e} = \exp\left(\ln(a) \cdot (\zeta^*_0)\right), \tag{5.4}$$

which coincides with the estimator of the median[1], proposed as an alternative to the lognormal kriging predictor by some authors (section 5.1). Consequently, the normal

[1]Note that in the last section, we introduced the possibility of using any basis a for the logarithm, which in fact does not change the results, only their scale of representation. However, $a = e$ in most of the cases.

kriging in $\mathbb{R}_+$ predictor (5.4) is a good estimator of the local conditional expectation [Roth, 1998], it is unsensitive to the fitting of the variogram sill and so robust against departures from normality on $\mathbb{R}_+$ as classical kriging is for departures from normality on $\mathbb{R}$. Furthermore, a 95% confidence interval built by assuming an underlying normal on $\mathbb{R}_+$ distribution is equivalent to (5.3), but here it can be considered to be optimal, in the sense that it has minimum logarithmic length. It is worth note that predictor (5.4) yields a weighted geometric average of the data,

$$\mathbf{z}^*(\vec{x}_0) = \mathbf{c} \cdot \prod_{i=1}^{N} \mathbf{z}^{\lambda_n}(\vec{x}_n).$$

Although this section explained only the univariate case, a kriging prediction of positive-valued RFs is not more difficult than classical multivariate kriging, once the observations and the predictors are expressed in coordinates. This coordinates are always real-valued, which allows for a direct application of the kriging techniques explained in chapter 3, and particularly universal kriging (section 3.3.1). Other multivariate cases are further explained in chapter 6.

5.4 Change-of-support problems

5.4.1 Lognormal change-of-support model

Consider a set of measurements of ammonia in the effluents of a waste-water treating plant (WWTP), regularly taken each minute during a week. Note that ammonia content is a positive variable, and a logarithmic scale may be considered fit to it. Due to its pernicious effects on the fluvial environment, this pollutant is strictly controlled, and WWTPs must report to the public management agencies any event of dumping an excess of ammonia to the river. However, agencies do not regulate the quasi-continuous measurement, but put an upper bound to the allowed average concentration of ammonia during an hour. The managers of the WWTP decide that they do not want to risk the fine in more than 5% of the cases. In such a case, we would have measurements in a 1-minute support (considered as point-support), but our questions would concern the probability distribution of an average in a time period of 60 minutes. Lognormal block kriging would give us estimates of the hourly averages, and a lognormal change-of-support model would allow us to approximate the distribution of hourly averages during a day, so that we can compute which is the probability of getting a fine each day. If this would be greater than the acceptable level, some corrections on the WWTP should be done.

The application of logarithmic transformations to block kriging (for local estimation) is a complex issue, which primarily involves the so-called *conservation of log-normality* assumption. Conservation of lognormality assumes that given a lognormal point-support RF, the linked block-support RF will be also a lognormal one. Although actually this cannot be strictly true, there is an empirical evidence of its fitness as

long as the blocks remain *small* compared with the range [Chilès and Delfiner, 1999, p. 433].

Recall that change-of-support models (for global estimation) put forward possible approximate relations between these point- and block-support RFs in order to approximate the distribution of the later with data from the former (section 3.5.3). In the case of lognormal RFs, one uses either a special case of the discrete Gaussian model, or a direct application of the multi-Gaussian model (definitions 3.21 and 3.22). Both cases are based on a transformation, which in this case is $\phi(\cdot) = \log(\cdot)$. By using the first model, the distribution of the block-support RF (3.31) is a lognormal distribution,

$$Z_v \sim \mathcal{L}\left(\log m_z - \frac{1}{2} r\sigma_\zeta, (r\sigma_\zeta)^2\right),$$

where r is the correlation coefficient of the discrete Gaussian model, computed as

$$r^2 = \frac{\log\left(1 + \sigma^2_{v,z}/m_z^2\right)}{\log\left(1 + \sigma_z^2/m_z^2\right)},$$

and m_z, σ_z^2 are the mean and variance of the original variable Z, σ_ζ^2 is the variance of its logarithmic transformation, and $\sigma^2_{v,z}$ is the variance of the block RF for Z. It is worth noting again that these models are approximations, which cannot strictly speaking be true. However, they were used to obtain approximate analytical expressions of the distribution of the RF in block support.

A classical way to deal with these block distributions is through the *selectivity curves*. These were introduced in mining applications by Lasky [1950]. Goovaerts et al. [1997] presented some applications to environmental issues. Selectivity curves are alternative ways to represent the estimated probability law in a block. These curves are represented in figure 5.1.

Definition 5.1 (Selectivity curves) *For any arbitrary positive distribution $F(\cdot)$ we may define the following associated functions:*

1. *the* total tonnage, *$T(z)$, corresponds to the complementary of the cumulative distribution, and explains the proportion of blocks in a deposit which average mineral content is above a cutoff value z,*

$$T(z) = \int_z^{+\infty} dF(u) = 1 - F(u);$$

   in this work, it has been already used to characterize the hazard of high conductivity on time units in section 3.6;

2. *the* quantity of metal, *$Q(z)$, is the expected metal (or pollutant) content above a certain threshold z*

$$Q(z) = \int_z^{+\infty} u \cdot dF(u);$$

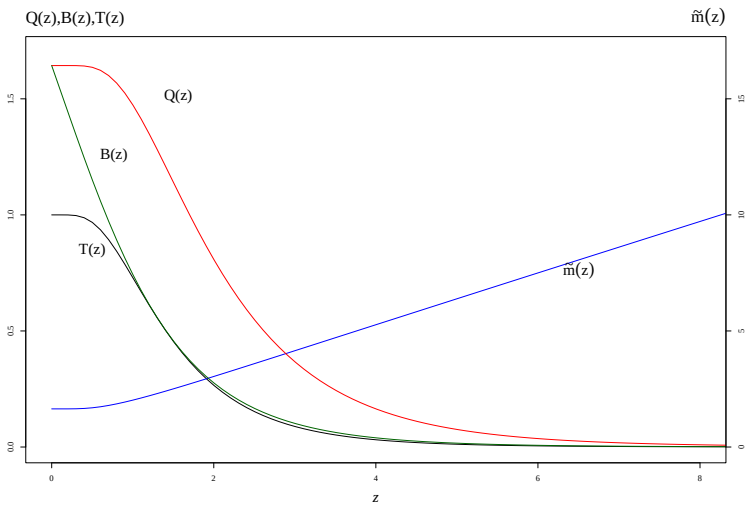

Figure 5.1: Selectivity curves for a lognormal distribution ($\mu = 0; \sigma = 1$). Decreasing ones, from top to bottom: quantity of metal, conventional income, tonnage. Increasing one: mean grade.

3. *the* mean grade, $\tilde{m}(z)$, *is the expected quantity of metal (or pollutant) conditional to be above the threshold* z

$$\tilde{m}(z) = \mathrm{E}\left[Z|Z \geq z\right] = \frac{Q(z)}{T(z)}; \tag{5.5}$$

4. *the* conventional income, $B(z)$, *is the expected income (or remediation cost) of removing and processing a rich block, either for its high mineral or pollutant content:*

$$B(z) = \int_z^{+\infty} (u - z) \cdot dF(u) = \int_z^{+\infty} T(u) \cdot du. \tag{5.6}$$

This curve is always continuous and non-increasing, independently of the nature of the RF. It is bounded by its value at the origin, the mean of the RF $m = B(0) \geq B(z)$, *and tends to zero with high values of* z. *Figure 5.2 shows that it decreases slower when the distribution is more selective (definition 3.23).*

Returning to our WWTP case example, we would obtain directly the sought hazard by plugging the legal threshold (say, $z_0 = 0.025\mathrm{mgNH_3}/l$) into $T(z_0)$. If we were interested in the average ammonia concentration of these hazardous events, these would be obtained with $\tilde{m}(z_0)$. Finally, we could assume that the fine could be proportional to the excess over the threshold, with proportionality constant $a\$l/mg\mathrm{NH_3}$. In this case, if the managers asked for the expected cost of the fine, we could look at $a \cdot B(z_0)$.

5.4.2 Normal on $\mathbb{R}_+$ change-of-support model

Note that in the example of the last section, we considered the *average* in the block as an arithmetic mean of point values. This is not necessarily a good model. Such an

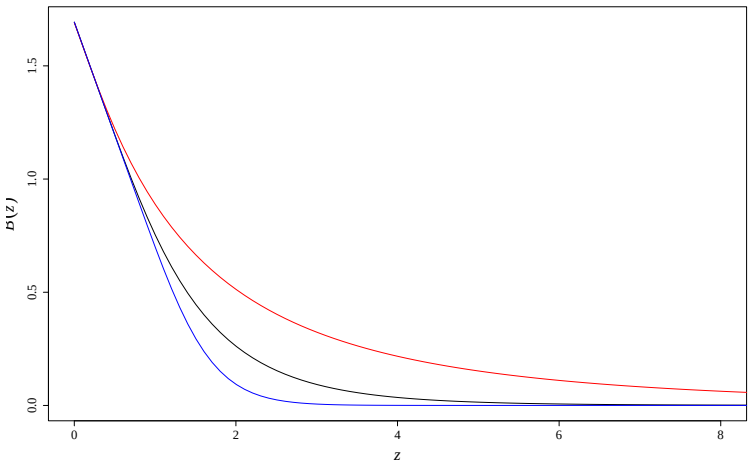

Figure 5.2: Comparison of conventional income curves expected for different supports: from top to bottom, variances of $4, 1$ and $1/4$.

average value in a block should reproduce what is observed in nature. For instance, the average permeability of a block should be defined as the value which gives exactly the same results in a flow simulation than the true heterogeneous block. In two-dimensional problems, for instance, it has been proven that this *representative average* is the geometric mean [Samper-Calvete and Carrera-Ramírez, 1990, Chilès and Delfiner, 1999, p. 598-599].

Once we put an Euclidean structure on the positive real line, the representative value of the block RF may be defined in terms of the coordinates, which means that instead of (3.31), we will work with the RF

$$\zeta_v(\vec{x}) = \int p_v(\vec{h})\zeta(\vec{x}+\vec{h})d\vec{h}, \tag{5.7}$$

defined on the coordinates $\zeta(\vec{x}) = \log \mathbf{Z}(\vec{x})$, and using the same sampling function introduced in section 3.5. Pay attention to the fact that the exponential of (5.7) corresponds to a weighted geometric mean of the point RF in the block, as was said to be advisable for permeability.

By definition, if $\mathbf{Z}$ follows a normal distribution on $\mathbb{R}_+$, then ζ is normally distributed, and consequently ζ_v too, in accordance with the fact that both ζ and ζ_v are Gaussian RFs. In this case, equation (5.7) defines a valid change-of-support model.

Property 5.2 *The RF (5.7) satisfies the conditions stated in property 3.5 (in the* addendum *of chapter 3).*

Proof

1. Both RFs have the same mean,

$$\mathrm{E}_+\left[\mathbf{Z}(\vec{x})\right] = \exp\left(\mu\right) = \exp\left(\mu_v\right) = \mathrm{E}_+\left[\mathbf{Z}_v(\vec{x})\right],$$

since $\mu = \mu_v$ is the mean of a Gaussian RF, which was already shown to be invariant by change-of-support in section 3.5;

2. Their variances and covariance functions are related through (3.33), which is satisfied directly due to the fact that covariances and variograms are defined in coordinates, and the coordinate of a normal on $\mathbb{R}_+$ RF is a Gaussian RF in $\mathbb{R}$,

3. The distribution of the point-support RF is more selective than that of the block-support RF, since they fulfill Cartier's relationship (3.38), which in $\mathbb{R}_+$ is

$$\mathrm{E}_+ [\mathbf{Z}|\mathbf{Z}_v] = \exp(\mathrm{E}[\zeta|\zeta_v]) = \exp(\zeta_v) = \mathrm{E}_+ [\mathbf{Z}_v];$$

the second equality holds due to the fact that Cartier's relation is always fulfilled by a Gaussian RF under the affine correction, as is applicable to the coordinates ζ and ζ_v. The other two equalities are simply the definition (2.6) of expectation in $\mathbb{R}_+$, as is expressed in equation (5.4).

□

It is very important to note that classical lognormal change-of-support models are approximations, whereas this normal in $\mathbb{R}_+$ change-of-support model is exact. However, this theoretical fitness has been obtained only for some cases: those where the geometric mean is a characteristic average in a block. This essentially means that both models, the lognormal and the normal in $\mathbb{R}_+$, are possible models to choose after answering the question: *which is the way to compute a characteristic value, according to the scale of my data?*. This issue affects too the selectivity curves, since they summarize some characteristic values of the proportion of blocks in the domain $\mathcal{D}$ with a value of Z above a threshold z, and do it for all possible values of the threshold. These characteristics are respectively

1. the number of selected blocks, which is equivalent to their volume or mass (assuming a constant specific weight and equal block volume),

2. the mass of Z in the selected blocks,

3. the average of Z in the selected blocks,

4. the benefit of exploitation (or the cost of remediation) of removing and processing the selected blocks.

These curves can be viewed as transformations of the distribution of Z. In this case, being integral transformations, they involve the measure assumed for $\mathbb{R}_+$, the sample space of Z. The question may be restated then as: *which is the measure associated to each one of these characteristics?*. And the answers may be the following.

1. The number of selected blocks defines the *total tonnage* to be extracted from the domain, made relative to the total tonnage of the whole domain; if we represent

by N the total number of blocks in $\mathcal{D}$, and by $N(Z > z)$ the number of selected blocks, the total tonnage is

$$T(z) = \frac{N(Z > z)}{N} = \frac{\sum_{n=1}^{N} I\{Z > z\}}{N} = \int_{Z}^{+\infty} dF(Z)$$

which finally does not depend on the chosen measure for $\mathbb{R}_+$; in fact, $F(z)$ itself acts as this measure; seen from that point of view, the total tonnage is the *measure* of the selected part of the domain. In the WWTP example, $T(z)$ would be proportional to the dumped liters of polluted water.

2. The mass of Z in the selected blocks defines the *quantity of metal*, again relative to the mass of metal in the whole domain; this property is additive in nature,

$$Q(z) = \int_{z}^{+\infty} Z \cdot dF(Z) = \int_{z}^{+\infty} Z \cdot f(Z) \cdot dZ,$$

thus depending on an additive structure. Consequently, here we should use a classical Lebesgue measure for Z. In the WWTP example, this would be the mass of ammonia dumped to the river only *during the hazardous events*.

3. The average of Z in the selected blocks define the *mean grade*, which depends on the measure selected for the space; this can be seen from definition 5.1.3 itself, which involve an expectation; its definition from an $\mathbb{R}_+$-Euclidean point of view yields

$$\tilde{m}_+(\mathbf{z}) = \mathrm{E}_+ \left[\mathbf{Z} | \mathbf{Z} > \mathbf{z}\right] = \frac{1}{T(\mathbf{z})} \int_{\mathbf{z}}^{+\infty} \log \mathbf{Z} \cdot f(\log \mathbf{Z}) \cdot d \log \mathbf{Z},$$

which is different from the classical definition (5.5) of mean grade; in particular, it has no longer a relationship with the quantity of metal above the threshold, since mass ($Z \in \mathbb{R}$) is assumed additive, whereas grade ($\mathbf{Z} \in \mathbb{R}_+$) is assumed multiplicative. Note that we have made explicit this difference by using in the mean grade definition a boldface character, the notation used when $\mathbf{Z} \in \mathbb{R}_+$ considered as an Euclidean space, and a regular character $Z \in \mathbb{R}_+ \subset \mathbb{R}$ when taking the positive real line as a subset of $\mathbb{R}$. In the WWTP example, this would be the characteristic mean concentratio of ammonia dumped to the river only *during the hazardous events*.

4. The benefit of exploitation of the selected blocks gives the *conventional income*, which is a monetary measure; benefits and costs follow an additive scale, thus the measure to use in this case should be again the Lebesgue measure on $\mathbb{R}$. the expression of conventional income (5.6) remains consequently the same, as its interpretation.

Of the four selectivity curves, one (tonnage) is not measure-dependent, two (quantity of metal and conventional income) should be constructed with a classical Lebesgue measure on $\mathbb{R}$, and one (grade) accept a dual approach, according to which is the characteristic average: an arithmetic mean (in the lognormal case) or a geometric one (in normal kriging in $\mathbb{R}_+$ case). These arguments here are nevertheless a first approximation to the subject. Further understanding of these considerations on selectivity curves is left for future work.

5.5 Case study: ammonia pollution risk

The concepts introduced in this chapter are here illustrated by using the ammonia system variables in the Gualba data set for the months of July 2002 and July 2003: acidity constant, acidity conditions, ammonia content and ammonium content. Recall that this data set is not fully observed: in fact, ammonia content is not present, due to its inherent sampling difficulties, and must be computed using equation (1.3). This system has been described in section 2.5.2, where a classical (non-geostatistical) analysis was applied, which took into account the positive nature of the data set by taking the potentials of these variables and working on the real random vector $\underline{Z} = (\zeta_1, \zeta_2, \zeta_3, \zeta_4) = (p\mathrm{H}, p\mathrm{K}_a, p\mathrm{NH}_4, p\mathrm{NH}_3)$. Its time dependence, specially the $24h$ periodic drift, is nevertheless self-evident in a time series plot (figure 5.3). Apart, the three observed components of $\underline{Z}$ (pH, $p\mathrm{K}_a$ and $p\mathrm{NH}_4$) are neither equally nor regularly sampled.

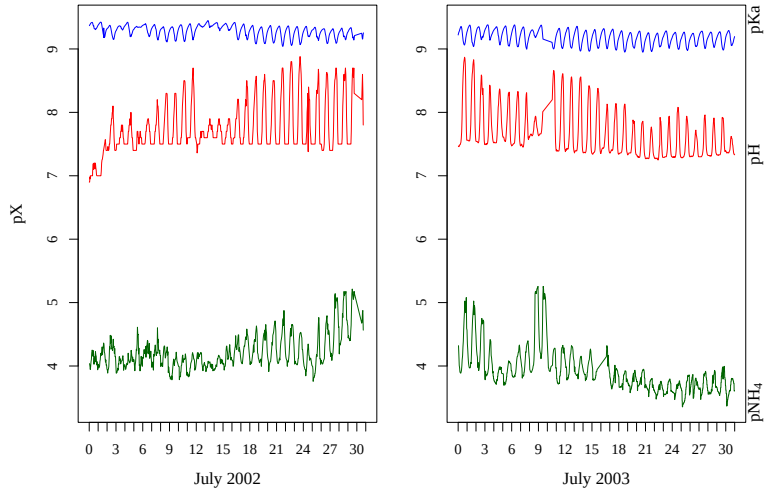

Figure 5.3: Time series of observed potential ammonium, potential hydrogen, and potential acidity constant.

As was explained in section 1.4.1, the drift of the full Gualba system is assumed to be controlled by the solar radiation dynamics, which were characterized through

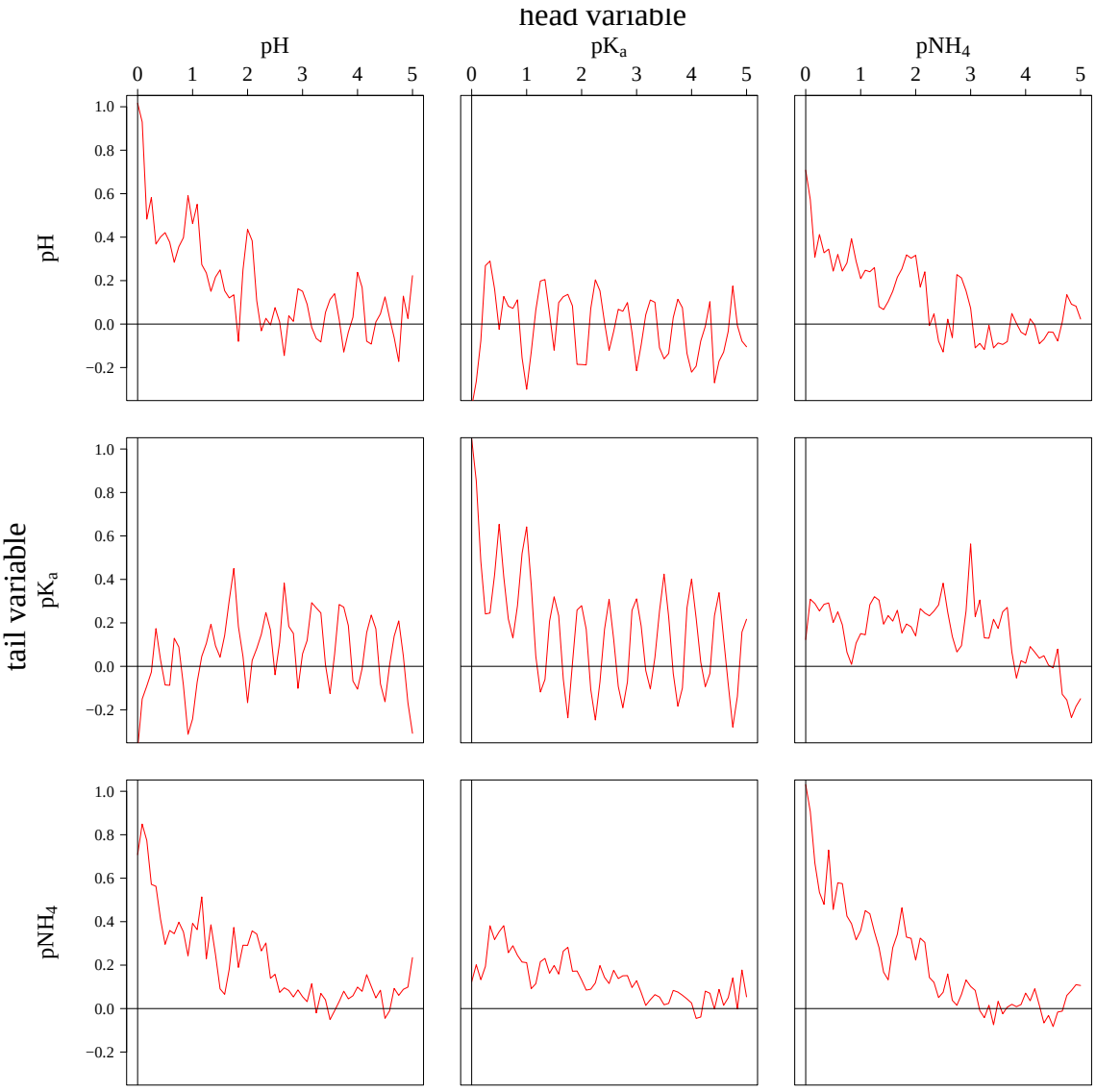

Figure 5.4: Experimental auto- and cross-correlation functions at short range. All plots share the same vertical scale.

Table 5.1: Fitted drift coefficients using classical regression for potential acidity constant ($p\mathrm{K}_a$), potential hydrogen ($p\mathrm{H}$) and potential ammonium ($p\mathrm{NH}_4$). The upper table recalls the periods obtained from the Fourier analysis of section 3.6, although here we have only used the first four and a constant term.

index	i	1	2	3	4	5	6	7
period(days)	τ_i	1	2.5	10	25	42	100	365

i	a_{2i-1}	a_{2i}	m_{2i}	α_{2i}
$p\mathrm{H}$ during July 2002				
0	0	7.75	7.75	0
1	-0.28	-0.31	0.42	-0.38
2	0.03	-0.05	0.05	1.04
3	-0.05	-0.02	0.06	-3.01
4	0.01	0.08	0.08	0.70
$p\mathrm{K}_a$ during July 2002				
0	0	9.28	9.28	0
1	0.03	0.10	0.11	0.04
2	-0.01	0	0.01	-0.64
3	0.01	0.03	0.03	0.37
4	0.04	-0.01	0.04	7.54
$p\mathrm{NH}_4$ during July 2002				
0	0	4.19	4.19	0
1	-0.13	-0.17	0.22	-0.40
2	0	0	0	0.45
3	-0.09	-0.12	0.14	-3.97
4	-0.14	0.01	0.14	-6.09

i	a_{2i-1}	a_{2i}	m_{2i}	α_{2i}
$p\mathrm{H}$ during July 2003				
0	0	7.66	7.66	0
1	-0.20	-0.32	0.38	-0.41
2	0	0.01	0.01	-0.29
3	-0.07	-0.03	0.08	-3.18
4	-0.15	0	0.15	-6.26
$p\mathrm{K}_a$ during July 2003				
0	0	9.17	9.17	0
1	0.02	0.13	0.13	0.02
2	0	0	0.01	-0.85
3	-0.01	-0.01	0.01	-3.61
4	-0.03	0.02	0.04	-3.80
$p\mathrm{NH}_4$ during July 2003				
0	0	3.87	3.87	0
1	-0.01	-0.17	0.17	-0.49
2	0.01	-0.01	0.02	0.89
3	-0.04	0.01	0.05	-2.00
4	-0.19	-0.05	0.19	-7.28

Fourier analysis of water temperature frequency spectrum (figure 3.6) in section 3.6. We take here the same trigonometric functions (3.36) to explain the drift of $\underline{Z}$ by using classical regression with uncorrelated residuals (eq. 3.20). The obtained coefficients are listed in table 5.1.

Note that strictly speaking, this regression method is not applicable here, because residuals are correlated (5.4). However, the experience and arguments developed in section 3.6, regarding the analysis of conductivity on the same control station, suggest that the approximation will be satisfactory. Therefore, the resulting regression was taken as the drift. This drift was subtracted from the observed data, and the residuals were used to compute the auto- and cross-covariance functions between pH, $p\mathrm{K}_a$ and $p\mathrm{NH}_4$, using equation (3.4), taking into account that the mean of all these residuals can be considered zero. Figure 5.4 shows the estimated correlation functions, without any assumption of symmetry: it is interesting to note the strong cross-correlation of pH and $p\mathrm{NH}_4$, and the periodic hole effect present in both pH and $p\mathrm{K}_a$. This figure shows also that there is no need to consider the covariance structure to be asymmetric. Attending to these facts, the covariances have been modeled with a linear model of coregionalization, displayed in figures 5.5 (short range) and 5.6 (long range) jointly with the estimated covariances, and corresponding to the following composite model:

$$
\begin{aligned}
\underline{\underline{C}}(h) &= \begin{pmatrix} \mathrm{Var}\,[p\mathrm{H}] & \mathrm{Cov}\,[p\mathrm{H}, p\mathrm{K}_a] & \mathrm{Cov}\,[p\mathrm{H}, p\mathrm{NH}_4] \\ \mathrm{Cov}\,[p\mathrm{K}_a, p\mathrm{H}] & \mathrm{Var}\,[p\mathrm{K}_a] & \mathrm{Cov}\,[p\mathrm{K}_a, p\mathrm{NH}_4] \\ \mathrm{Cov}\,[p\mathrm{NH}_4, p\mathrm{H}] & \mathrm{Cov}\,[p\mathrm{NH}_4, p\mathrm{K}_a] & \mathrm{Var}\,[p\mathrm{NH}_4] \end{pmatrix} = \\
&= \begin{pmatrix} 0.04 & 0 & 0.01 \\ 0 & 0.00125 & 0 \\ 0.01 & 0 & 0.022 \end{pmatrix} \cdot \mathrm{Exp}(\mathrm{a}=2) + \\
&+ \begin{pmatrix} 0.0075 & 0 & 0.0045 \\ 0 & 1e-06 & 0 \\ 0.0045 & 0 & 0.008 \end{pmatrix} \cdot \mathrm{Hol}(\mathrm{a_t}=7, \mathrm{a_d}=\infty) + \\
&+ \begin{pmatrix} 0.01 & 0 & 0.0065 \\ 0 & 0.00025 & 0 \\ 0.0065 & 0 & 0.007 \end{pmatrix} \cdot \mathrm{Hol}(\mathrm{a_t}=1, \mathrm{a_d}=\infty) + \\
&+ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 5e-04 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \mathrm{Hol}(\mathrm{a_t}=0.5, \mathrm{a_d}=\infty)
\end{aligned}
$$

Note the four components considered: an exponential covariance of $a = 2$ days (6-days effective range), and three non-dampened hole effects of periods: a week, a day and half a day. Note also that this last half-day-period is only present in the $p\mathrm{K}_a$ series, but it is its most important structure. Recall that the lineal model of corregionalization is only valid if the matrices involved are symmetric positive (semi-) definite, as is in our case.

With this covariance model we may attempt prediction of the value of the vector $\mathbf{z}$ at each full hour. From this vector, we are particularly interested in the fourth

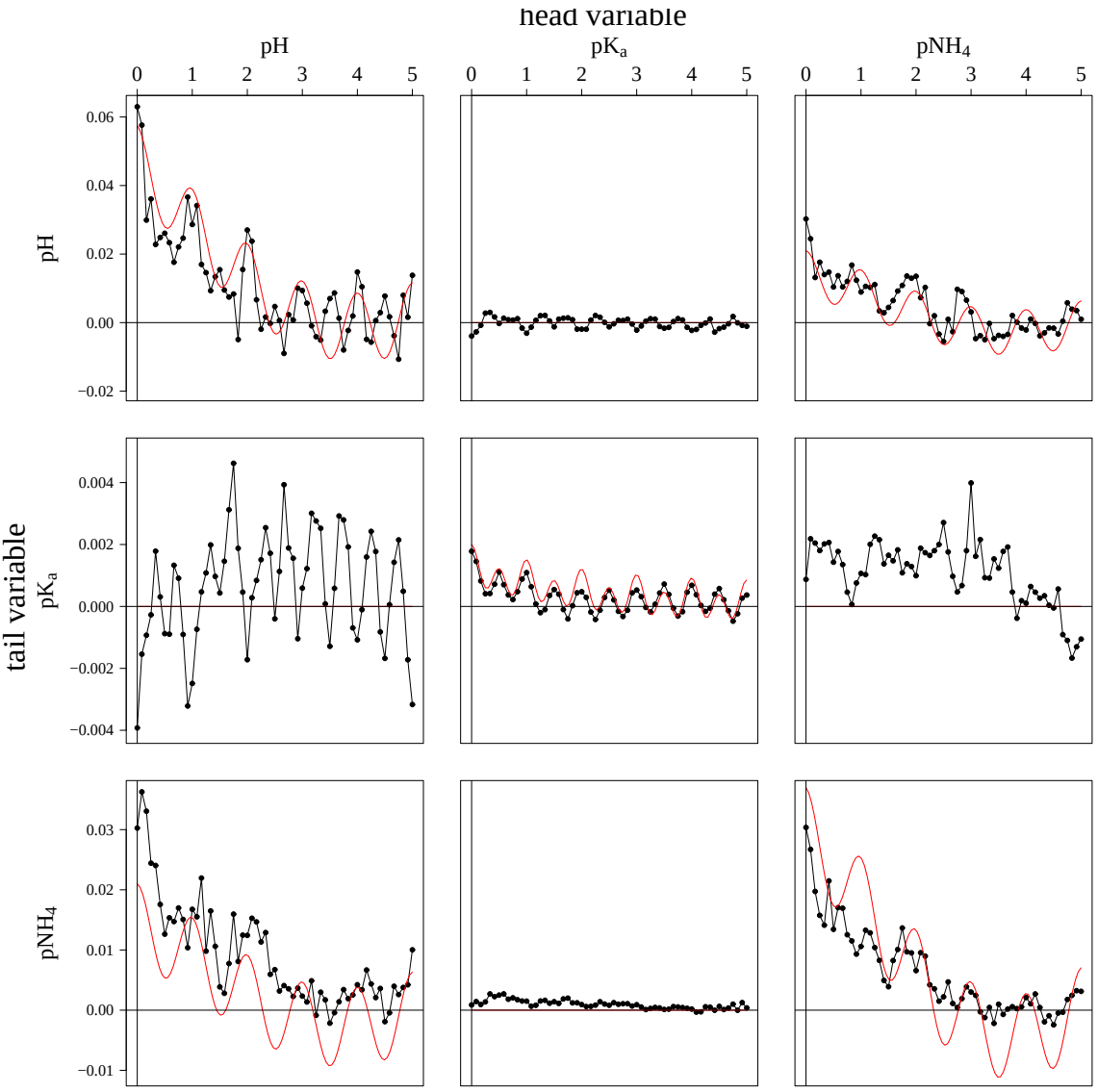

Figure 5.5: Experimental (dots and black line) and fitted (red line) auto- and cross-covariance functions at short range. All plots in a row share the same vertical scale.

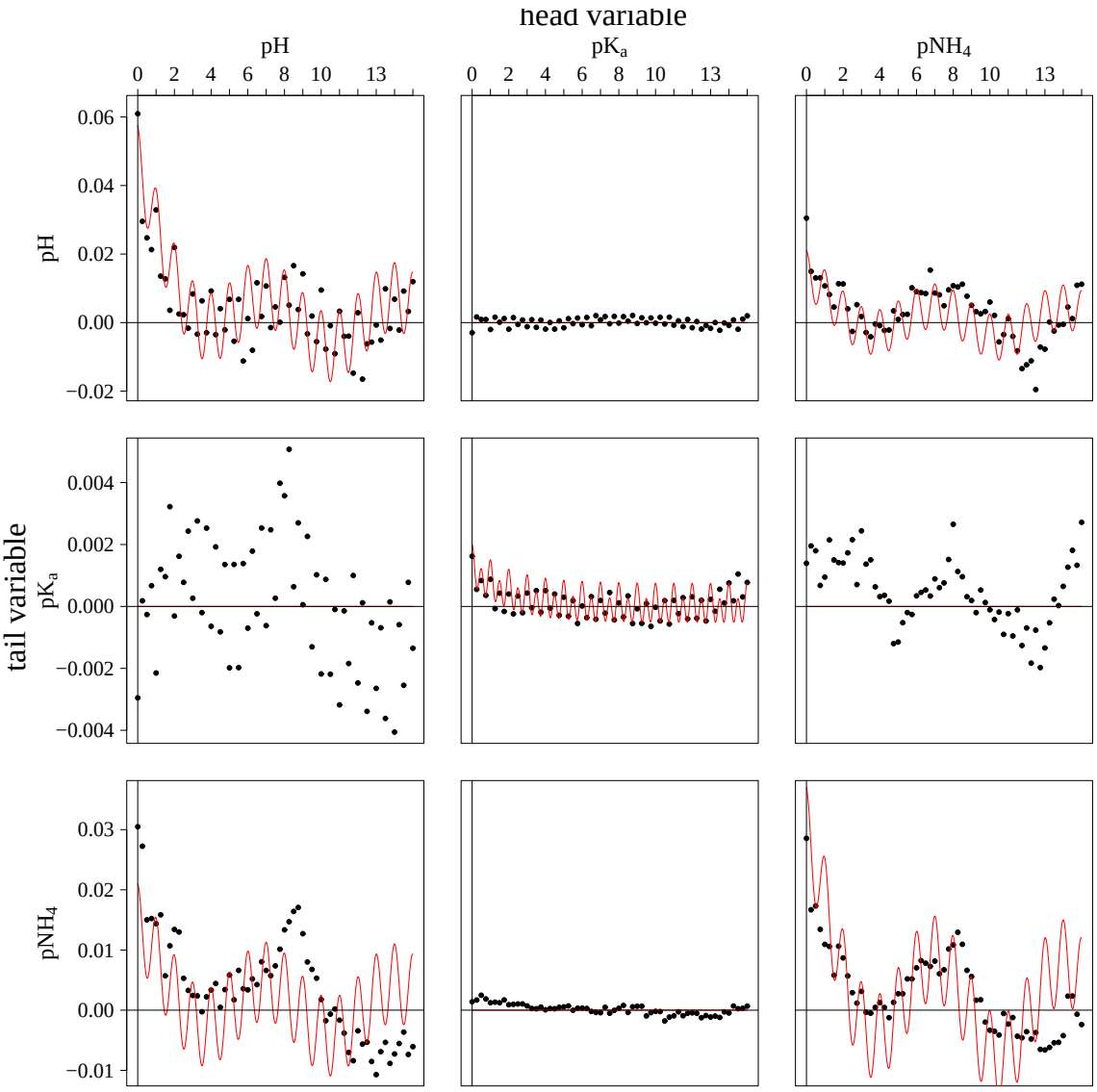

Figure 5.6: Experimental (dots) and fitted (line) auto- and cross-covariance functions at long range. Note that all plots in a row share the same vertical scale.

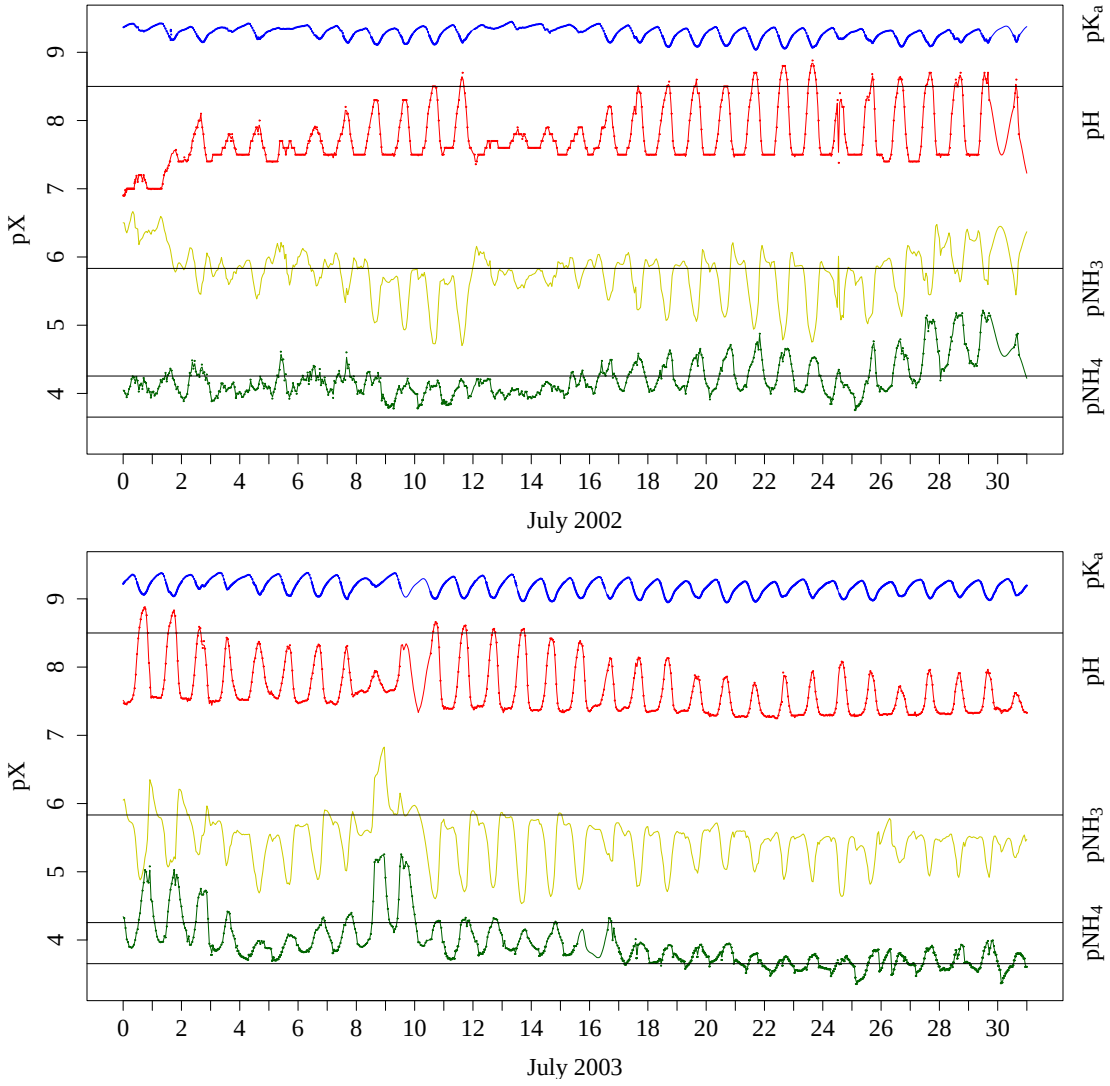

Figure 5.7: Predicted coordinates of the vector $\mathbf{z}(t_0)$ for each predicted time moment using simple kriging, between the beginning of July 1 and the end of July 31, both for the years 2002 (top) and 2003 (bottom). In each plot, from top to bottom, the series are: $\zeta_2 = p\mathrm{K}_a$, $\zeta_1 = p\mathrm{H}$ (with a reference level at $p\mathrm{H} = 8.5$), $\zeta_4 = p\mathrm{NH}_3$ (with a reference level corresponding to 0.025 mg/l of NH_3), and $\zeta_3 = p\mathrm{NH}_4$ (with two reference levels of 1mg/l and 4 mg/l of NH_4^+). Note that, due to the definition of the coordinate as a potential ($pX = -\log_{10}[X]$), an *increase* in the value of pX represents a *decrease* in the concentration of X.

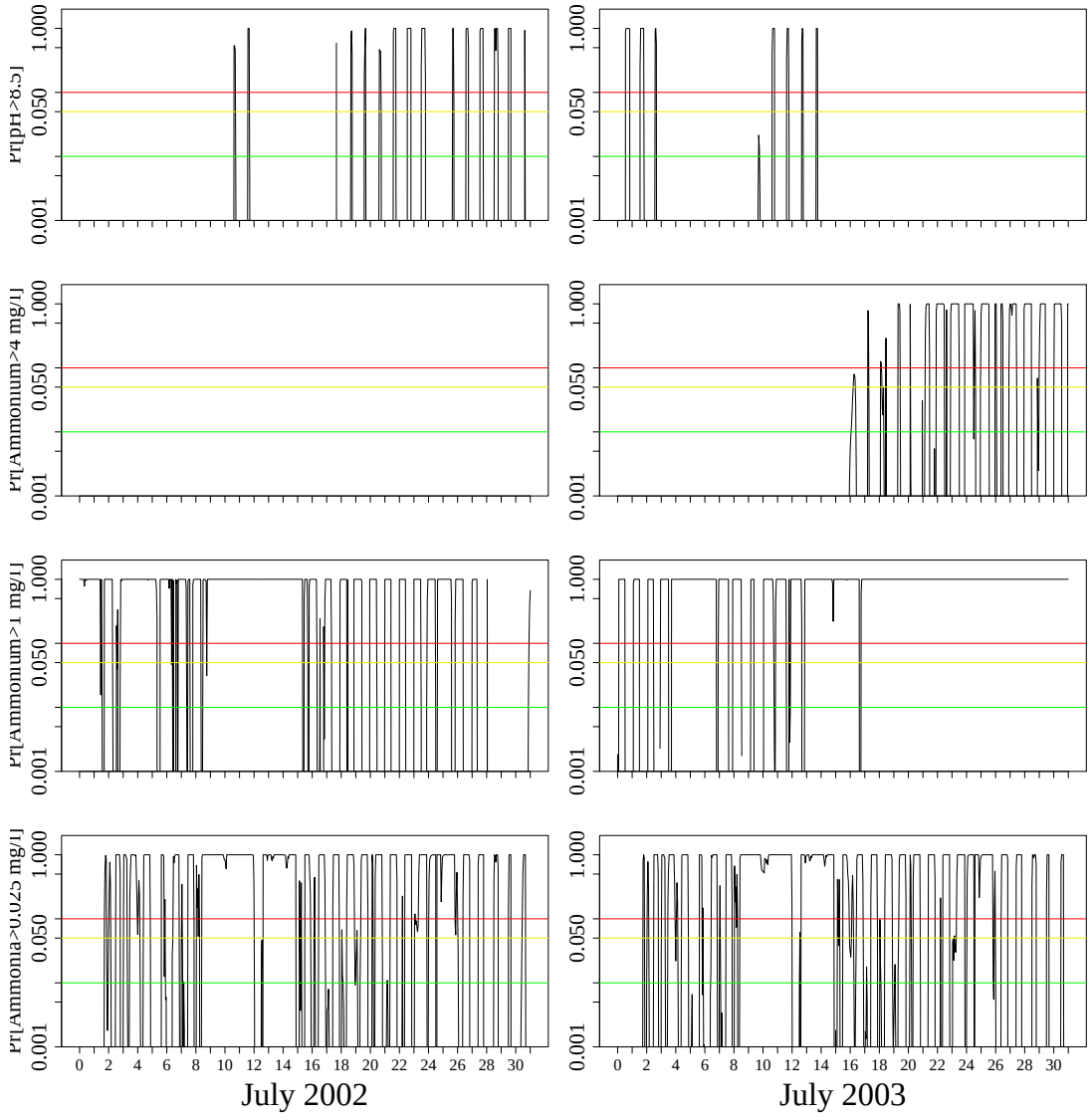

Figure 5.8: Log-hazard of exceeding each of the thresholds defining the water quality categories of table 1.1. Values are computed at each hour, between the beginning of July 1 and the end of July 31, both for the years 2002 (left) and 2003 (right). In each of these plots, three reference lines mark the probability levels of 0.01, 0.05 and 0.10.

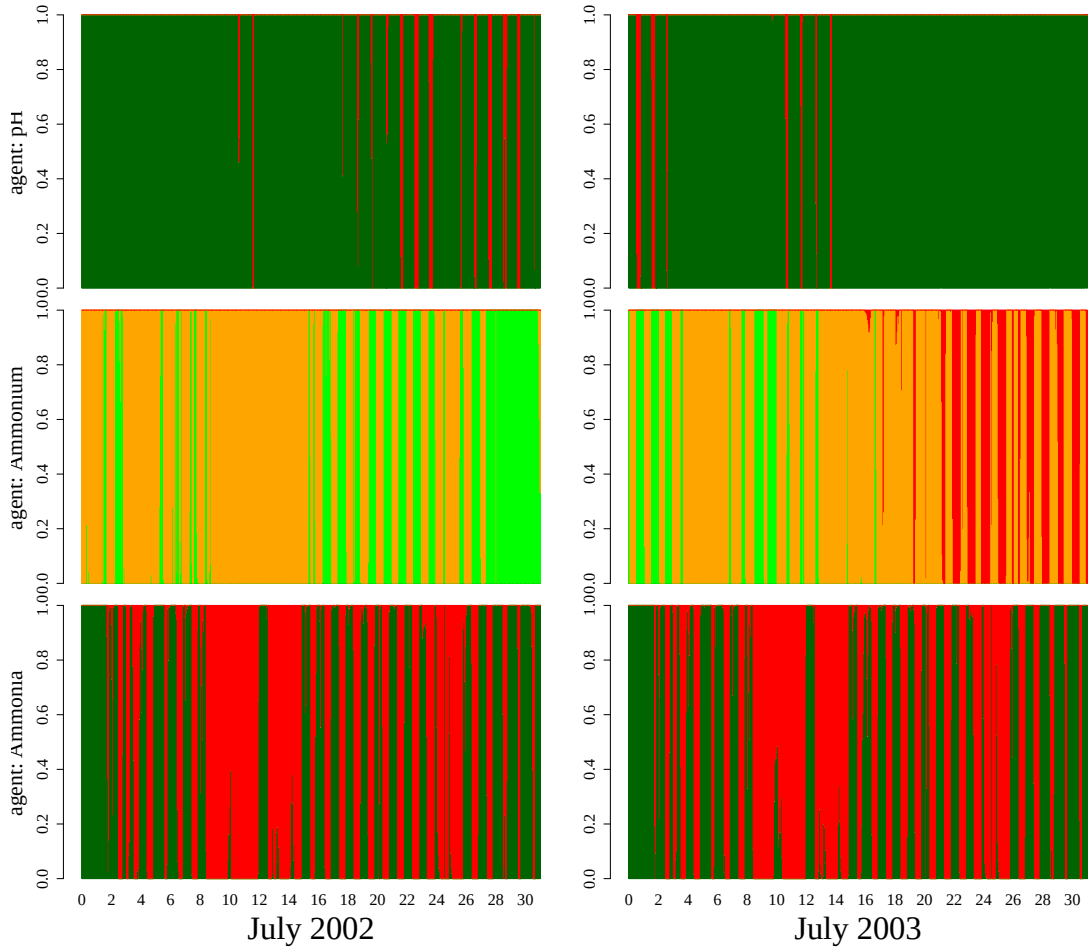

Figure 5.9: Probability of being in each category of water quality, as a function of a single component: $\zeta_1 = p\text{H}$ (top), $z_3 = \left[\text{NH}_4^+\right]$ (middle) and $z_4 = [\text{NH}_3]$ (bottom), computed for all hours between the beginning of July 1 and the end of July 31, both for 2002 (left) and 2003 (right). The colors and reference levels are summarized in the following table:

			color		categories	
plot	variable	reference level	below	above	below	above
top	pH	8.5			1,2,3	4
middle	$\left[\text{NH}_4^+\right]$	1 mg/l			1,2	3,4
middle	$\left[\text{NH}_4^+\right]$	4 mg/l			1,2,3	4
bottom	$[\text{NH}_3]$	0.025 mg/l			(admissible)	(inadmissible)

coordinate, $\zeta_4 = p\mathrm{NH}_3$. Though not observed, this variable is deterministically related to the observed ones through a linear equation (1.3), of the general form

$$\zeta_4(t_0) = \alpha_1 \cdot \zeta_1(t_0) + \alpha_2 \cdot \zeta_2(t_0) + \alpha_3 \cdot \zeta_3(t_0),$$

where $(\alpha_1, \alpha_2, \alpha_3) = (-1, 1, 1)$. Since no simultaneous measure of the three real variables $(\zeta_1, \zeta_2, \zeta_3)$ is available, we will predict at each moment the residual of potential ammonia replacing the other residual values by their simple kriging predictors, available in function `predict.gstat` of the `gstat` [Pebesma and Wesseling, 1998] package for `R` [R Development Core Team, 2004].

The obtained kriging predictions are displayed in figure 5.7. This figure displays also some reference levels extracted from table 1.1. Recall that we want to compute the probability that hourly averages of the chemical variables fall in certain fields of their domain. We have to assume then a change-of-support model, from which we choose the normal in $\mathbb{R}_+$ one. Therefore, using the obtained predictions and their kriging variances as moments of the Normal on $\mathbb{R}_+$ distributions of the true $\mathbf{Z}(t_0)$ for each predicted t_0, we may compute the probability of exceeding each of the interesting thresholds (figure 5.8), and finally the sought probability of being in each category from those defined in table 1.1 (figure 5.9).

The impact of the daily influence on all these results is evident. Figure 5.7 shows for instance that Ammonium tends to be maximal at noon ($p\mathrm{NH}_4$ is minimal) and minimal at midnight, whereas Ammonia behavior is usually the opposite. Ammonium content is classically used as a proxy for Ammonia content (which is not directly measured due to sampling problems), since accordingly with equation (1.3) they are proportionally related. However, in a river like the studied one, pH (maximal after midnight also) and $p\mathrm{K}_a$ (minimal after midnight) fluctuations are much more regular, stronger, and both favor the ammonia behavior. Consequently, we see that Ammonium daily fluctuations are mainly produced by fluctuations of pH and temperature, which propagate to $p\mathrm{NH}_4$ due to the equilibrium reaction (1.1), and not by direct dumping of NH_4^+ to the river. This has a very important management implication: Ammonium is not enough in this river to account for Ammonia content, and pH variations become an essential information.

Probability oscillations show also this strong daily control: in all plots of figure 5.8 we almost always see sharp pulses with a 1 day-periodicity. There are very few situations in which the probability of exceeding each threshold is between the reference levels of 0.01 and 0.1. This strong certainty is also evident in figure 5.9. This is an effect of the small kriging variances, something related to the high density of the sampling.

Finally, a last word is due with regard to the choice of a geometric mean as characteristic average. One might argue that the total mass of Ammonia flowing through the station must be measured in an additive scale, and thus we should use a lognormal approach. However, interest lies on the effects that this pollutant might have on living beings, and this is related to its chemical behavior. Some points suggest that the choice of a normal on $\mathbb{R}_+$ change-of-support model (thus, a weighted geometric mean) might be better in these conditions:

- the river is an open system, and Ammonia is an extremely volatile compound; thus, it is expected that its mass will leave the system;
- the amount of Ammonia leaving the system will depend on its equilibrium with pH, which we know to be linear in terms of the species involved in the reaction (equation 1.3),
- linearity in the logarithms of $\mathbf{Z}$ imply that the model for $\mathbf{Z}$ itself is a multiplicative one.

However, these are suggestions, and not true arguments. In absence of a better model, both the arithmetic and the geometric approach might be possible.

5.6 Remarks

Positive variables—either in classical or regionalized problems—are usually treated through a logarithm, which is expected to yield unbounded scores with a higher symmetry, and a better fit to a normal model. This simple procedure defines the log-normal distribution, and in the geostatistical case, a log-normal RF. Usually, the mean of this distribution—or this RF—is estimated as the exponent of the transformed mean, multiplied by a correction factor which depends on the variance and accounts for the different measures considered in these spaces. Such a correction has been related in this chapter to some problems in the behavior of these estimators, specially in the geostatistical case: the local conditional expectation character of predictions is unclear, and change-of-support models and block kriging are theoretically inconsistent.

However, this logarithm is not an arbitrary transformation, but a decision on the scale to be used to compare positive observations. Furthermore, it coincides with the way to compute the coordinates of the observations with respect to a basis of the positive real line regarded as an Euclidean space on its own. Such a natural coordinate system arises in the definition of pH, and has some algebraic advantages, *e.g.*it linearizes equilibrium equations. In this Euclidean space structure, we used the normal distribution on $\mathbb{R}_+$, an alternative model to the log-normal one, to naturally define normal RFs on $\mathbb{R}_+$. Estimation and prediction in these RFs is theoretically robust against departures of the normality on $\mathbb{R}_+$ of the data. Moreover, it yields a valid change-of-support model, once the conditions used to validate a model are adequately interpreted.

Taking into account this natural structure and the expected time dependence, we treated the Gualba Ammonia series already introduced in section 1.4.1, and described from an univariate point of view in section 2.5.2. In that section, we derived a global probability of 0.6 of being above the threshold of 0.025 mg NH_3/l, something considered as too much in that basin, with a relatively low pollution history. Taking into account clear periodic drifts and a strong auto-correlation structure underlying the series of Ammonium content, pH and water temperature (and its surrogate, the acidity

constant potential, $p\mathrm{K}_a$), we predicted the Ammonia content, and saw that Ammonium fluctuations might not be due to the human action, but to fluctuations of pH. According to Tolosana-Delgado [2004], these fluctuations are probably linked to the photosynthesis-respiration processes in the river.

Here, the final interest was the assessment of water quality, and not the causes of these fluctuations. Roughly, the predictions of actual concentration of Ammonia suggest that water quality is better during the night, but becomes poorer during daylight. To complement this, we computed the probability of being above each threshold of interest for each variable used. This was achieved thanks to the fact that kriging in $\mathbb{R}_+$ predictor and its kriging variance are the moments of the normal on $\mathbb{R}_+$ distribution describing the conditional expectation of the predicted variables. These results mainly showed a strong daily drift, and a clear definition of each category: almost always, one of the categories was highly sure, thus reducing the usefulness of this probabilistic approach. It is nevertheless expected to be more useful in situations with higher kriging variance.

These results were nevertheless only obtained after discussing two different geometries for the Ammonia concentration in the river. Both the classical arithmetic average and the suggested geometric average of measurements in a block were considered. The second one was chosen, due to the characteristics of the problem: mainly, the fact that in an open system (as is a river with Ammonia, which is highly volatile) mass is not conservative, and therefore there are no arguments in favor of using an additive scale. In contrast, a multiplicative scale is consistent with the equilibrium reaction governing Ammonia stability in the dissolution. These considerations on which is the underlying process should be always done before choosing a characteristic average.

Chapter 6

Geostatistics in the Simplex

6.1 Kriging of compositions

Let $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$ be a point (or the center of a block v) in a domain $\mathcal{D}$ of the space-time real space, with $p \in \{1, 2, 3, 4\}$. Let $\mathbf{Z}(\vec{x}) = (Z_1(\vec{x}), Z_2(\vec{x}), \ldots, Z_D(\vec{x})) \in \mathbb{S}^D$ be a D-part compositional random function (RF), and call Z_i the i-th *component* of the composition. Due to the nature of compositional data, at any location $\vec{x}$ all the components are non-negative and *closed*,

$$Z_i(\vec{x}) \geq 0 \quad \text{and} \quad \sum_{i=1}^{D} Z_i(\vec{x}) = \kappa, \tag{6.1}$$

with κ equal to the value of the whole composition. Usually, $\kappa = 100$ (percentages) or $\kappa = 1$ (proportions). For the sake of simplicity, we work with the latter, without loss of generality. Note that the concept of compositional RF we use here is exactly equivalent to the regionalized compositions of Pawlowsky-Glahn and Olea [2004].

Let $\mathbf{z}(\vec{x}_1), \mathbf{z}(\vec{x}_2), \ldots, \mathbf{z}(\vec{x}_N)$ be an observed sample of this RF, which is assumed to be fully observed. The goal will be the estimation of the composition $\mathbf{Z}(\vec{x}_0)$ at an unsampled location, and of its error variance-covariance matrix. As happened with positive RFs (chapter 5), the fact that compositions are constrained should discourage us to assume $\mathbf{Z}(\vec{x})$ to be a Gaussian RF, which is defined on the whole real space $\mathbb{R}^D$. Moreover, it is well-known that classical statistical methods based on the covariance matrix of compositional data sets cannot be interpreted in the usual way due to a negative bias introduced by the closure in the covariances [Chayes, 1960, 1971]. Although geostatistical methods are also based on the covariance matrix [Pawlowsky-Glahn, 1984], today most of the references found in the literature on structural analysis and kriging of compositional data simply do not consider the problem, *e.g.*Wackernagel [1998]; an extensive list may be found in Pawlowsky-Glahn and Olea [2004]. It is argued that when the total sum of the interesting/available variables is far from one, the *closure* has no noticeable effect, but the fact is that it might yield non-sense estimations, *e.g.*negative or not summing up to one [Isaaks and Srivastava, 1989, Pawlowsky

et al., 1994, 1996]. Furthermore, the addition of any non-relevant variable or the forced closure of the system can lead to totally inconsistent conclusions. Two different and incompatible approaches have been developed to deal with this: transforming the data set, and constraining the admissible solutions.

Walwoort and de Gruijter [2001] propose a method, based on constraining the predictions of D *simultaneous kriging systems* (3.16) to sum up to one and have positive values. This is achieved first by adding to the universality conditions (3.18) a further constraint with the constant sum condition (6.1). Second, the positivity of the components is ensured by using the concept of active constraint [Wismer and Chattergy, 1978]. This is a further non-linear constraint which acts only when the linear kriging prediction would be negative. In this case, this new constraint simply ensures that the negative predictions will be zero, and consequently changes the kriging variance. The authors warn that as a result, kriging offers no more an estimate of the conditional distribution. This solution is a non-linear interpolation procedure, which may be difficult to compare with a linear one. In [Walwoort and de Gruijter, 2001] opinion, it should be taken simply as an interpolation technique.

Based on the characterization of the Simplex by Aitchison [1986, also detailed in section 2.5.3], Pawlowsky [1986] developed a transformation strategy, which relies on similar ideas to those used in lognormal kriging. First, the data set is transformed to a set of real vectors $\left\{\underline{\zeta}(\vec{x}_n)\right\}$; afterwards geostatistical techniques are applied to the transformed scores to obtain, *e.g.*a prediction at an unsampled location $\underline{\zeta}^*(\vec{x}_0)$; finally the kriging prediction is back-transformed to obtain a prediction $\mathbf{z}^*(\vec{x}_0)$ of the composition itself at the unsampled location. The transformations studied by Pawlowsky [1986] and Pawlowsky-Glahn and Olea [2004] are three:

- the centered log-ratio transformation [Aitchison, 1986, clr], which maps $\mathbb{S}^D$, the Simplex of D-parts, into an hyperplane of $\mathbb{R}^D$ through

$$\zeta_i = \mathrm{clr}_i(\mathbf{Z}) = \log \frac{Z_i}{g(\mathbf{Z})}, \quad i \in \{1, 2, \ldots, D\}, \tag{6.2}$$

  where $g(\mathbf{Z})$ represent the geometric mean of all components of $\mathbf{Z}$. The fact that the image space of this mapping is again constrained hinders its use in geostatistical applications, essentially due to the singularity of the resulting kriging system matrix; consequently, Pawlowsky-Glahn and Olea [2004] discouraged its use;

- the additive log-ratio transformation [Aitchison, 1986, alr], a mapping between $\mathbb{S}^D$ and the space $\mathbb{R}^{D-1}$ through

$$\zeta_i = \mathrm{alr}_i(\mathbf{Z}) = \log \frac{Z_i}{Z_D}, \quad i \in \{1, 2, \ldots, D-1\}, \tag{6.3}$$

  is presented as the most-adequate, due to its unconstrained character. Following the procedure of lognormal kriging, we can assume the RF to follow an *additive logistic normal model*, and obtain an unbiased predictor for the composition

by using in a direct way a numerical Hermitian integration method [Aitchison, 1986], which should be consequently called *additive logistic normal kriging* (ALN-kriging). However, this predictor is seldom used in practice, due to the complication of such a numerical integration;

- the basis method, which can be applied only when an *independent* auxiliary RF $t(\vec{x})$ representing the size of the composition is available. Then the product $t(\vec{x}) \cdot Z_i(\vec{x})$, called a *basis*, defines a strictly positive RF, which should be treated with the techniques introduced in chapter 5. Note that the concept of basis is here different from the algebraic-geometric concept introduced in chapter 2.

The clear connection of these methods with lognormal kriging may help us to understand them and their limitations. Again, the numerically-obtained ALN-kriging predictor is considered optimal, as the lognormal predictor was. However, the same family of problems found on lognormal kriging are expected for ALN-kriging. In particular, the conclusion of Roth [1998] about the inadequateness of lognormal kriging for local conditional expectation is also applicable here. To construct confidence regions, one can attempt to obtain also an estimate of the covariance matrix of $\mathbf{Z}$ by using again Hermitian numerical integration; afterwards, this covariance matrix may be used to build D-dimensional confidence regions around the ALN-kriging predictor. However, this confidence region contains mostly values which cannot be compositions, since their components are not necessarily positive, nor sum they up to one [Pawlowsky-Glahn and Olea, 2004]. Following the same reasoning applied to lognormal kriging, it is better to build a confidence region around the prediction $\underline{\zeta}^* = (\zeta_i^*(\vec{x}_0))$ by using the kriging covariance matrix, and afterwards back-transform the obtained region through the *additive generalized logistic transformation*,

$$\mathbf{z} = \mathrm{agl}(\underline{\zeta}) = \mathcal{C}\left(\exp(\zeta_1), \exp(\zeta_2), \ldots, \exp(\zeta_{D-1}), 1\right), \tag{6.4}$$

the inverse transformation to (6.3). Simulation of compositional RFs become also straightforward by simulating $(D-1)$-dimensional real RFs and transforming them to compositions through (6.4). Regarding block kriging, the same problems arising in lognormal kriging around *conservation of lognormality* are also valid here (section 5.4).

Following this strategy, Tjelmeland and Lund [2003] present an implementation of the generalized linear predictor (section 3.4.2) based on the additive-logistic normal distribution. Their work is essentially a straightforward application of model-based geostatistical techniques [Diggle et al., 1998] on the alr-transformed values, where a Gaussian assumption is fully natural and consequently MCMC techniques are expected to be efficient. Afterward, results are back-transformed to compositions through (6.4). The advantage of their approach lies on the ability to process compositions with lost components, and estimate them as well as predict unsampled locations with a bayesian account of their uncertainty. The main disadvantage of this method is the need of the MCMC simulation, which require usually extensive computations.

6.2 Simplex space structure

Following the arguments of section 5.1 regarding the problems of lognormal kriging, it is interesting to consider the Euclidean structure of the Simplex, denoted by $\mathbb{S}^D$, in order to understand the known and expected problems of ALN-kriging and its possible solutions. This structure has been detailed throughout chapter 2, and specially in section 2.5.3. Recall that it is based on the vector space operations called perturbation and power transformation, in the Aitchison distance (2.23) and its associated scalar product. Given an orthonormal basis of $\mathbb{S}^D$, the coefficient vector $\underline{\zeta}$ of any element $\mathbf{z}$ of the space can be computed with

$$\underline{\zeta} = \underline{\underline{\varphi}} \cdot \underline{\mathrm{clr}}(\mathbf{z}) = \underline{\underline{\varphi}} \cdot \underline{\log}(\mathbf{z}), \tag{6.5}$$

where $\underline{\log}(\mathbf{z}) = (\log z_1, \log z_2, \ldots, \log z_D)^t$, and $\underline{\underline{\varphi}}$ is a matrix associated to the used basis, with $(D-1)$ rows by D columns which rows sum up to zero

$$\sum_{j=1}^{D} \varphi_{ij} = 0, \quad i = 1, 2, \ldots, D-1, \tag{6.6}$$

and satisfies a kind of orthogonality conditions, coming from the orthonormality of the basis,

$$\begin{aligned} \underline{\underline{\varphi}} \cdot \underline{\underline{\varphi}}^t &= \underline{\underline{I}}_{(D-1)} \\ \underline{\underline{\varphi}}^t \cdot \underline{\underline{\varphi}} &= \underline{\underline{I}}_D - \frac{1}{D} \cdot \underline{\underline{1}}_D. \end{aligned} \tag{6.7}$$

In these expressions, $\underline{\underline{I}}$ represents the identity matrix and $\underline{\underline{1}}$ a matrix with all elements equal to 1, while the subindexes show their dimension. For these reasons, the coefficients $\underline{\zeta}$ in equation (6.5) are called *isometric log-ratio coordinates* (ilr), and define another log-ratio transformation to use as the clr and alr (6.2-6.3).

In particular, the elements of the basis themselves can be retrieved from matrix $\underline{\underline{\varphi}}$ by taking the closed exponential of each one of its rows $\underline{\varphi}_i$, $i = 1, 2, \ldots, D-1$,

$$\mathbf{e}_i = \mathcal{C}\left(\exp\left(\underline{\varphi}_i\right)\right) = \mathcal{C}\left(\exp\varphi_{i1}, \exp\varphi_{i2}, \ldots, \exp\varphi_{iD}\right). \tag{6.8}$$

With coefficients (6.5), a Lebesgue measure (2.3) can be defined in $\mathbb{S}^D$, which allows to introduce the normal distribution on $\mathbb{S}^D$ as an alternative model to the ALN distribution [Mateu-Figueras et al., 2003]. This normal distribution on $\mathbb{S}^D$ is illustrated in figure 2.7.

In this context, ALN-kriging can be considered to follow a two-step procedure like lognormal kriging:

1. first, estimate the log-ratios of the vector RF at the unsampled location, by minimizing the distance on $\mathbb{S}^D$ between the true value of $\underline{\zeta}(\vec{x}_0)$ and its estimate $\underline{\zeta}^*_K$;

2. second, estimate the compositional RF at the unsampled location $\mathbf{z}(\vec{x}_0)$ as the mean of an ALN distribution, with reference to the Lebesgue measure on $\mathbb{R}^D$, using multivariate Hermite integration.

This is a mixture of two mutually-inconsistent criteria, based respectively on Aitchison distance and Lebesgue measure.

6.3 Structural analysis

Pawlowsky-Glahn and Olea [2004] present four different specifications of the covariance structure of a regionalized composition, plus their corresponding variogram structures.

spatial covariance structure is the basic specification which includes D^4 functions, showing which is the relationship between all the possible log-ratios at the two locations,

$$\sigma_{ij\cdot kl}(\vec{h}) = \mathrm{Cov}\left[\log\left(\frac{z_i(\vec{x})}{z_k(\vec{x})}\right), \log\left(\frac{z_j(\vec{x}+\vec{h})}{z_l(\vec{x}+\vec{h})}\right)\right];$$

this is mostly superfluous, since it can be specified with a small number of functions, without assuming any kind of symmetry;

lr cross-covariance is the $D \times D$ matrix of covariance functions of each log-ratio

$$\tau_{i\cdot j}(\vec{h}) = \mathrm{Cov}\left[\log\left(\frac{z_i(\vec{x})}{z_j(\vec{x})}\right), \log\left(\frac{z_i(\vec{x}+\vec{h})}{z_j(\vec{x}+\vec{h})}\right)\right] = \sigma_{ii\cdot jj}, \qquad \underline{\underline{\mathrm{T}}}(\vec{h}) = \left(\tau_{i\cdot j}(\vec{h})\right),$$

it is characterized by the fact that $\tau_{i\cdot i} = 0$, and it describes the whole covariance structure, given that

$$\sigma_{ij\cdot kl} = \frac{1}{2}\left(\tau_{i\cdot l} + \tau_{j\cdot k} - \tau_{i\cdot j} - \tau_{k\cdot l}\right);$$

clr cross-covariance is the $D \times D$ matrix of cross-covariance functions of the clr-transformed data (6.2),

$$\xi_{i\cdot j}(\vec{h}) = \mathrm{Cov}\left[\log\left(\frac{z_i(\vec{x})}{g(\mathbf{z}(\vec{x}))}\right), \log\left(\frac{z_j(\vec{x}+\vec{h})}{g(\mathbf{z}(\vec{x}+\vec{h}))}\right)\right], \qquad \underline{\underline{\Xi}}(\vec{h}) = \left(\xi_{i\cdot j}(\vec{h})\right), \tag{6.9}$$

it is a singular matrix for any $\vec{h}$, and it univocally specifies the covariance structure through

$$\sigma_{ij\cdot kl} = \xi_{i\cdot j} + \xi_{k\cdot l} - \xi_{i\cdot l} - \xi_{k\cdot j};$$

alr cross-covariance is the $(D-1)\times(D-1)$ matrix of cross-covariance functions of the alr-transformed data (6.3),

$$\sigma_{ij}(\vec{h}) = \mathrm{Cov}\left[\log\left(\frac{z_i(\vec{x})}{z_D(\vec{x})}\right), \log\left(\frac{z_j(\vec{x}+\vec{h})}{z_D(\vec{x}+\vec{h})}\right)\right] = \sigma_{ij\cdot DD}, \qquad \underline{\underline{\Sigma}}(\vec{h}) = \left(\sigma_{ij}(\vec{h})\right);$$

this is the favored specification, because it has the lowest number of elements, the resulting matrix is non-singular, and it still completely specifies the covariance structure, since

$$\sigma_{ij\cdot kl} = \sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{kj}.$$

These definitions, the properties of all these covariance specifications and its variogram counterparts, as well as the relationships between them were exposed by Pawlowsky-Glahn and Olea [2004]. From the point of view of the Simplex space structure, these covariance structures will be replaced by the coordinate cross-covariance.

Definition 6.1 (coordinate cross-covariance) *The matrix of auto- and cross-covariance functions*

$$C_{ij}(\vec{h}) = \mathrm{Cov}\left[\zeta_i(\vec{x}), \zeta_j(\vec{x}+\vec{h})\right].$$

of the coordinates of the data set with respect to a basis (6.5) form a $(D-1)\times(D-1)$ *matrix, called* coordinate cross-covariance function matrix.

Property 6.1 *The coordinate cross-covariance function matrix (definition 6.1) and the clr cross-covariance matrix (equation 6.9) satisfy*

$$\underline{\underline{C}}(\vec{h}) = \underline{\underline{\varphi}} \cdot \underline{\underline{\Xi}}(\vec{h}) \cdot \underline{\underline{\varphi}}^t,$$

which can be inverted to

$$\underline{\underline{\Xi}}(\vec{h}) = \underline{\underline{\varphi}}^t \cdot \underline{\underline{C}}(\vec{h}) \cdot \underline{\underline{\varphi}},$$

in the case of an orthonormal basis.

Proof: In matrix terms, this covariance specification can be related to the clr covariance structure,

$$\begin{aligned}
\underline{\underline{C}}(\vec{h}) &= \mathrm{Cov}\left[\underline{\zeta}(\vec{x}), \underline{\zeta}(\vec{x}+\vec{h})\right] = \mathrm{E}\left[\left(\underline{\zeta}(\vec{x}) - \underline{\mu}\right)^t \cdot \left(\underline{\zeta}(\vec{x}+\vec{h}) - \underline{\mu}\right)\right] = \\
&= \mathrm{E}\left[\underline{\underline{\varphi}} \cdot \left(\underline{\mathrm{clr}}(\mathbf{z}(\vec{x})) - \underline{\mathrm{clr}}(\mathbf{m})\right)^t \cdot \left(\underline{\mathrm{clr}}(\mathbf{z}(\vec{x}+\vec{h})) - \underline{\mathrm{clr}}(\mathbf{m})\right) \cdot \underline{\underline{\varphi}}\right] = \\
&= \underline{\underline{\varphi}} \cdot \mathrm{E}\left[\left(\underline{\mathrm{clr}}(\mathbf{z}(\vec{x})) - \underline{\mathrm{clr}}(\mathbf{m})\right)^t \cdot \left(\underline{\mathrm{clr}}(\mathbf{z}(\vec{x}+\vec{h})) - \underline{\mathrm{clr}}(\mathbf{m})\right)\right] \cdot \underline{\underline{\varphi}}^t = \\
&= \underline{\underline{\varphi}} \cdot \mathrm{Cov}\left[\underline{\mathrm{clr}}(\mathbf{z}(\vec{x})), \underline{\mathrm{clr}}(\mathbf{z}(\vec{x}+\vec{h}))\right] \cdot \underline{\underline{\varphi}}^t = \\
&= \underline{\underline{\varphi}} \cdot \underline{\underline{\Xi}}(\vec{h}) \cdot \underline{\underline{\varphi}}^t.
\end{aligned}$$

The inverse relationship is proven by considering equation 6.7:

$$\begin{aligned}
\underline{\underline{C}} &= \underline{\underline{\varphi}} \cdot \underline{\underline{\Xi}} \cdot \underline{\underline{\varphi}}^t \\
\underline{\underline{\varphi}}^t \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}} &= \underline{\underline{\varphi}}^t \cdot \underline{\underline{\varphi}} \cdot \underline{\underline{\Xi}} \cdot \underline{\underline{\varphi}}^t \cdot \underline{\underline{\varphi}} \\
\underline{\underline{\varphi}}^t \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}} &= \left(\underline{\underline{I}} - \frac{1}{D}\underline{\underline{1}}\right) \cdot \underline{\underline{\Xi}} \cdot \left(\underline{\underline{I}} - \frac{1}{D}\underline{\underline{1}}\right)^t \\
\underline{\underline{\varphi}}^t \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}} &= \underline{\underline{\Xi}} - \frac{1}{D}\left(\underline{\underline{1}} \cdot \underline{\underline{\Xi}} + \underline{\underline{\Xi}} \cdot \underline{\underline{1}}\right) + \frac{1}{D^2}\underline{\underline{1}} \cdot \underline{\underline{\Xi}} \cdot \underline{\underline{1}} \\
\underline{\underline{\varphi}}^t \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}} &= \underline{\underline{\Xi}},
\end{aligned}$$

given the fact that $\underline{\underline{\Xi}}$ sums up to zero both by rows and columns. □

The advantage of the coordinate cross-covariance structure lays on the fact that, since the coordinates are real-valued RFs, their covariance functions can be interpreted in a classical way. However, the relationship between each original part in the composition and the coordinates is not one-to-one, which hinders a direct interpretation in terms of the parts.

Recall that the choice of a basis does not affect such structural concepts as range, independence, anisotropy and validity of a cross-covariance model, as was shown in section 4.3. Note also that matrix $\underline{\underline{\varphi}}$ allows us to pass from clr transformation scores (6.2) to ilr coordinates (6.5), whereas in section 4.3 it represented an arbitrary change-of-basis matrix of full rank, *e.g.*between two different coordinate systems.

6.4 Kriging in the Simplex

In the case of the D-part Simplex $\mathbb{S}^D$, the so-called isometric log-ratio (ilr) coordinates are taken first by computing either the clr (6.2) or the logarithms of the parts, and afterwards applying the matrix operation described in equation (6.5). This yields

$$\underline{\zeta}_0^* = \underline{\underline{\varphi}} \cdot \underline{\mathrm{clr}}(\mathbf{z}_0^*) = \underline{c} + \sum_{n=1}^{N} \underline{\underline{\lambda}}_n \cdot \underline{\zeta}(\vec{x}_n), \tag{6.10}$$

with $\underline{c} = \underline{\underline{\varphi}} \cdot \underline{\mathrm{clr}}(\mathbf{c})$ ensuring unbiasedness by setting $\underline{c} = \underline{0}$—in universal kriging, thus satisfying the unbiasedness conditions of equation (4.11)—or $\underline{c} = \left(\underline{\underline{I}}_D - \sum \underline{\underline{\lambda}}_n\right) \cdot \underline{\mu}$ (in simple kriging) and $\underline{\mu}$ the vector of known mean values of the ilr coordinates.

The optimal predictor is obtained by minimizing the kriging error variance (eq. 4.4) using covariance functions of definition 6.1, subject to the classical universality conditions (eq. 3.18), if universal kriging is used. Properly expressed in coordinates, the system is reduced to a collocated real kriging, like that explained in chapter 3. The final predictor is obtained applying the result of (6.10) to the basis of the space

$$\mathbf{z}_0^* = \mathcal{C}\left(\exp\left(\underline{\underline{\varphi}} \cdot \underline{\zeta}_0^*\right)\right). \tag{6.11}$$

Consider now the simplest case: the weight matrices as scalars ($\underline{\underline{\lambda}}_n = \lambda_n \cdot \underline{\underline{1}}$). Aitchison [1997] showed—in the non-geostatistical case—that predictor (6.11) is equivalent to a *closed weighted geometric* average of the data set. As a kriging system, it is equivalent to the ALN kriging system of Pawlowsky [1986], although ALN-Kriging yielded an estimator which optimality properties were unclear, while (6.11) has been shown in chapter 4 to be the optimal predictor in the Euclidean structure of $\mathbb{S}^D$. In particular, it minimizes the expected Aitchison distance (2.23) between the true composition $\mathbf{Z}(\vec{x}_0)$ and its prediction $\mathbf{z}_0^*$ (property 4.3). Finally, the predictor (6.10) and its kriging variance are the parameters of the normal distribution on the Simplex, conditional to the observed data set (proposition 4.5).

Regarding block kriging and global change-of-support models, the same considerations expressed for positive variables (section 5.4) apply to kriging in the Simplex. Block kriging (for local estimation) does not yield any new problem as those already treated in section 5.4. The multivariate character of compositions must nevertheless be taken into account when applying any change-of-support model (for global estimation), like *e.g.*the affine correction (3.34). Consider for instance the Cholesky decomposition of the covariance matrix, denoted as $\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}^{1/2} \cdot \underline{\underline{\Sigma}}^{t/2}$. Then, the affine correction can be expressed in the Simplex as

$$\underline{\underline{\Sigma}}^{-1/2} \odot (\mathbf{Z} \ominus \mathbf{m}) \sim \underline{\underline{\Sigma_v}}^{-1/2} \odot (\mathbf{Z}_v \ominus \mathbf{m}) \sim \mathcal{N}_{\mathbb{S}^D}\left(\underline{0}, \underline{\underline{I}}_{D-1}\right),$$

which summarizes both the relationship between the coordinates and the compositions, and the fact that the affine correction is applied to the coordinates due to its Gaussian distribution.

6.5 Case study: air quality index

The data set of moss pollution by Fe, Pb and Hg, at the Ukrainian Carpathian Range is an example of regionalized composition. Its compositional character was introduced in section 2.5.3, where we ignored its spatial dependence. Figure 6.1 shows this data set, with each element in a different map.

To analyze this composition, we will take into account the fact that its sample space is the 3-part Simplex. So, let $\mathbf{Z} = (\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg}) \in \mathbb{S}^3$ be the random composition indicating the proportion of these three elements in each sample. Its sample space, $\mathbb{E} = \mathbb{S}^3$, can be given an Euclidean structure [Billheimer et al., 2001, Pawlowsky-Glahn and Egozcue, 2001], characterized by the operations explained in section 2.5.3. This Euclidean structure captures a scale for compositions as closed relative vectors: they only convey information on the relative importance of a part within a whole, and the total amount of this whole is irrelevant.

Switching to the notation in expression (6.5), the orthonormal coordinates used

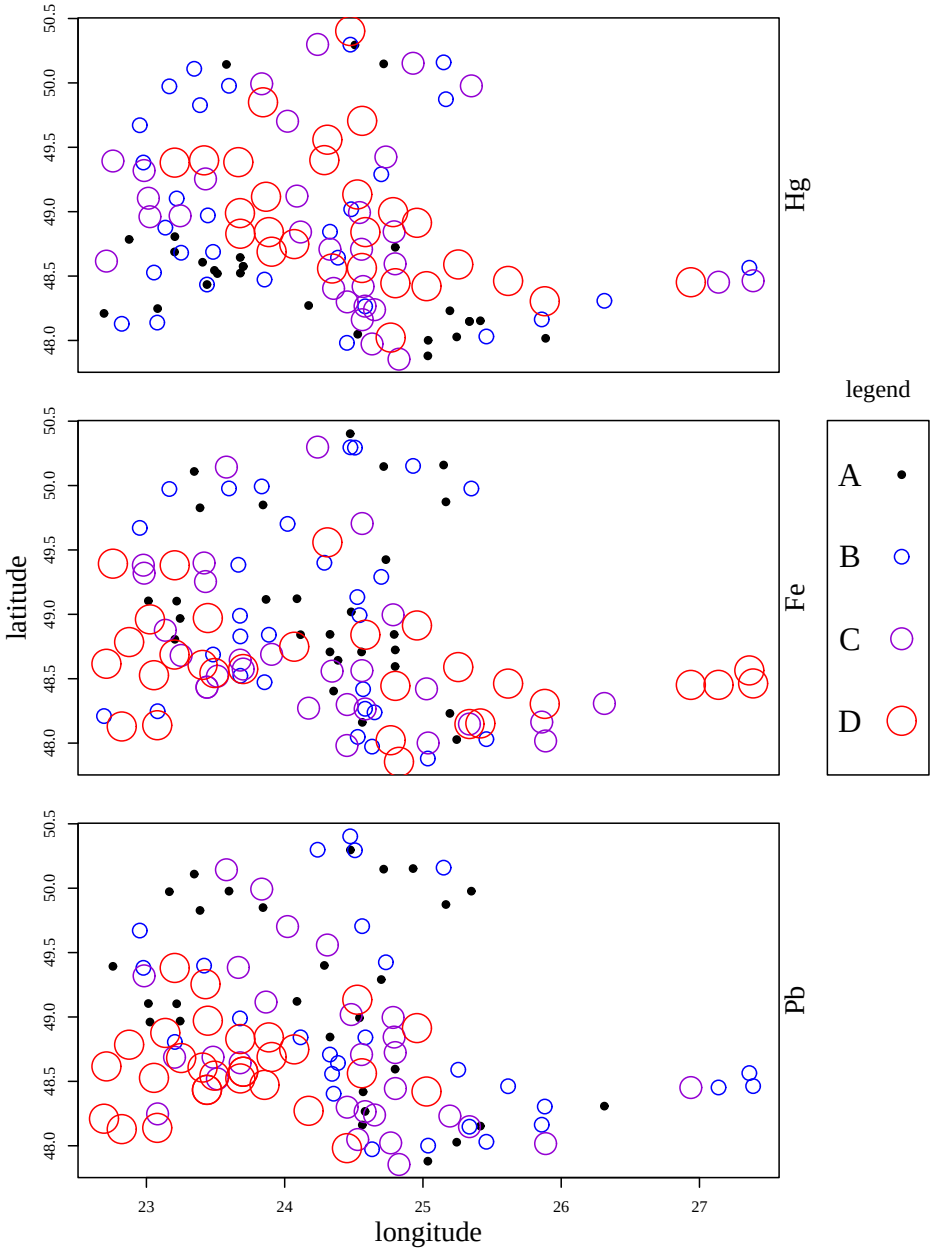

Figure 6.1: Maps of concentration of each element at the sampled locations. The circles represent the experimental quartiles for each element, according to the following table (units in $mg/100g$):

	A	B	C	D
Hg	0.008-0.024	0.024-0.039	0.039-0.054	0.054-0.114
Fe	66-198	198-311	311-496	496-1326
Pb	2.3-4.2	4.2-6.8	6.8-10.9	10.9-32.7

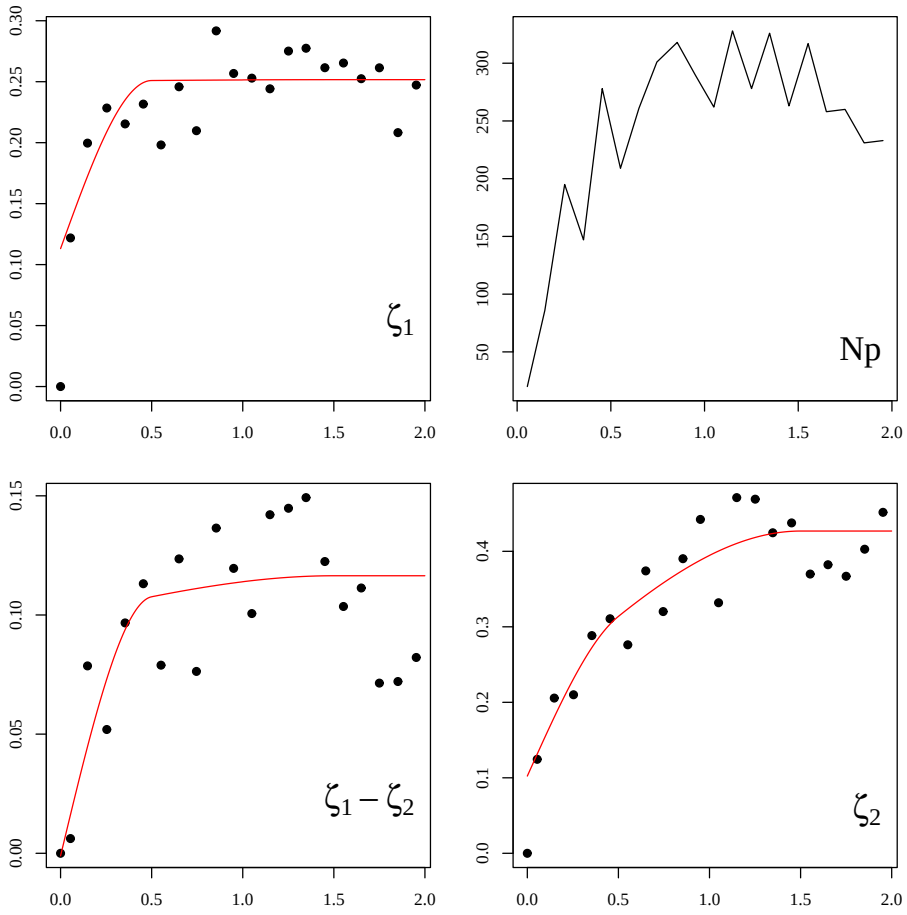

Figure 6.2: Omnidirectional variograms (diagonal plots) and cross-variogram (left bottom plot) for the coordinates of the moss pollution system, jointly with the number of pairs (top right plot) used in the computations.

Table 6.1: Parameters of definition of the auto- and cross-variograms used for (ζ_1, ζ_2). Both contain a nugget, and two nested spherical structures, with a short ($a = 0.5$) and a long ($a = 1.5$) range. Note that they are used into a linear model of corregionalization, which needs these variances and covariances to define a valid covariance matrix for each structure.

variable	nugget	sill (a=0.5)	sill (a=1.5)
ζ_1	0.113	0.137	0.001
ζ_2	0.102	0.107	0.218
ζ_1 vs. ζ_2	-0.001	0.101	0.017

here can be expressed as

$$\underline{\zeta} = \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} = \underline{\underline{\varphi}} \cdot \begin{pmatrix} \ln(\mathrm{Fe}) \\ \ln(\mathrm{Pb}) \\ \ln(\mathrm{Hg}) \end{pmatrix} = \underline{\underline{\varphi}} \cdot \ln(\mathbf{Z}), \quad \underline{\underline{\varphi}} = \begin{pmatrix} \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}.$$

Note that the rows of matrix $\underline{\underline{\varphi}}$ are the clr transforms of each component of the two elements of the basis used, according with equation (6.8)

$$\mathbf{e}_1 = \mathcal{C}\left(\exp\frac{2}{\sqrt{6}}, \exp\frac{-1}{\sqrt{6}}, \exp\frac{-1}{\sqrt{6}}\right) \quad \mathbf{e}_2 = \mathcal{C}\left(1, \exp\frac{1}{\sqrt{2}}, \exp\frac{-1}{\sqrt{2}}\right).$$

These coordinates are chosen because we expect Fe to be more noisy than Pb, which in turn should be less continuous than Hg; to better keep the smoothness of Hg, we balance it against Pb in the second coordinate, and these two against Fe in the first coordinate. We will then interpret the first coordinate as a balance between the influences of short- and long-range pollution, and the second coordinate will inform us on which is the dominant character of the long-range pollution.

Once the basis and the coordinates are fixed, covariance or variograms are computed. In this case, we do not know the mean of the RF, and we expect to use ordinary kriging in the predictions. Variograms are thus preferable, although they would not be able to capture an asymmetric structure. However, we will assume that the covariance function is isotropic, therefore it must be symmetric [Chilès and Delfiner, 1999, p. 324] Using the classical formula for estimation of variograms and cross-variograms (3.5), we estimate the omni-directional variogram and cross-variogram functions displayed in figure 6.2, jointly with the fitted model, which is described in table 6.1. Recall that this model is a linear corregionalization. Thus each nested structure must be linked to a valid variance-covariance matrix.

The ordinary kriging predictor is:

$$\underline{\zeta}^*(\vec{x}_0) = \begin{pmatrix} \zeta_1^*(\vec{x}_0) \\ \zeta_2^*(\vec{x}_0) \end{pmatrix} = \sum_{n=1}^{N} \begin{pmatrix} \lambda_{11}(\vec{x}_n) & \lambda_{12}(\vec{x}_n) \\ \lambda_{21}(\vec{x}_n) & \lambda_{22}(\vec{x}_n) \end{pmatrix} \cdot \begin{pmatrix} \zeta_1(\vec{x}_n) \\ \zeta_2(\vec{x}_n) \end{pmatrix} = \sum_{n=1}^{N} \underline{\underline{\lambda}}(\vec{x}_n) \cdot \underline{\zeta}(\vec{x}_n),$$

subject to the universality conditions

$$\sum_{n=1}^{N} \begin{pmatrix} \lambda_{11}(\vec{x}_n) & \lambda_{12}(\vec{x}_n) \\ \lambda_{21}(\vec{x}_n) & \lambda_{22}(\vec{x}_n) \end{pmatrix} = \underline{\underline{I}}_2,$$

which yields as solution of the system (3.19) the matrix [Myers, 1982]

$$\begin{pmatrix} \underline{\underline{\lambda}}(\vec{x}_1) \\ \vdots \\ \underline{\underline{\lambda}}(\vec{x}_N) \\ \underline{\underline{\nu}} \end{pmatrix} = \begin{pmatrix} \underline{\underline{\Gamma}}(\vec{x}_1, \vec{x}_1) & \cdots & \underline{\underline{\Gamma}}(\vec{x}_1, \vec{x}_N) & \underline{\underline{1}}_2 \\ \vdots & \ddots & \vdots & \vdots \\ \underline{\underline{\Gamma}}(\vec{x}_N, \vec{x}_1) & \cdots & \underline{\underline{\Gamma}}(\vec{x}_N, \vec{x}_N) & \underline{\underline{1}}_2 \\ \underline{\underline{1}}_2 & \cdots & \underline{\underline{1}}_2 & \underline{\underline{0}}_2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \underline{\underline{\Gamma}}(\vec{x}_1, \vec{x}_0) \\ \vdots \\ \underline{\underline{\Gamma}}(\vec{x}_N, \vec{x}_0) \\ \underline{\underline{I}}_2 \end{pmatrix}.$$

Here, the Lagrange multipliers used to satisfy the universality conditions are in the matrix $\underline{\underline{\nu}}$, whereas the 2×2 matrices $\underline{\underline{\Gamma}}$ are formed by the variograms $\gamma_{ij}(\vec{x}_n, \vec{x}_m)$. Results of this kriging system are displayed in figure 6.3. They were obtained with function `predict.gstat` of package `gstat` [Pebesma and Wesseling, 1998] for `R` [R Development Core Team, 2004].

Afterwards, these results are applied to the basis to obtain the final estimates, computed as

$$\mathbf{z}^*(\vec{x}_n) = \zeta_1^*(\vec{x}_n) \odot \mathbf{e}_1 \oplus \zeta_2^*(\vec{x}_n) \odot \mathbf{e}_2.$$

Figure 6.4 represents respectively the predicted values for Hg, Fe and Pb in clr scale. These maps might be interpreted as showing the relative influence of each pollution process. The higher relative influence of corrosion processes (clrFe map) is at the west of the city of Chernovtsí (with a lesser area near Drogóbich), whereas combustion pollution (clrPb map) is relatively most important in the southwestern basin, around the urban center of Uzhgorod. Finally, the highest relative impact of regional pollution (clrHg map) is found in the low polluted area at the northeast of the Carpathians, in the rainy north face of the mountains.

6.6 Remarks

This chapter presented a straightforward generalization of the log-normal and normal in $\mathbb{R}_+$ kriging techniques to deal with compositional data. It essentially served as illustration of the theory in chapter 4 and, as we will see, to present some tools for the next chapter. With the first aim, we derived maps of influence of each of three pollution processes, which were deemed *relative*. This relative character comes from the assumption that the total size of the composition is only an irrelevant artifact of the sampling strategy.

Kriging techniques have been shown to provide optimal predictions of RFs which sample space may be given a meaningful Euclidean structure, in particular the Simplex and the positive real line. An Euclidean space structure allows us to select an orthonormal basis, compute the *real* coordinates of any object in this space with respect to this basis, and apply known methods to them. Thus, we defined variogram or covariance functions on the coordinates, and used kriging techniques applied to the coordinates themselves. Afterwards, predictions were applied to the basis to recover objects of the Euclidean space itself. The same techniques could be applied both to RFs with point-support and with block-support.

Alternatively, we showed that the kriging predictor and its error variances provided the characteristic measures of central tendency and dispersion needed to define a normal distribution on the Simplex. This was regarded as the distribution of the true value of the RF conditional to the observed data, which allowed us to compute probabilities of various hazardous events. This powerful result is however valid under the assumption that the RF is Gaussian, and provided that we know exactly its mean and covariance structure. If the RF was not Gaussian, there were some techniques still applicable,

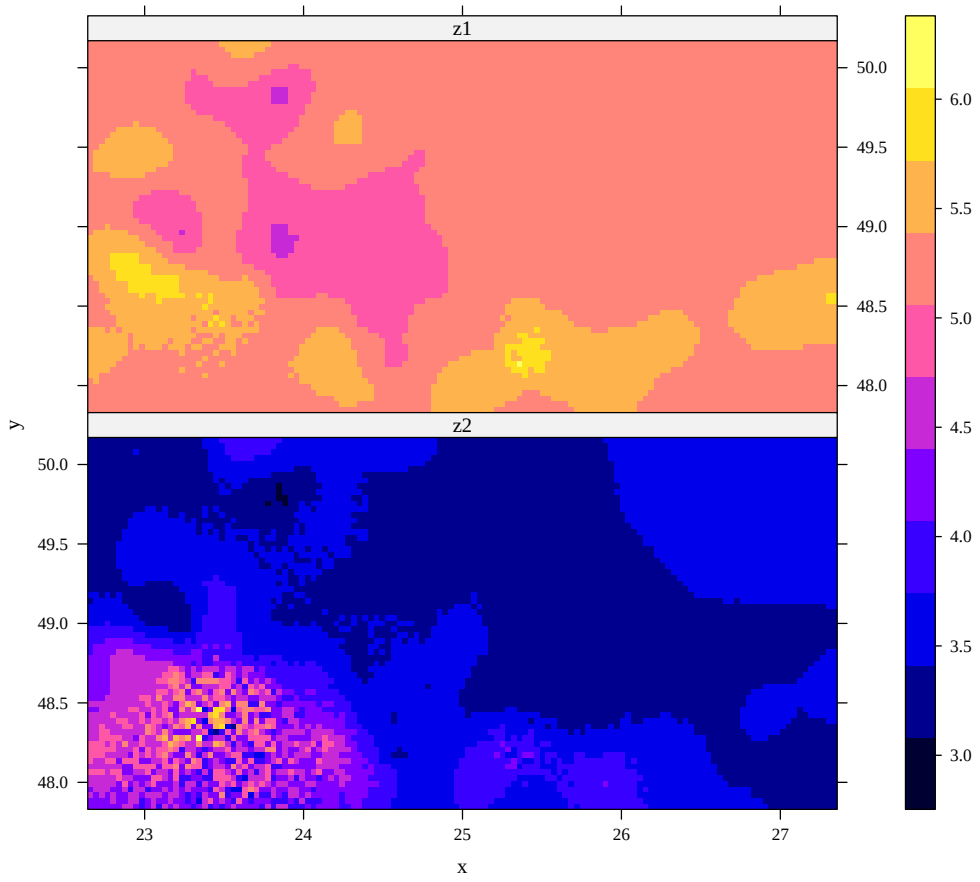

Figure 6.3: Maps of predicted values for each coordinate (here represented as z_1 and z_2) in the moss pollution system. Longitude corresponds to direction x, and latitude to direction y.

which relied upon transforming the observed data to have a Gaussian marginal and *assuming* the joint distribution to be Gaussian.

Given the importance of the information provided by the distribution on the characterization of hazard, some other techniques have been developed in the last years, which focus on directly estimating probabilities. They are addressed in the next chapter.

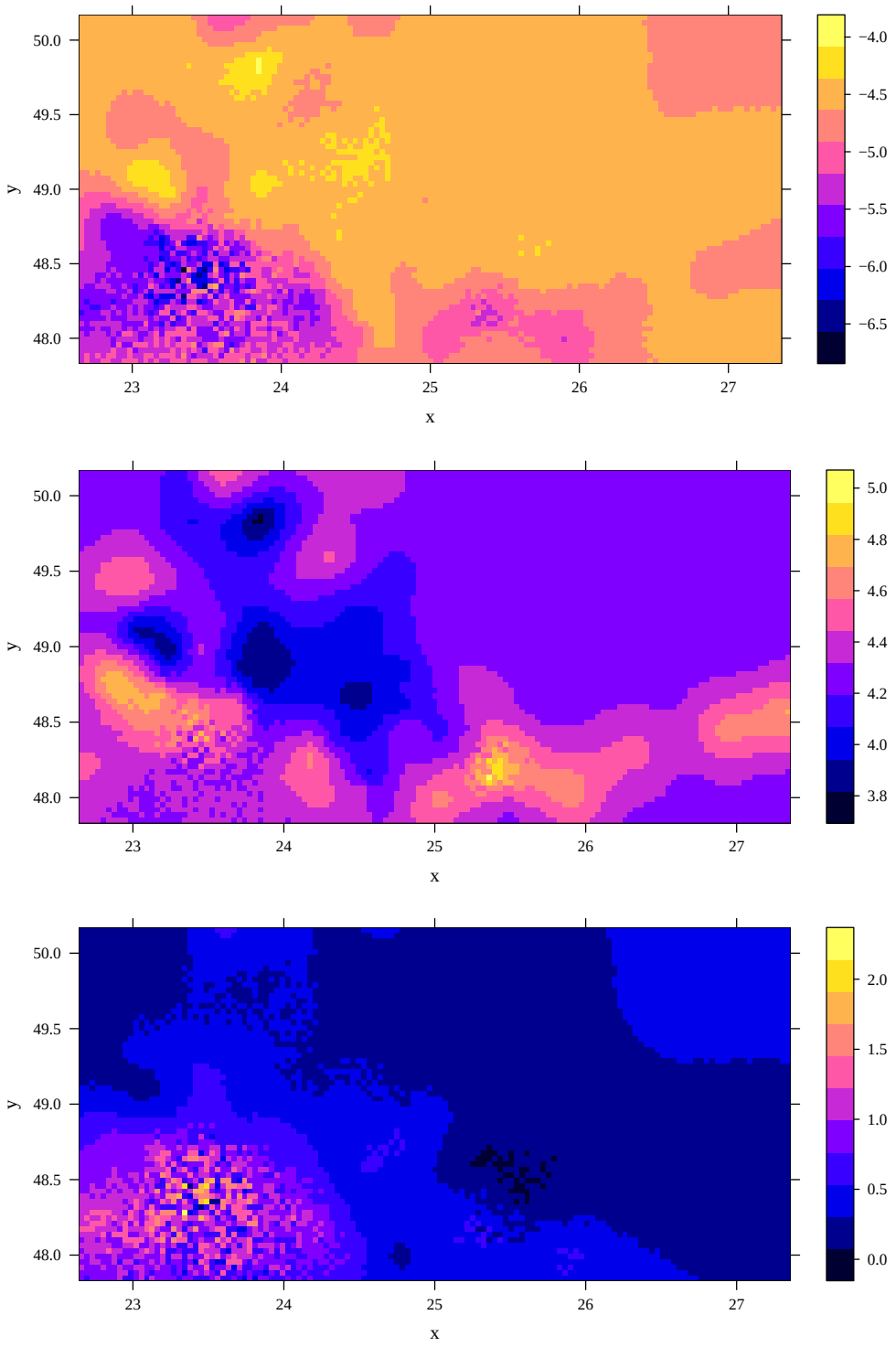

Figure 6.4: Predicted values for each element in clr scale: from top to bottom clrHg, clrFe and clrPb. Longitude corresponds to direction x, and latitude to direction y.

Chapter 7

Geostatistics for probability functions

Geostatistics offers some techniques aimed to estimate the probability distribution of a random function (RF) without assuming its joint Gaussianity—with a possible preliminary transformation—, inherent in the lognormal and ALN-methods (and the alternatives we put forward) of the last chapters. Most of these non-Gaussian techniques are based on the application of indicator functions, which transform the original RF into a boolean one: at any location, the transformed RF is one if the original fulfills a certain condition (*e.g.*"*being above a given threshold*"), and zero otherwise. Then, this boolean RF is treated with classical geostatistics. Kriging results are finally interpreted as the conditional probability of fulfilling the stated condition at an unsampled location. However, this probability may be negative or larger than one. We propose a solution based on considering probabilities as compositions, and applying to them kriging in the Simplex, as explained in the last chapter. However, compositions with zeroes and ones, as obtained from indicator transforms, are placed at the infinity of the Simplex when seen from the log-ratio point of view. Consequently, we cannot use these transforms to estimate the sought conditional probability. To overcome this problem, we first develop kriging in the Simplex for probabilities which are assumed to be directly observed. Afterwards, we put forward a method—which does not yield zeroes or ones—to replace the classical indicator transform. Then, this generalized indicator transform is used in conjunction with kriging in the Simplex, and we study its properties in detail. Finally, we put forward a global method, integrating these two steps into a joint estimation method, which bears more resemblances to bayesian estimation than to geostatistical techniques.

7.1 Indicator kriging and other probability estimation techniques

7.1.1 Random function

Let $\vec{x} \in \mathcal{D} \subset \mathbb{R}^p$ be a point (or the center of a block v) in a domain $\mathcal{D}$ of the space-time real space, with $p \in \{1, 2, 3, 4\}$. Let $Z(\vec{x}) \in \mathbb{A}$ be a RF in a set $\mathbb{A}$ admitting a partition. Let $z(\vec{x}_1), z(\vec{x}_2), \ldots, z(\vec{x}_N)$ be an observed sample of this RF. The goal will be the estimation of the probability distribution of Z at an unsampled location $\vec{x}_0$, either its density (pdf, both for continuous and discrete variables) or its cumulative distribution (cdf). To do so, $\mathbb{A} = \bigcup_{i=1}^{D} A_i$ is partitioned in D disjoint sets $\{A_i\}$, *e.g.*according to one of the following cases:

1. $\mathbb{A}$ is a (totally) ordered set (not necessarily with a specific space structure); then, we can define $\{z_0, z_1, \ldots, z_D\}$, a set of $D+1$ cutoffs or reference levels for Z which in turn define a set of intervals $A_i = (z_{i-1}, z_i]$ partitioning $\mathbb{A}$; using them, the RF Z is transformed to the set of indicator functions

$$I_i(\vec{x}) = I_i(Z(\vec{x})) = \begin{cases} 1, & Z(\vec{x}) \leq z_i \\ 0, & Z(\vec{x}) > z_i, \end{cases} \quad i \in \{1, 2, \ldots, D-1\}, \tag{7.1}$$

   which define a vector $\mathbf{I}(\vec{x})$ of $D-1$ *ordered* boolean RFs, since $i < j \rightarrow I_i(\vec{x}) \leq I_j(\vec{x})$ necessarily for any $\vec{x} \in \mathcal{D}$;

2. $Z(\vec{x})$ is a categorical RF, and $\mathbb{A}$ is the finite set of its outcomes; we can define then a set of D indicator functions

$$J_i(\vec{x}) = J_i(Z(\vec{x})) = \begin{cases} 1, & Z(\vec{x}) = A_i \\ 0, & Z(\vec{x}) \neq A_i, \end{cases} \quad i \in \{1, 2, \ldots, D\}, \tag{7.2}$$

   which define a vector $\mathbf{J}(\vec{x})$ of D *disjunctive* boolean RFs, where necessarily at each sampled location $\vec{x} \in \mathcal{D}$ *one and only one* of the J_i is equal to one and the rest are zero; if these categories can be ordered, *e.g.*A_i is the event "Z' *belongs to the interval* $(z_{i-1}, z_i]$" (with Z' an auxiliary RF in an ordered set fulfilling the conditions of the first point), then the trivial relation holds

$$I_i(\vec{x}) = \sum_{j<i} J_j(\vec{x}); \tag{7.3}$$

3. $\mathbb{A}$ is a subset of a multidimensional vector space admitting a partition, *e.g.*the set of wind directions (the unit circle) or the metamorphic phase pressure-temperature space (the positive quadrant $\mathbb{R}^2_+$); in this case, the RF Z is transformed through (7.2), since usually no cumulative relation is naturally defined on multidimensional spaces.

These three cases contain most of the possible situations, although other possibilities could be devised: the key condition is the possibility to univocally partition the image set $\mathbb{A}$, and define an associated categorical RF satisfying the second case. Then the following techniques can be applied to this associated RF, and the computed probabilities may be finally transferred to the original. This procedure will obviously be meaningful only when that partition has a physical sense.

7.1.2 Structural analysis

For a given index i, the real expectations of both functions I_i (7.1) and J_i (7.2) are related respectively to the cdf $F_Z(z)$ (specially in the continuous case) and the pdf $f_Z(z)$ (specially in the discrete case):

$$\mathrm{E}\left[I_i(\vec{x})\right] = \Pr\left[Z(\vec{x}) < z_i\right] = F_Z(z_i) \quad \text{and} \quad \mathrm{E}\left[J_i(\vec{x})\right] = \Pr\left[Z(\vec{x}) = A_i\right] = f_Z(A_i). \tag{7.4}$$

Journel [1983] put forward the possibility to apply geostatistical methods to the ordered indicator transforms of the observed data set, denoted by $\mathbf{i}(\vec{x}_n)$, to estimate $F_Z(z) = \mathrm{E}\left[\mathbf{I}(\vec{x})\right]$. To do so, he defined cross-covariance $\{C_{ij}\}$ and variogram $\{\gamma_{ij}\}$ functions for the indicator transforms, and showed that they satisfied the following relationships with the bivariate cumulative probability distribution:

$$\begin{aligned} K_{ij}(\vec{h}) &= \mathrm{E}\left[I_i(\vec{x}) \cdot I_j(\vec{x}+\vec{h})\right] = \Pr\left[\{Z(\vec{x}) \le z_i\} \cap \{Z(\vec{x}+\vec{h}) \le z_j\}\right] \\ &= F_{ZZ}(z_i, z_j), \end{aligned} \tag{7.5}$$

$$\begin{aligned} C_{ij}(\vec{h}) &= \mathrm{Cov}\left[I_i(\vec{x}), I_j(\vec{x}+\vec{h})\right] = F_{ZZ}(z_i, z_j) - F_Z(z_i) \cdot F_Z(z_j), \\ C_{ii}(\vec{0}) &= \mathrm{Cov}\left[I_i(\vec{x}), I_i(\vec{x})\right] = \mathrm{Var}\left[I_i(\vec{x})\right] = F_Z(z_i) - F_Z^2(z_i) = \\ &= \mathrm{E}\left[I_i(\vec{x})\right]\left(1 - \mathrm{E}\left[I_i(\vec{x})\right]\right) = \gamma_{ii}(\infty), \\ \gamma_{ij}(\vec{h}) &= F_Z(z_i)\delta_{ij} - F_Z(z_i) \cdot F_Z(z_j). \end{aligned} \tag{7.6}$$

Equation (7.5) shows the non-centered cross-covariance of the random variables $I_i(\vec{x})$ and $I_j(\vec{x}+\vec{h})$, which was used by Carle and Fogg [1996] to define the *transition probability*,

$$t_{ij}(\vec{h}) = \mathrm{E}\left[I_j(\vec{x}+\vec{h}) | I_i(\vec{x})\right] = \frac{\Pr\left[\{Z(\vec{x}) \le z_i\} \cap \{Z(\vec{x}+\vec{h}) \le z_j\}\right]}{\Pr\left[Z(\vec{x}) \le z_i\right]} = \frac{F_{ZZ}(z_i, z_j)}{F_Z(z_i)}$$

Equation (7.6) shows that knowledge of the sill of the variogram or the covariance at the origin of a single indicator is equivalent to the knowledge of its mean value. Similar relationships can be established between the bivariate discrete probability density function of the events A_i and covariance functions between disjunctive indicators J_i. For instance, a non-centered cross-covariance can be also defined as

$$\begin{aligned} K_{ij}(\vec{h}) &= \mathrm{E}\left[J_i(\vec{x}) \cdot J_j(\vec{x}+\vec{h})\right] = \Pr\left[\{Z(\vec{x}) = A_i\} \cap \{Z(\vec{x}+\vec{h}) = A_j\}\right] \\ &= f_{ZZ}(A_i, A_j), \end{aligned} \tag{7.7}$$

and a transition probability as

$$t_{ij}(\vec{h}) = \mathrm{E}\left[J_j(\vec{x}+\vec{h})|J_i(\vec{x})\right] = \frac{\Pr\left[\{Z(\vec{x})=A_i\}\cap\{Z(\vec{x}+\vec{h})=A_j\}\right]}{\Pr\left[Z(\vec{x})\leq z_i\right]} = \frac{f_{ZZ}(A_i,A_j)}{f_Z(A_i)} \tag{7.8}$$

7.1.3 Linear prediction

Journel [1983] proposed also to estimate $F_Z(z_i) = \mathrm{E}\left[\mathbf{I}(\vec{x})\right]$ by kriging of the transformed data $\mathbf{i}(\vec{x}_n)$. Given the fact that knowledge of the variogram or the covariance system implies knowledge of the mean of $\mathbf{I}(\vec{x})$, simple kriging predictor

$$\mathbf{i}^*_{SK} = \sum_{n=1}^{N} \underline{\underline{\lambda}}_n \cdot \mathbf{i}(\vec{x}_n) - \left(\underline{\underline{I}} - \sum_{n=1}^{N} \underline{\underline{\lambda}}_n\right) \cdot \mathrm{E}\left[\mathbf{I}\right] \tag{7.9}$$

is the most suitable linear technique to use. This is a clear case of collocated kriging, since at each location the whole vector $\mathbf{i}(\vec{x}_n)$ is defined. Consequently, the error variance to minimize is defined as

$$\sigma^2_{SK} = \mathrm{Tr}\left(\mathrm{Var}\left[\mathbf{i}^*_{SK} - \mathrm{E}\left[\mathbf{I}(\vec{x}_0)\right]\right]\right)$$

and the kriging system to solve is equivalent to (3.24), taking $C(\vec{x}_n,\vec{x}_m) = (C_{ij}(\vec{x}_n,\vec{x}_m))$ to be full matrices of cross-covariance,

$$\sum_{m=1}^{N} \underline{\underline{\lambda}}_m \cdot \underline{\underline{C}}(\vec{x}_n,\vec{x}_m) = \underline{\underline{C}}(\vec{x}_n,\vec{x}_0), \quad n=1,2,\ldots,N, \tag{7.10}$$

As a simpler alternative, indicator kriging can be also implemented, by using the kriging predictor (3.23) independently for each level i, which is equivalent to let $\underline{\underline{C}}(\vec{x}_n,\vec{x}_m)$ be a diagonal matrix; resulting weights $\underline{\underline{\lambda}}$ are also diagonal matrices in this simplified case. Again, an equivalent approach to (7.9) can be implemented on the $\mathbf{J}(\vec{x})$ to compute estimates of $f_Z(A_i) = \mathrm{E}\left[\mathbf{J}(\vec{x})\right]$. All these techniques are usually called Indicator Kriging (IK).

Simultaneous independent indicator kriging (the variant with diagonal weight and covariance matrices) has as its best advantage its simplicity over the full indicator co-kriging (*i.e.*the simultaneous kriging of all indicator levels, accounting for their cross-covariance) and many other techniques presented later, whereas full indicator co-kriging is expected to yield better predictions, in the sense of lower error variance. However, it is reported that the arbitrary fitting of the cross-covariance functions needed for full co-kriging counters this theoretical advantage [Goovaerts, 1994]. Furthermore, the sufficient conditions which a set of cross-covariances must satisfy to be an admissible model for probabilities are unknown: only some necessary conditions were exposed by Journel and Posa [1990]. In particular, Bogaert [2002] points that the linear model

of coregionalization is only valid with a proportional covariance, which turns out to simplify the co-kriging system to a set of independent kriging systems of each variable, thus spoiling the advantage of co-kriging. But the worse flaw of indicator kriging techniques is the fact that usually they yield results which are impossible to interpret as probabilities. In the predictor (7.9) or in the system of equations (7.10), nothing guarantees that the result remains positive, or lower than 1, conditions which should be satisfied by any individual probability. Furthermore, even in the kriging case, the estimates of $\mathbf{I}(\vec{x}_0)$ are not necessarily ordered which, due to expression (7.3), implies that the estimates of $\mathbf{J}(\vec{x}_0)$ may be negative or larger than one. Finally, if the system is built on $\mathbf{J}(\vec{x}_0)$, its component estimates do not necessarily sum up to one. These are usually addressed as the *order relation problems* of indicator kriging, the major source of concern about the adequateness of this technique.

7.1.4 Indicator kriging family techniques

To solve these problems, a set of techniques have been developed which essentially take into account complementary information in the indicator kriging system, with the hope of reducing uncertainty and consequently order relation problems. In this context, *probability kriging* [Sullivan, 1984] complements the indicator transform with a rank transform, which takes into account the order of the samples. The resulting technique is median unbiased but still presents order problems [Carr and Mao, 1993], which gives ground to Carr [1994] and Bogaert [1999] to suggest two different smoothing approaches of the estimated quantiles through trans-Gaussian-type curves which automatically correct most order-relations, although they have a rather weak theoretical basis. In a very similar fashion, the technique called *cumulative distribution function of order statistics kriging* [Juang et al., 1998] takes into account the deviation between the observed value and the cutoff, like the rank transform, and offer similar predictions with similar flaws. Another way of reducing the order relation problems is based on the better characterization of the covariance structure between several cutoffs. In this way, *indicator principal component kriging* [Suro-Perez and Journel, 1991] looks for a principal component decomposition of the covariance structure in order to skip the modeling of cross-covariances, and *successive kriging of indicators* [Vargas-Guzman and Dimitrakopoulos, 2003] does the same by considering successively the data samples by groups, and applying kriging to estimate the residual probabilities with respect to the already-used data. Transition probabilities [Carle and Fogg, 1996] were also shown to reduce the number of order relation occurring in a classical indicator kriging. Pardo-Igúzquiza and Dowd [2005] proposed an easy correction, based on fitting a logistic regression model using generalized least squares (*e.g.*see section 3.3.2) to the kriged quantiles, and replacing those quantiles which violate order-relation by their prediction using the obtained logistic regression.

The most-used straightforward patch applied to indicator kriging to reduce the number of order relation problems is using the same covariance for all the estimated cutoff levels. This solution, already proposed by Journel [1983] in his seminal paper

on indicator kriging, is equivalent to assume a *mosaic model* for the spatial structure [Chilès and Delfiner, 1999, p. 384]. In this simple model, $Z(\vec{x}_n)$ and $Z(\vec{x}_m)$ are assumed either equal with probability $\rho(\vec{x}_n - \vec{x}_m)$ and the distribution of their common value is $F(z)$, or independent and each has the same $F(z)$ distribution. In other words, the space is partitioned in cells where the RF is constant. This picture does not always fit the many applications of indicator kriging.

7.1.5 Disjunctive kriging

Another classical alternative to indicator kriging is *disjunctive kriging* [Matheron, 1976]. Assume that $Z(\vec{x})$ is a continuous variable in a totally ordered set (usually a subset of the real line), satisfying the first case in page 134. As exposed by Chilès and Delfiner [1999], disjunctive kriging amounts to the joint kriging of all indicator functions associated with all possible cutoff levels z_i. To model the infinite number of covariances associated with it, disjunctive kriging must assume one of a particular group of models for the joint bivariate distribution: those accepting an isofactorial decomposition. This means that there exists a series of functions $\{\psi_n\}$ indexed by a natural number $n \in \mathbb{N}$, which form an orthogonal system with respect to the bivariate probability distribution $F(z_1, z_2)$,

$$\int \psi_n(z)\psi_m(z+\vec{h})dF(z(\vec{x}), z(\vec{x}+\vec{h})) = \delta_{nm}T_n(\vec{h}),$$

where δ_{nm} is the Kronecker delta function, and the $T_n(\vec{h})$ functions are the coefficients of the bivariate density function in that orthogonal system. Note that the series of functions $\{\psi_n\}$ are an orthogonal basis of $L^2(\mathbb{R})$, the Hilbert space of square-integrable real functions of a real variable [Berberian, 1961].

Then the joint bivariate distribution of two random variables can be expressed as a product of their marginals scaled by a given simple mixing function. In a Gaussian case with correlation coefficient $\rho(\vec{h})$, it holds that $T_n(\vec{h}) = \rho^n(\vec{h})$, and consequently

$$F(z_1, z_2) = \sum_{n=0}^{\infty} \rho^n(\vec{h})\psi_n(z_1)\psi_n(z_2)\phi(z_1)\phi(z_2),$$

being $\psi_n(z)$ the Hermite polinomial of n-th degree, and $\phi(z)$ the univariate normal density law. There are other isofactorial decompositions for the most-frequently used probability models [Armstrong and Matheron, 1986a,b]. The K-order approximation to the true probability distribution offered by disjunctive kriging is

$$F^*(Z_0) = \left(1 + \sum_{k=1}^{K}\sum_{n=1}^{N} \psi_k(Z_0)\lambda_{kn}\psi_k(y_n)\right)\Phi(Z_0),$$

where the weights λ_{kn} for each one of the k orders are obtained from an independent simple kriging system built with a covariance function $C(h) = T_n(h)$. The result of

disjunctive kriging $F^*(Z_0)$ is regarded as the best approximation to $F(Z_0)$ of order K in the mean square sense. This points to one of the flaws of disjunctive kriging, since nothing guarantees that $F^*(Z) > 0$, and indeed for moderate degrees of approximation K, the tails of the estimated distribution present negative values. In summary, although much more theoretically grounded (and rather complicated) than indicator kriging, disjunctive kriging may offer also impossible estimates of the probability distribution.

These problems of indicator and disjunctive kriging might be related to the fact that both deliver the best approximation in the sense that the $L^2(\mathbb{R})$ distance between the estimated function and the true distribution is minimal. This distance is defined as

$$d^2(f, f^*) = \int_{\mathbb{R}} (f(z) - f^*(z))^2 dz,$$

which is a flawed distance between densities, given that it does not take into account their specific properties, *i.e.*they must be positive and integrate to one.

7.1.6 Bayesian/maximum entropy method

The Bayesian-Maximum Entropy formalism [Christakos, 1990] offers also an alternative to kriging of indicators, particularly suited for the case of disjunctive indicators. Following the reasoning exposed in section 3.4.3, Bogaert [2002] argues that the best way to encode the relationship between two categorical variables placed at two locations separated a lag distance $\vec{h}$ is the two-way contingency table, defined in the context of disjunctive indicator variables as the non-centered covariance at lag $\vec{h}$ (7.7). Then, he estimates the maximum-entropy $(N+1)$-way contingency table by fitting a non-saturated log-linear model

$$\begin{aligned} \log f_Z(Z_0 = z_0, Z_1 = z_1, \quad \ldots, \quad Z_N = z_N) = \\ \eta_0 \quad + \quad \sum_{i=1}^{N} \left(\eta_1(Z_i = z_i) + \sum_{j \neq i}^{N} \eta_2(Z_i = z_i, Z_j = z_j) \right) \end{aligned}$$

involving only a constant term (ensuring that the final probability estimates sum up to one) and first and second-order interactions. This fitting is done using standard numerical algorithms, given the nonexistence of analytical forms for these distributions. The final estimate of the distribution is the vector of values $f_Z(Z_0, Z_1 = z_1, \ldots, Z_N = z_N)$, with Z_0 varying among all its possible outcomes.

The main advantage of BME in general, and of this application in particular, is the extreme flexibility of the method. As the author points out, *e.g.*if we had three-way contingency tables as data, we should simply add a third-order interaction, and so on with higher orders. Also, it is important to notice that it does not estimate a probability by minimizing a squared error, as did indicator kriging and disjunctive kriging; instead it uses a logarithmic criterion, which implies that the obtained results will be always

valid positive probabilities. On the side of the disadvantages, BME methods have a difficult implementation, and most usually involve extensive computation. In the special case of BME for categorical variables, one has to add a theoretical inconsistency: according to van den Boogaart [2005, pers. comm.], given N_o data, the contingency tables of order $N_o + 1$ and $N_o + N_p$ (say, with one single predicted location and with N_p simultaneously-predicted locations) are not compatible.

7.2 Kriging in the Simplex for multinomial probability vectors

7.2.1 Random function

From now on, we will focus on the disjunctive indicator approach (eq. 7.2), the second case in page 134. Let $\mathbf{P}(\vec{x})$ be a vector RF satisfying at all locations $\vec{x} \in \mathcal{D}$

- the components of the vector RF are positive: $P_i(\vec{x}) \geq 0$ for all $i = 1, 2, \ldots, D$,
- the sum of all components is one: $\sum_{i=1}^{D} P_i(\vec{x}) = 1$.

Then let $Z(\vec{x}) \sim \mathcal{M}(\mathbf{P}(\vec{x}); 1)$ be the realization of a single-trial multinomial distribution at location $\vec{x}$, where each component of the RF indicates the probability of obtaining a given category: $P_i(\vec{x}) = \Pr[Z(\vec{x}) = A_i]$. Thus we are assuming here a two-step stochastic process, where first a RF of probability vectors is realized, and afterwards at each location an *independent realization* of a multinomial distribution is drawn. From this point of view, the vector of disjunctive indicators $\mathbf{J}(\vec{x})$ (7.2) is nothing else than a different encoding of the result of this multinomial realization.

The two-step random process suggested here was already used by Diggle et al. [1998] to model the relative incidence of campylobacter infections among other enteric infections. They assumed $\mathbf{P}(\vec{x})$ to be a logistic-normal RF, which is equivalent to transforming the function through a logistic transformation and assuming a normal distribution for the transformed scores. However, they had more than one single replication of the associated binomial variable.

7.2.2 The D-part Simplex, a space structure for multinomial probability vectors

Let us first head to the case where we observe directly the probability vector RF: $\mathbf{p}(\vec{x}_1), \mathbf{p}(\vec{x}_2), \ldots, \mathbf{p}(\vec{x}_N)$ is an observed sample. From the conditions stated above on the positive and closed character of these vectors, it is evident that their sample space is the D-part Simplex ($\mathbb{S}^D$). The following considerations suggest that the Euclidean space structure explained in section 2.5.3 is applicable to probability vectors, because they describe a meaningful scale of this vectors.

The closure operation $\mathcal{C}(\cdot)$ converts a likelihood vector into a probability vector. Perturbation is equivalent to the updating of two probability vectors $\mathbf{p}, \mathbf{q}$ (or likelihoods) to obtain another one $\mathbf{r}$, as shown by Aitchison [1982],

$$\mathbf{r} = (r_1, r_2, \ldots, r_D) = \frac{(p_1 q_1, p_2 q_2, \ldots, p_D q_D)}{\sum_{i=1}^{D} p_i q_i} = \mathcal{C}\left(p_1 q_1, p_2 q_2, \ldots, p_D q_D\right) = \mathbf{p} \oplus \mathbf{q}. \quad (7.11)$$

Power operation is interpreted then as the self-updating of a probability vector or likelihood.

Log-odds, with a long history in the fitting of probability models (*e.g.*logistic regression), are obtained as the result of the alr transformation (6.3). Following this, the coordinate vector $\underline{\pi}$ with respect to a basis $\mathbf{E} = \{\mathbf{e}_i\}$ has $D-1$ components, computed with scalar products like

$$\pi_i = \langle \mathbf{p}, \mathbf{e}_i \rangle_A = \underline{\varphi}_{i\cdot} \cdot \underline{\mathrm{clr}}\left(\mathbf{p}\right) = \underline{\varphi}_{i\cdot} \cdot \underline{\log}\left(\mathbf{p}\right) = \log \prod_{j=1}^{D} p_j^{\varphi_{ij}}, \quad (7.12)$$

where $\underline{\varphi}_{i\cdot}$ represents the vector of the $i-th$ row of the matrix $\underline{\underline{\varphi}}$ characterizing the basis of the Simplex used (6.5). This vector components sum up to zero (6.6), which means that the last part of this equality is the logarithm of a fraction, and we may consider it a *generalized log-odd*. Aitchison introduced them in the standard analysis of compositional data under the name of *log-contrasts*. Recall that the vectors of the basis themselves can be retrieved from the matrix by equation (6.8).

Also, Egozcue [2002] suggested that the Aitchison norm of a probability vector may be interpreted as a measure of the *information* carried by this vector, thus offering a connection between the metric structure of the Simplex and entropy and information formalism [Shannon, 1948]. Furthermore, Egozcue and Díaz-Barrero [2003] extended this Euclidean space structure for discrete multinomial probabilities to a Hilbert space structure for bounded continuous densities [Egozcue et al., 2006]. Finally, van den Boogaart [2004] established an equality between the Aitchison distance between distributions and the mean information for discrimination between them [Kullback, 1997].

7.2.3 Linear prediction

Given the results of the preceding sections, we can predict the $\mathbf{P}(\vec{x})$ RF at an unsampled location $\vec{x}_0$ conditionally on the observed sample $\mathbf{p}(\vec{x}_n)$, by using any coordinate system and the kriging technique for multi-dimensional Euclidean spaces of chapter 4. The optimization criterion will be the minimization of the error variance, defined as

$$\sigma_k^2 = \mathrm{Tr}\left[\mathrm{Var}\left[\mathbf{p}_K^* \ominus \mathrm{E}\left[\mathbf{P}(\vec{x}_0)\right]\right]\right] = \mathrm{Tr}[\underline{\underline{\Sigma}}_{K,\mathbb{S}^D}] = \mathrm{E}\left[d_{\mathbb{S}}^2(\mathbf{p}_K^*, \mathbf{P}(\vec{x}))\right],$$

where $\mathbf{p}_K^*$ is the kriging estimator, and $\mathbf{P}(\vec{x}_0)$ the true value of the RF $\mathbf{P}(\vec{x})$ at the unsampled location $\vec{x}_0$. Recall that $\underline{\underline{\Sigma}}_{K,\mathbb{S}^D} = \mathrm{Var}\left[\mathbf{p}_K^* \ominus \mathrm{E}\left[\mathbf{P}(\vec{x}_0)\right]\right]$ is the variance-covariance matrix between the errors of the coordinates, according to property 4.2.4, showing us

the *stochastic dependence* of the uncertainty about the different components of predictor vector.

The coordinate vector of the kriging predictor itself $\underline{\pi}^*_K$ is a weighted linear combination (4.2) of the coordinates of the observed data, where the weights are derived from a system of equations like (3.19). Thus they are $D-1$ *algebraically independent* quantities, in contrast with what is obtained in classical indicator kriging, where they must either be ordered or sum up to one. However, the results of kriging in the Simplex are not probabilities before they are applied to the basis used through

$$\mathbf{p}^*_K = \bigoplus_{i=1}^{D-1} \pi^*_i(\vec{x}_0) \odot \mathbf{e}_i. \tag{7.13}$$

This prediction $\mathbf{p}^*_K$ is the best unbiased linear approximation to the true $\mathbf{P}(\vec{x}_0)$ in the Euclidean structure of the Simplex, as it is a kriging predictor on the coordinates. Furthermore, assuming a joint normal distribution on the Simplex for the RF, we can be sure that

$$\mathbf{P}(\vec{x}_0) \sim \mathcal{N}_{\mathbb{S}^D}(\mathbf{P}^*_k, \Sigma_{k,\mathbb{S}^D}), \quad \text{since} \quad \underline{\Pi}(\vec{x}_0) \sim \mathcal{N}_{\mathbb{R}^{D-1}}(\underline{\pi}^*_K, \underline{\underline{\Sigma}}_{K,\mathbb{S}^D}),$$

being $\underline{\Pi}(\vec{x})$ the vector of coordinates of the RF $\mathbf{P}(\vec{x})$. Thus, it yields the conditional distribution of the unknown $\mathbf{P}(\vec{x}_0)$. In other words, all the optimality properties attached to the different kriging techniques for point and block-support will be preserved under the very special assumption of joint Gaussianity of the RF of $\mathbf{P}(\vec{x})$.

7.3 Kriging in the Simplex for generalized indicators

7.3.1 The generalized indicator function

General case

In section 7.2.1, we stated that this chapter would consider $Z(\vec{x}) \sim \mathcal{M}(\mathbf{P}(\vec{x}); 1)$ independently drawn from multinomial distributions which probability vectors formed a RF, in a two-step stochastic process. Afterwards, we derived a kriging technique based on log-odds of this probability RF $\mathbf{P}(\vec{x})$, and assuming that this was partially observed. This is an unrealistic assumption, since what we have is the value of $Z(\vec{x})$ at the sampled locations $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$, or alternatively its disjunctive encoding through equation (7.2): $\mathbf{j}(\vec{x}_1), \mathbf{j}(\vec{x}_2), \ldots, \mathbf{j}(\vec{x}_N)$. Thus, before any kriging method shall be applied, we need to obtain at each sampled location $\vec{x}_n$ the value of $\mathbf{p}(\vec{x}_n)$ from the *single* observation $\mathbf{j}(\vec{x}_n)$. Classical indicator kriging does so by simply taking

$$\mathbf{p}^*(\vec{x}_n) = \mathbf{j}(\vec{x}_n), \quad n = 1, 2, \ldots, N$$

and Journel [1983] interprets it as the conditional probability

$$\mathbf{p}^*(\vec{x}_n) = \Pr\left[Z(\vec{x}_n) = A_i | z(\vec{x}_m),\ m = 1, 2, \ldots, N\right],$$

which afterwards is to be introduced in the kriging procedure. For our purposes, we cannot do so, because log-odds are not defined when the probability vector has zeros. Also, this estimation is inconsistent with the two-step stochastic nature of $Z(\vec{x})$, since it states that another realization of this process would yield the same result at a sampled location.

Thus, we look for an alternative procedure. Under the two-step stochastic model, after observing the occurrence of category A_k at a sampled location, we may *subjectively* attach to each category A_i an own probability $p^*_{i|k}$. Organizing these probabilities in an array, we obtain what is called a *design matrix*: after observing event A_k, column k of this matrix tells us the conditional probability of occurrence of all the events at that location, both the observed and the non-observed. With this matrix we could, for instance, determine that the observation of category A_k would give to this category a 70% of the probability, the two nearest categories (if there is an order, A_{k-1} and A_{k+1}) would equally share a 20%, and the rest of the categories would equally share the exceeding 10%. Note that such an approach would give the analyst a big deal of freedom in modeling the relationship between the categories, but demands the determination of many parameters. Instead, we introduce here a simpler approach, the generalized indicator function,

$$p^*_{i|k} = \begin{cases} a, & i = k \\ \frac{1-a}{D-1}, & i \neq k \end{cases} . \tag{7.14}$$

This depends only on one parameter $a \in (0, 1)$, which logically should be $a > 1/D$ in order to obtain a higher probability for the observed category. Indicator kriging in its classical form uses the extreme case $a = 1$. Note that all the categories which were not observed receive exactly the same estimated probability, regardless of whether they are "near" or "far" from the observed one. A Bayesian justification of this generalized indicator is discussed in this chapter *addendum* (page 175), and it is also used to ground section 7.4.

Results of expression (7.14) form a compositional vector $\mathbf{p}^*(\vec{x})$, which falls inside the D-part Simplex. Thus, the prediction approach developed in section 7.2 for unsampled locations of the RF $\mathbf{P}(\vec{x})$ can be also applied to this case. The next properties detail how to compute the coordinates of the generalized indicator function, and its expectation, both in coordinates and as a probability.

Property 7.1 (Coordinates of the generalized indicator function) *Assuming the k-th category to be observed, the coordinate vector of the generalized indicator function (7.14) is proportional to the k-th column of the matrix of* ilr *coordinates $\underline{\underline{\varphi}}$,*

$$\underline{\pi}^* = \alpha \cdot \underline{\varphi}_k = \alpha \cdot \underline{\underline{\varphi}} \cdot \boldsymbol{j}_k. \tag{7.15}$$

Proof:

$$\begin{aligned}\pi_i^* &= \log\prod_{j=1}^{D}\left(p_j^*\right)^{\varphi_{ij}} = \log\left(a^{\varphi_{ik}}\left(\frac{1-a}{D-1}\right)^{\sum_{j\neq k}\varphi_{ij}}\right) = \\ &= \log\left(a^{\varphi_{ik}}\left(\frac{1-a}{D-1}\right)^{-\varphi_{ik}}\right) = \varphi_{ik}\log\frac{a}{\frac{1-a}{D-1}} = \varphi_{ik}\cdot\alpha,\end{aligned}$$

The first equality comes from (7.12), the second is the application of the predictor (7.14) once we consider that we observed the k-th category ($j_k = 1$), the third equality holds due to the fact that all rows of matrix $\underline{\underline{\varphi}}$ sum up to zero (6.6), and the rest of the equalities are simple algebraic manipulations. □

Definition 7.1 (Canonical generalized indicator transformation) *We consider* canonical *the generalized indicator transformation where $\alpha = 1$.*

Property 7.2 *The expectation of the generalized indicator function expressed in coordinates, a measure of central tendency, is*

$$\mathrm{E}\left[\pi_i^*\right] = \sum_{k=1}^{D}\varphi_{ik}\cdot\alpha\cdot p_k, \tag{7.16}$$

where p_k is the true *probability of observing the k-th category.*

Property 7.3 *The expectation in the Simplex of the generalized indicator function itself, a the central tendency element, is*

$$\mathrm{E}_A\left[\boldsymbol{p}^*\right] = \mathcal{C}\left(\left(\frac{a(D-1)}{1-a}\right)^{p_1},\left(\frac{a(D-1)}{1-a}\right)^{p_2},\ldots,\left(\frac{a(D-1)}{1-a}\right)^{p_D}\right). \tag{7.17}$$

Proof:

$$\begin{aligned}\mathrm{E}_A\left[\mathbf{p}^*\right] &= \bigoplus\nolimits_{i=1}^{D-1}\left(\sum_{k=1}^{D}\varphi_{ik}\alpha p_k\right)\odot\mathbf{e}_i = \bigoplus\nolimits_{i=1}^{D-1}\left(\sum_{k=1}^{D}\varphi_{ik}\alpha p_k\right)\odot\mathcal{C}\left(\exp\left(\varphi_{ij}\right)\right) = \\ &= \mathcal{C}\left(\exp\left(\sum_{k=1}^{D}\sum_{i=1}^{D-1}\varphi_{ij}\varphi_{ik}\alpha p_k\right)\right) = \mathcal{C}\left(\exp\left(\sum_{k=1}^{D}\left(\delta_{kj}-\frac{1}{D}\right)\alpha p_k\right)\right) = \\ &= \mathcal{C}\left(\exp\left(\alpha p_j - \frac{1}{D}\alpha\sum_{k=1}^{D}p_k\right)\right) = \mathcal{C}\left(\exp\left(\alpha p_j\right)\cdot\exp\left(-\frac{1}{D}\alpha\right)\right) = \\ &= \mathcal{C}\exp\left(p_j\log\frac{a(D-1)}{1-a}\right)\oplus\mathcal{C}\left(\exp\left(-\frac{\alpha}{D}\right)\right) = \mathcal{C}\left(\left(\frac{a(D-1)}{1-a}\right)^{p_j}\right),\end{aligned}$$

The first equality comes from (7.16), the second one is implied by the definition of each element of the basis (6.8), the third is an application of standard manipulation of coordinates in an Euclidean space, the fourth is true due to the orthogonality properties of matrix $\underline{\underline{\varphi}}$ (6.7), the fifth and sixth come from the properties of Dirac delta function and standard algebraic manipulation, the seventh is the definition of perturbation-updating (7.11), and the last holds due to the fact that $\left(\exp\left(-\frac{\alpha}{D}\right), \exp\left(-\frac{\alpha}{D}\right), \ldots, \exp\left(-\frac{\alpha}{D}\right)\right)$ is a constant vector, which is the neutral element of perturbation. □

Two-categories case

Note that in the simplest two-categories case, the matrix of definition of the orthonormal basis is $A = \left(1/\sqrt{2}, -1/\sqrt{2}\right)$, which means that the unique basis element $\mathbf{e} = \mathcal{C}\left(\exp\left(1/\sqrt{2}, -1/\sqrt{2}\right)\right)$. This yeilds as the unique coordinate

$$\pi = \langle \mathbf{p}, \mathbf{e} \rangle_A = \langle (p_1, p_2), \left(e^{1/\sqrt{2}}, e^{-1/\sqrt{2}}\right) \rangle_A \frac{1}{\sqrt{2}} \log \frac{p_1}{p_2} = \frac{1}{\sqrt{2}} \log \frac{p}{1-p}. \tag{7.18}$$

In this simplest case, with two mutually-exclusive categories or a single cutoff, the generalized indicator function is

$$p_2^* = \begin{cases} a, & j_2(\vec{x}) = 1 \\ 1-a, & j_2(\vec{x}) = 0 \end{cases}, \tag{7.19}$$

and $p_1^* = 1 - p_2^*$. This yields, by using coordinate (7.18), the expectation of the generalized indicator coordinate equal to

$$\mathrm{E}\left[\pi^*\right] = \alpha \frac{1}{\sqrt{2}}(p_1 - p_2), \tag{7.20}$$

where p_1, p_2 are the true probabilities of obtaining respectively categories A_1 and A_2 in the binomial experiment. The expectation as a probability vector is

$$\mathrm{E}_A\left[\mathbf{p}^*\right] = \mathcal{C}\left(\left(\frac{a}{1-a}\right)^{p_1}, \left(\frac{a}{1-a}\right)^{p_2}\right). \tag{7.21}$$

7.3.2 Structural analysis

General case

Let us now head to the structural characterization of the RF $\mathbf{P}(\vec{x}), \vec{x} \in \mathcal{D}$ from the available information on the observations of $\mathbf{j}(\vec{x}_n)$, *i.e.*from the estimates $\mathbf{p}^*(\vec{x}_n)$ and their coordinates $\underline{\pi}^*(\vec{x}_n)$ obtained with equation (7.15).

Property 7.4 (Covariance of the generalized indicator function) *The covariance function of the generalized indicator function, represented by $\underline{\underline{C}}(\vec{h})$, can be obtained from the covariance of the disjunctive indicator function, represented with $\underline{\underline{C}}^J(\vec{h})$, through*

$$\underline{\underline{C}}(\vec{h}) = \alpha^2 \underline{\underline{\varphi}} \cdot \underline{\underline{C}}^J(\vec{h}) \cdot \underline{\underline{\varphi}}^t.$$

This relationship can be inverted to

$$\underline{\underline{C}}^J = \alpha^{-2} \underline{\underline{\varphi}}^t \cdot \underline{\underline{C}} \cdot \underline{\underline{\varphi}}. \tag{7.22}$$

Proof: We are interested in the set of coordinate cross- and auto-covariances like

$$C_{ij}(\vec{h}) = \mathrm{E}\left[\pi_i(\vec{x}) \cdot \pi_j(\vec{x}+\vec{h})\right] - \mathrm{E}\left[\pi_i(\vec{x})\right] \cdot \mathrm{E}\left[\pi_j(\vec{x}+\vec{h})\right].$$

The negative terms of this expression are provided by equation (7.16), whereas the positive part is, as an expectation, equal to

$$\begin{aligned} \mathrm{E}\left[\pi_i(\vec{x}) \cdot \pi_j(\vec{x}+\vec{h})\right] &= \sum_{k_1=1}^{D}\sum_{k_2=1}^{D} \alpha\varphi_{ik_1} \Pr\left[\{j_{k_1}(\vec{x})=1\} \cap \{j_{k_2}(\vec{x}+\vec{h})=1\}\right] \alpha\varphi_{jk_2} = \\ &= \sum_{k_1=1}^{D}\sum_{k_2=1}^{D} \alpha^2 \varphi_{ik_1}\varphi_{jk_2} K_{k_1k_2}(\vec{h}), \end{aligned} \tag{7.23}$$

with $K_{k_1k_2}(\vec{h})$ the non-centered cross-covariance for lag $\vec{h}$ (7.7) between categories k_1 and k_2. Consequently, the covariance function is

$$\begin{aligned} C_{ij}(\vec{h}) &= \sum_{k_1=1}^{D}\sum_{k_2=1}^{D} \alpha^2 \varphi_{ik_1}\varphi_{jk_2} K_{k_1k_2}(\vec{h}) - \sum_{k_1=1}^{D} \alpha\varphi_{ik_1} p_{k_1}(\vec{x}) \sum_{k_2=1}^{D} \alpha\varphi_{jk_2} p_{k_2}(\vec{x}+\vec{h}) = \\ &= \sum_{k_1=1}^{D}\sum_{k_1=1}^{D} \alpha^2 \varphi_{ik_1}\varphi_{jk_2} (K_{k_1k_2}(\vec{h}) - p_{k_1}(\vec{x}) p_{k_2}(\vec{x}+\vec{h})) = \\ &= \sum_{k_1=1}^{D}\sum_{k_1=1}^{D} \alpha^2 \varphi_{ik_1}\varphi_{jk_2} C^J_{k_1k_2}(\vec{h}), \end{aligned} \tag{7.24}$$

where $C^J_{k_1k_2}(\vec{h})$ represents the covariance function at lag $\vec{h}$ computed with disjunctive indicators for the categories k_1 and k_2.

The inverse relatioship (7.24) is derived from property 6.1, given the fact that the matrix $\underline{\underline{C}}^J$ is the raw covariance of a composition, which sums up to zero both by rows and columns, like the matrix of clr-transformed compositions,

$$\begin{aligned} \sum_{i=1}^{D} C^J_{ij}(\vec{h}) &= \sum_{i=1}^{D} (f_{ZZ}(z_i, z_j) - f_Z(z_i) f_Z(z_j)) = \\ &= \sum_{i=1}^{D} f_{ZZ}(z_i, z_j) - f_Z(z_j) \sum_{i=1}^{D} f_Z(z_i) = \\ &= f_Z(z_j) - f_Z(z_j) \sum_{i=1}^{D} f_Z(z_i) = f_Z(z_j) \left(1 - \sum_{i=1}^{D} f_Z(z_i)\right) = 0. \end{aligned}$$

□

Properties 7.2 and 7.4 show that the value of α only scales means and covariance functions, thus the structural analysis does not depend on which generalized indicator function of type (7.14) we are effectively using. In other words, in this step we can simply choose the canonical one (definition 7.1) and do all computations for covariances with it, as well as fit models. The influence of α will be further discussed in the next section.

Another interesting issue shown in this section is the relationship between classical indicator structural functions (7.5-7.8) and both non-centered and centered covariance functions on coordinates (7.23-7.24). Given a set of non-centered cross-covariances for disjunctive indicators (7.8) and a basis of $\mathbb{S}^D$, all the structural functions of the coordinate approach can be computed with simple linear combinations of disjunctive indicator covariances.

The use of the coordinate covariance functions nevertheless clarifies the structural analysis. Covariance functions of disjunctive indicators must sum up to zero at each lag, while covariance of the coordinates are free of this limitation: relation (7.22) ensures that any coordinate indicator covariance will be linked to a disjunctive indicator covariance which sums up to zero both by rows and columns. Both covariances tend to zero beyond their range, and the linearity of (7.22) ensures that if this happens with one of them (either the coordinate or the disjunctive indicator), it happens with the other. The knowledge of the mean of the RF (7.17) allows the direct computation of the covariance at the origin, which must be

$$C_{ij}(\vec{0}) = \sum_{k=1}^{D} \varphi_{ik} p_k \varphi_{jk} - \left(\sum_{k=1}^{D} \varphi_{ik} p_k \right) \cdot \left(\sum_{k=1}^{D} \varphi_{jk} p_k \right)$$

in order to keep consistency with the covariance of disjunctive indicators. In this expression p_k represents the marginal probability of category A_k. Furthermore, ensuring that the $(D-1)^2$ cross-covariances form a valid covariance system ensures the validity of the D^2 disjunctive covariances, due to the linearity of the Fourier Transform. Last but not least, a covariance system which does not satisfy relationship (7.24) may still be a valid covariance model for the original compositional RF (that of section 7.2.3), although it is directly related neither to a generalized indicator nor to a classical disjunctive indicator RF.

Two-categories case

In the simplest case, with two mutually exclusive categories or a single cutoff, the coordinate covariance can be computed using the relation between covariances (7.24)

and the coordinate expression (7.18), which yields

$$\begin{aligned} C(\vec{h}) &= \alpha^2 \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} p_{11}-p_1^2 & p_{12}-p_1p_2 \\ p_{21}-p_1p_2 & p_{22}-p_2^2 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} = \\ &= \frac{\alpha^2}{2}\left(p_{11}-p_1^2+p_{22}-p_2^2-p_{12}-p_{21}+2p_1p_2\right) = \\ &= \frac{\alpha^2}{2}\left(p_{11}+p_{22}-(p_{12}+p_{21})-(p_1-p_2)^2\right), = \\ &= \frac{\alpha^2}{2}\left(1-2p_{12}-2p_{12}-(1-2p_2)^2\right) = \frac{\alpha^2}{2}\left(1-4p_{12}+4p_2-1-4p_2^2\right) = \\ &= 2\alpha^2\left(-p_{12}+p_2-p_2^2\right) = 2\alpha^2\left(p_{22}-p_2^2\right), \end{aligned} \tag{7.25}$$

with $p_i = f_Z(A_i)$ and $p_{ij} = f_{ZZ}(A_i, A_j)$ as in (7.7). Note the validity of the last step because $p_2 = p_{12} + p_{22}$. Considering the intermediate step of expression (7.25), the covariance of the coordinate is the difference between the mean deviations of independence observed between the diagonal and the terms outside it in the contingency table. This expression is still easily adapted to the non-stationary case, whereas in the final simplifications we took into account that with only two categories, assuming stationarity implies symmetry of the contingency table. Note that this final covariance expression for the coordinate of the generalized indicator function is proportional to the covariance of the $J_2(\vec{x})$ disjunctive indicator function.

7.3.3 Linear prediction

General case

Finally, we can head to the prediction of the distribution of $Z(\vec{x}_0)$ at an unsampled location, or the value of $\mathbf{P}(\vec{x}_0)$ conditional to the observed generalized indicators $\mathbf{p}^*(\vec{x}_n), n = 1, 2, \ldots, N$ (7.14), by using exactly the same kriging procedure explained in section 7.2.3. Consequently this will not be repeated here. Some considerations need nevertheless to be taken into account, in order to understand the implications of such a procedure:

- which is the influence of the value of α in the kriging results?
- which is the final predictor for the sought probability distribution, namely $\mathbf{p}^*(\vec{x}_0)$?
- how can we choose a value of α, and most important, which is its interpretation?

As seen in equations (7.16) and (7.24), the parameter α derived from the generalized indicator function (7.14) only scales the problem, and does change neither the shape of the covariances nor the proportions among their sills or the means. It is then straightforward to show that the kriging weigths and results satisfy some proportionality relationships with this factor α.

Property 7.5 *Given a generalized indicator transformation (7.14), with a suite of coordinates proportional to α, and the* canonical *generalized transformation (definition 7.1), it holds:*

1. *the coordinates (7.15) of both suites ($\underline{\pi}^*_{(\alpha)}$ and $\underline{\pi}^*$ respectively) satisfy*

$$\underline{\pi}^*_{(\alpha)}(\vec{x}_n) = \alpha \cdot \underline{\pi}^*(\vec{x}_n),$$

   for all sampled locations $\vec{x}_n$,

2. *the expectations (7.17) of both suites are proportional, with proportionality constant α*

$$\mathrm{E}\left[\underline{\pi}^*_{(\alpha)}(\vec{x})\right] = \alpha \cdot \mathrm{E}\left[\underline{\pi}^*(\vec{x}_0)\right]$$

3. *the covariance functions (7.24) of both suites are proportional, with proportionality constant α^2*

$$\mathrm{Cov}\left[\underline{\pi}^*_{(\alpha)}(\vec{x}_0), \underline{\pi}^*_{(\alpha)}(\vec{x}_0 + \vec{h})\right] = \alpha^2 \cdot \mathrm{Cov}\left[\underline{\pi}^*(\vec{x}_0), \underline{\pi}^*(\vec{x}_0 + \vec{h})\right],$$

4. *the matrix of kriging weights attached to each datum $\vec{x}_n$ by both system suites are identical*

$$\underline{\underline{\lambda}}_{n,(\alpha)} = \underline{\underline{\lambda}}_n$$

5. *the Lagrange multipliers of both system suites are proportional, with proportionality constant α^2*

$$\nu_{(\alpha)} = \alpha^2 \cdot \nu,$$

   provided that the system to solve is a universal kriging system,

6. *the kriging predictors (4.2) of both suites are proportional, with proportionality constant α*

$$\underline{\pi}^*_{(\alpha)}(\vec{x}_0) = \alpha \cdot \underline{\pi}(\vec{x}_0),$$

7. *the kriging error variance (4.9) of both suites are proportional, with proportionality constant α^2*

$$\sigma^2_{K(\alpha)} = \alpha^2 \cdot \sigma^2_K.$$

Sketch of a proof: Items 1, 2 and 3 come directly from their definitions, according with equations (7.15), (7.17) and property 7.22 respectively.

Items 4 and 5 are a classical result in kriging with proportional covariance systems. They can be easily proven using the same matrix notation of property 4.7: Λ and $\Lambda_{(\alpha)}$ are the $D \times (D \cdot N)$ matrices of all kriging weights obtained with a covariance system proportional to 1 and to α respectively; thanks to item 3, C and $\alpha^2 \cdot C$ are the $(D \times N) \times (D \cdot N)$ matrices of covariances used in the two kriging system, and c

and $\alpha^2 \cdot c$ the corresponding right-side terms, matrices of $(D \times N) \times D$ elements. It is straightforward to show that

$$\Lambda_{(\alpha)} = \left(\alpha^2 \cdot C\right)^{-1} \cdot \left(\alpha^2 \cdot c\right) = \alpha^{-2}\alpha^2 \cdot C^{-1} \cdot c = 1 \cdot \Lambda.$$

Eventually, adding row and column block matrices of basis functions (also scaled by α^2 to keep consistency) to matrices C and c, like in solving the UK system (page 91), yields again a matrix of weights Λ which do not depend on α. However, Λ includes now a row block matrix with the Lagrange multipliers ν to be multiplied by the basis functions. The α^2 factor which artificially scaled the basis functions is transferred to the Lagrange multipliers, in order to keep the basis functions themselves unchanged.

The coordinate predictor is affected by α exactly as the data set or the mean, as seen comparing items 1, 2 and 6. This is simply proven when considering that this predictor is of the form

$$\underline{\pi}^*(\vec{x}_0) = \left(\underline{\underline{I}}_D - \sum_{n=1}^{N} \underline{\underline{\lambda}}_n\right) \cdot \mathrm{E}\left[\underline{\pi}^*(\vec{x})\right] + \sum_{n=1}^{N} \underline{\underline{\lambda}}_n \cdot \underline{\pi}^*(\vec{x}_n), \tag{7.26}$$

applying items 1, 2 and 4. Regarding the kriging error variance, item 7 is also directly obtained by applying to property 4.2.4 items 3 and 4. □

Proposition 7.1 (Probability vector predictor) *The final predictor of the probability vector is*

$$\boldsymbol{p}^*(\vec{x}_0) = \mathcal{C}\left(\left(\frac{a(D-1)}{1-a}\right)^{\xi_1}, \left(\frac{a(D-1)}{1-a}\right)^{\xi_2}, \ldots, \left(\frac{a(D-1)}{1-a}\right)^{\xi_D}\right), \tag{7.27}$$

with

$$\xi_k = \sum_{i=1}^{D-1} \pi_i^* \varphi_{ik},$$

where π_i^* is the predicted value for i-th coordinate using the canonical generalized indicator transform. A vectorial expression for ξ_k is also

$$\underline{\xi} = \underline{\underline{\varphi}}^t \cdot \left(\underline{\underline{\lambda}}_{(0)} \cdot \underline{\underline{\varphi}} \cdot \boldsymbol{p} + \sum_{k=1}^{D} \underline{\underline{\lambda}}_{(k)} \cdot \underline{\underline{\varphi}} \cdot \boldsymbol{j}_k\right), \tag{7.28}$$

with $\underline{\underline{\lambda}}_{(k)}$ the sum of all kriging weights attached to the locations where we observed the k-th category, $\boldsymbol{j}_k$ the classical *disjunctive indicator function result of observing this k-th category, $\boldsymbol{p}$ the mean probability (extracted* e.g.*from the sills of the variograms in a structural analysis), and $\underline{\underline{\lambda}}_0 = \underline{\underline{I}} - \sum_{k=1}^{D} \underline{\underline{\lambda}}_{(k)}$.*

Proof: Replacing the definition of the basis (6.8) in the final predictor (7.13), one obtains

$$\mathbf{p}^*(\vec{x}_0) = \bigoplus_{i=1}^{D} \alpha \cdot \pi_i^*(\vec{x}_0) \odot \mathcal{C}\left(\exp\varphi_{i1}, \exp\varphi_{i2}, \ldots, \exp\varphi_{iD}\right),$$

being $\pi_i^*(\vec{x}_0)$ the predictor obtained with the canonical indicator transformation. Then the k-th element is (up to the closure) equal to

$$p_k^*(\vec{x}_0) = \prod_{i=1}^{D} (\exp \varphi_{ik})^{\alpha \cdot \pi_i^*(\vec{x}_0)} = (\exp \alpha)^{\sum_{i=1}^{D} \varphi_{ik} \cdot \pi_i^*(\vec{x}_0)},$$

which yields exactly the desired expressions by realizing that $\exp \alpha = \frac{a(D-1)}{1-a}$.

To proof the vectorial expression (7.28), we shall recall that at every location $\vec{x}_n$ where the k-th category was observed, the data used is exactly the same, offered by equation (7.15). Then, representing by $\underline{\underline{\lambda}}_{(k)}$ the sum of all kriging weights attached to these locations, and given the linearity of the kriging predictor (7.26), we immediately therefrom derive

$$\underline{\underline{\varphi}} \cdot \underline{\xi} = \underline{\underline{\lambda}}_{(0)} \cdot \underline{\underline{\varphi}} \cdot \mathbf{p} + \sum_{k=1}^{D} \underline{\underline{\lambda}}_{(k)} \cdot \underline{\underline{\varphi}} \cdot \mathbf{j}_k,$$

by simply taking into account equation (7.16), regarding the coordinates of the mean of the RF, and equation (7.15) about the coordinates of the generalized indicator function at those places where the k-th category was observed. The "inversion" of the $\underline{\underline{\varphi}}$ function is achieved using its quasi-orthonormality properties of equation (6.7), which yields

$$\underline{\xi}' = \underline{\underline{\varphi}}^t \cdot \left(\underline{\underline{\lambda}}_{(0)} \cdot \underline{\underline{\varphi}} \cdot \mathbf{p} + \sum_{k=1}^{D} \underline{\underline{\lambda}}_{(k)} \cdot \underline{\underline{\varphi}} \cdot \mathbf{j}_k \right) = \underline{\underline{\varphi}}^t \cdot \underline{\underline{\varphi}} \cdot \underline{\xi} = \left(\underline{\underline{I}} - \frac{1}{D} \underline{\underline{1}} \right) \cdot \underline{\xi} = \underline{\xi} - \frac{1}{D} \underline{\underline{1}} \cdot \underline{\xi} = \underline{\xi} - \underline{C}$$

where $\underline{C} = \frac{1}{D} \underline{\underline{1}} \cdot \underline{\xi}$ is a constant vector, and we can write $\underline{\xi} = \underline{\xi}' + \underline{C}$. Then, taking exponentials and closing the result we would obtain

$$\mathcal{C}\left(\exp\left(\alpha \cdot \xi_j\right)\right) = \mathcal{C}\left(\exp\left(\alpha \cdot \xi_j'\right)\right) \oplus \mathcal{C}\left(\exp\left(\alpha \cdot C\right), \exp\left(\alpha \cdot C\right), \ldots, \exp\left(\alpha \cdot C\right)\right),$$

which shows us that $\underline{C}$ evolves into a perturbation by the neutral element of the Simplex due to the effect of the closure. Consequently, $\underline{\xi} \equiv \underline{\xi}'$, proving equation (7.28). □

Note that expression (7.27) results always in a positive set of D values which sum up to one (thanks to the closure operator), thus automatically satisfying the conditions of a multinomial vector. This implies that kriging in the Simplex applied to generalized indicator transformations does never present order problems as indicator kriging did.

Two-categories case

In the case of two categories, or a single cut-off, the kriging predictor of the coordinate (7.18) by using covariance function (7.25) and simple kriging predictor,

$$\pi^*(\vec{x}_0) = \sum_{n=1}^{N} \lambda_n \pi^*(\vec{x}_n) + \left(1 - \sum_{n=1}^{N} \lambda_n\right) \mathrm{E}\left[\pi^*\right], \tag{7.29}$$

is equivalent to classical kriging of the disjunctive indicator $j_1(\vec{x})$ (7.2). By replacing expressions (7.19-7.21) on (7.29), we get

$$\alpha\varphi_{11}(2j_1^*(\vec{x}_0)-1)=\sum_{n=1}^{N}\lambda_n\alpha\varphi_{11}(2j_1^*(\vec{x}_n)-1)+\left(1-\sum_{n=1}^{N}\lambda_n\right)\mathrm{E}\left[\alpha\varphi_{11}(2j_1^*(\vec{x})-1)\right],$$

which is reduced to simple kriging of $j_1(\vec{x})$.

Furthermore, (7.29) is reduced to

$$\pi^*(\vec{x}_0)=\frac{\alpha}{\sqrt{2}}\left(\lambda_1-\lambda_2+(1-\lambda_1-\lambda_2)(p_1-p_2)\right)=\frac{\alpha}{\sqrt{2}}\left(p_1(1-2\lambda_2)-p_2(1-2\lambda_1)\right)$$

where $\lambda_1=\lambda_{(1)}$ represents the sum of the kriging weights for all those locations where category A_1 was observed, and so $\pi^*(\vec{x}_n)=\frac{\alpha}{\sqrt{2}}$. Equivalently $\lambda_2=\lambda_{(2)}$ is the sum of all those location where A_2 was observed, and $\pi^*(\vec{x}_n)=-\frac{\alpha}{\sqrt{2}}$. Applying this to the basis used, with some manipulations parallel to those in expression (7.21), we obtain as final predictor of the probability vector $\mathbf{P}(\vec{x}_0)$

$$\begin{aligned}\mathbf{p}^*(\vec{x}_0) &= \mathcal{C}\left(\left(\frac{a}{1-a}\right)^{p_1(1-2\lambda_2)},\left(\frac{a}{1-a}\right)^{p_2(1-2\lambda_1)}\right)\\ &= \mathcal{C}\left(\left(\frac{a}{1-a}\right)^{p_1},\left(\frac{a}{1-a}\right)^{p_2}\right)\oplus\mathcal{C}\left(\left(\frac{a}{1-a}\right)^{-2p_1\lambda_2},\left(\frac{a}{1-a}\right)^{-2p_1\lambda_2}\right)=\\ &= \mathrm{E}_A\left[\mathbf{p}^*\right]\oplus\mathcal{C}\left(\left(\frac{a}{1-a}\right)^{2\lambda_1p_2},\left(\frac{a}{1-a}\right)^{2\lambda_2p_1}\right),\end{aligned}$$

which in terms of the operations described in section 7.2.2 correspond to the bayesian updating of the mean probability vector by the evidence in favor of one or another category.

We will consider now three simple possibilities.

1. When we try *prediction at an actually-sampled location*, it is known that simple kriging prediction gives a single weight $\lambda=1$ to that location and zero to all others, since it is an exact interpolator. Assume for instance that we observed A_1, thus $\lambda_1=1$ and $\lambda_2=0$. Then the final prediction will be

   $$\tilde{\mathbf{p}}(\vec{x}_0)=\mathcal{C}\left(\left(\frac{a}{1-a}\right)^{p_1},\left(\frac{a}{1-a}\right)^{-p_2}\right)=\mathcal{C}\left(\frac{a}{1-a},1\right)=\mathcal{C}\left(a,1-a\right),$$

   which is the result of (7.19), and consequently this kriging method is also an exact interpolator.

2. If we try to *predict a location out of the range of the covariance*, which implies that all weights will be equal to zero, $\lambda_1=\lambda_2=0$, then the prediction reduces to the mean.

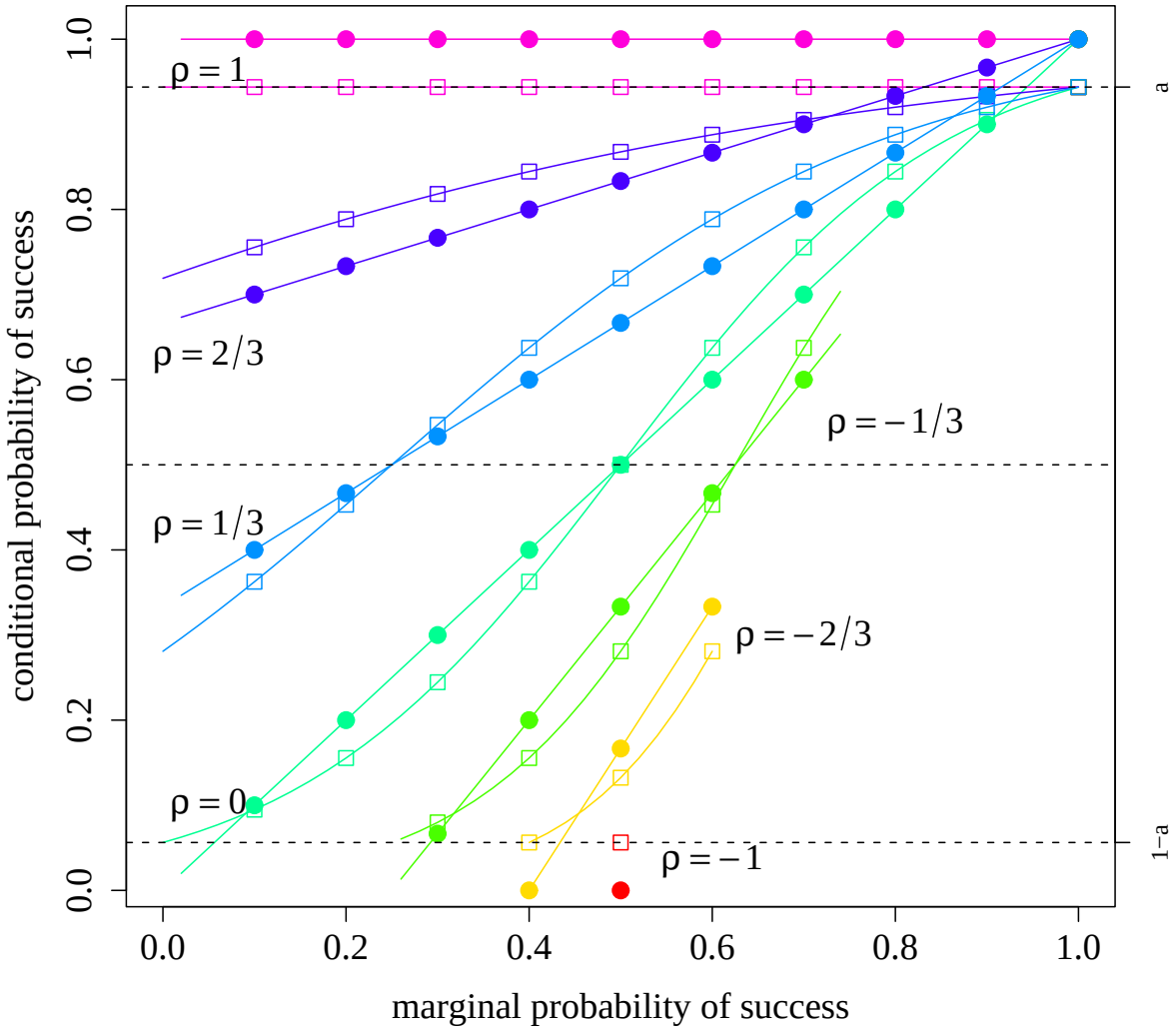

Figure 7.1: Predicted (empty squares) and true (filled circles) conditional probability as a function of the true probability of success in the datum location, for several correlation coefficients, for a value of $a = 0.95$.

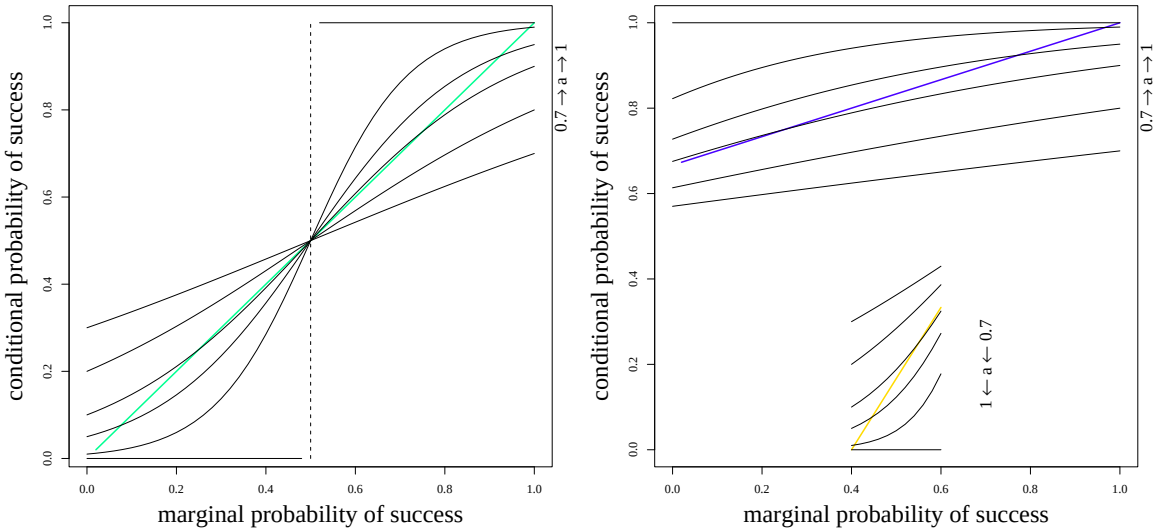

Figure 7.2: Predicted (black lines) and true conditional probability (color line), as a function of the true probability of success in the datum location, for several values of the a parameter. Predicted lines correspond to values of a equal to $0.7, 0.8, 0.9, 0.95, 0.99$ and 1.0. The left-side plot (cyan line) shows the case $\lambda = 0$, whereas the right-side plot explains the cases with correlation coefficient $\lambda = 0.66$ (blue) and $\lambda = -0.66$ (yellow).

3. In the situation of *prediction with a single datum* placed at $\vec{x}_1$, the simple kriging weight used to predict an unsampled location $\vec{x}_0$ from the information in this single sample is the correlation coefficient between them, which ranges between the two previously-described cases: $\lambda = 1$ for a perfect correlation, and $\lambda = 0$ for a null correlation. Finally, a moderate degree of negative correlation can also be modelled, which amounts to slightly weighting up the probability of the non-observed category. Perfect negative correlation is only compatible with $p_1 = p_2 = 0.5$: figure 7.1 shows the relation between the true conditional probability and the estimated probability obtained with logistic indicator kriging with a value $a = 0.95$ ($\alpha = \ln \frac{a}{1-a} \approx 2.8$), for several values of the correlation coefficient. This figure shows also which marginal probabilities and correlation coefficients are compatible with a valid two-way contingency table, *i.e.*with all cells containing a positive probability: the non-allowed combinations are simply not plotted. Figure 7.2 represents some selected predicted probabilities, compared with the true conditional probabilities, for several values of $\alpha = \ln \frac{a}{1-a}$. It can be seen that the final predicted distribution of kriging in the Simplex is *smoother* than the true distribution, in the sense that predictions tend to a when the true conditional probability is above 0.5, and to $1-a$ when it is below. Note in figure 7.2 that the case $a = 1$ yields a degenerate estimator, with events at the predicted location either absolutely sure or impossible.

Bayesian interpretation of the generalized indicator scaling

The last considerations regarding the single datum problem, as well as the definition itself of the generalized indicator function (7.14), show us that essentially, α is a *subjective* assessment of the log-odds of *the observed category against the unobserved ones*. Classical indicator kriging places an infinite odd to this relationship, whereas coordinate indicator kriging leaves it open to the analyst. This value can be fixed previously to kriging, either as a guess or following one of the criteria detailed in this chapter addendum (section 7.7). In this case, kriging will offer a way to combine these odds through a series of weights which minimize an Aitchison error between the prediction and the true probability vector.

But we can also consider leaving α free for all the kriging procedure, since the kriging results do not depend on the value of this parameter, given the proportionality property 7.5. Then, *assuming that the kriging results have no error*, proposition 7.1 and particularly equation (7.28), provides us with D quantities, the vector $\underline{\xi}$, that *may be reasonably interpreted as* the equivalent number of *relative* observed samples in each category: these values always sum up to zero, thus some of them must be negative; then, if we select the minimum of them and add its opposite to all the ξ_k values, we will obtain a set of D positive values (and one of them equal to zero), which we will represent by ξ'_k. In this situation, we can apply the bayesian framework of section 7.7 and update the chosen prior $f^0_{\mathbf{P}}(\mathbf{p})$ by the derived likelihood to obtain a posterior

$$f_{\mathbf{P}}(\mathbf{p}) \propto f^0_{\mathbf{P}}(\mathbf{p}) \prod_{k=1}^{D} p_k^{\xi'_k}.$$

Assuming the prior distribution $f^0_{\mathbf{P}}(\mathbf{p})$ to be of the Dirichlet type, with parameters $\underline{b} = (b_1, b_2, \ldots, b_D)$, then the posterior distribution is also of the same type with parameters $\underline{b} + \underline{\xi'}$. Again, we will have to choose a value from this distribution by taking into account a loss criterion. So, the decision on the value of α is not avoided, but only delayed to this moment.

Note in passing that the global *equivalent sample size* $\xi'_T = \xi'_1 + \xi'_2 + \cdots + \xi'_D$ has no sense, since it is a function of the minimum. Instead, we can consider the relative number of observations in the most-observed category, as the length of the range of $\underline{\xi}$. In particular, at a sampled location, the range is exactly one. This implies that in this model, a sampled location does not receive any information from the samples around it, as was shown in the first simple case of page 152. This is an effect of the exact-interpolator character of kriging, which in this case we regard as a limitation: it would be better to have a way to accumulate concomitant information of nearby samples, *e.g.*the probability of obtaining result A_1 at a sampled location where we observed A_1 should be greater (the event, surer) if this location is surrounded by many sampled locations where the same A_1 was also observed.

7.4 A Bayesian method

Following the ideas of bayesian methods, like those applied by Diggle et al. [1998] to binomial counts (section 3.4.2) or Tjelmeland and Lund [2003] to compositional data, we may want to assume a global *a priori* distribution for the whole compositional RF $\mathbf{P}(\vec{x})$, and update it by all the observations. In contrast, the model presented in section 7.3 (and further justified in this chapter addendum) assumed an individual prior for each sampled location, and updated independently each one of them by the observation at that location.

Haas and Formery [2002] propose a similar model, applied to simulation of proportions of facies, where at each location a Dirichlet distribution is assumed as the marginal model, and the joint model is implicitly constructed by specifying a valid set of auto- and cross-covariance models. Their model is however not used for estimation, but as a simulation tool. To our knowledge, there is no direct and easy definition of a Dirichlet N-variant model, which would in fact be a $(D-1) \times N$-variant distribution.

Instead of using a Dirichlet distribution, with its strong independence structure and the problems to build a joint model for all locations of interest, we propose to use a joint N-variate normal model in the D-part Simplex. Let $\mathbf{p}_N$ denote a block vector of N compositions $\mathbf{p}_n, n = 1, 2, \ldots, N$, where each one has D components. Assume this vector to be jointly normally distributed in the D-part Simplex (definition 2.26), with a density like

$$\log f^0(\mathbf{p}_N|\underline{\theta}_N, \underline{\underline{\phi}}_N) = \sum_{n=1}^{N} \underline{\theta}_n^t \cdot \underline{\pi}_n + \sum_{n=1}^{N}\sum_{m=1}^{N} \underline{\pi}_n^t \cdot \underline{\underline{\phi}}_{nm} \cdot \underline{\pi}_m,$$

where $\underline{\theta}_N = \underline{\underline{\Sigma}}_N^{-1} \cdot \underline{\mu}_N$, being $\underline{\mu}_N = (\underline{\mu}_n)$, where each $\underline{\mu}_n$ is the vector of expectations of $\underline{\pi}_n = (\pi_i)_n = \left(\log \frac{p_i}{p_D}\right)_n$, the vector of $(D-1)$ alr transformations (6.3) of the composition $\mathbf{p}$. Also, with $\underline{\underline{\Sigma}}_N$ the block matrix of cross-covariances $(\sigma_{ij})_{nm} = \mathrm{Cov}\,[(\pi_i)_n, (\pi_j)_m]$ already used in the proof of property 4.7, we may express by $\underline{\underline{\phi}}_N = -\underline{\underline{\Sigma}}_N^{-1}/2$, where matrix $\underline{\underline{\phi}}_{nm} = (\phi_{ij})_{nm}$ represents the interaction between the components of $\underline{\pi}_n = (\pi_i)_n$ and those of $\underline{\pi}_m = (\pi_j)_m$. In geostatistical terms, n, m are called *locations*, and i, j *variables*, or *coordinates* in the approach of this Thesis. If we denote by $\underline{\pi}_N = (\underline{\pi}_n)$ the block vector of these coordinates, then we can write the density also as

$$\log f^0(\mathbf{p}_N|\underline{\theta}_N, \underline{\underline{\phi}}_N) = \underline{\theta}_N^t \cdot \underline{\pi}_N + \underline{\pi}_N^t \cdot \underline{\underline{\phi}}_N \cdot \underline{\pi}_N, \tag{7.30}$$

Before considering the next step, one has to obtain or fix the auto- and cross-covariances of this prior distribution. In most of the situations, this must be estimated, for instance using the considerations of section 7.3.2. However, a much more rigorous approach would be to follow Tjelmeland and Lund [2003] and apply a hierarchical bayesian approach, parametrizing the covariance structure with prior distributions for the range and the covariance at the origin. This will be left for further studies.

We encode the observations in $\underline{J} = (J_{in}) = (I(z_n = A_i))$, a vector with $(D-1)N$ elements, where the first $D-1$ elements are a set of zeroes and ones (*bits*) indicating which category was observed at location $n = 1$, the second block of $D-1$ bits indicates which was observed at $n = 2$, and so on. Note that at most, only one of these bits may be 1 in each block, and even all can be zero if we observed category A_D or if we had no observation there. Moreover, $\underline{M}$ represents a vector of N elements, where M_n shows how many observations we obtained at the n-th location (so, usually at most one).

Proposition 7.2 (Likelihood of the joint sample) *The likelihood of the sample assuming an independent multinomial model at each location is the cell of the N-way contingency table defined by crossing all the observed categories at each location,*

$$L(\boldsymbol{p}_N|\underline{J},\underline{M}) = p_{\underline{J},\underline{M}} = \prod_{n=1}^{N} p_{Dn}^{M_n - \sum_{j=1}^{D-1} J_{jn}} \prod_{i=1}^{D-1} p_{in}^{J_{in}}.$$

Proof: Note that this expression is simply the product of those marginal p_{in} observed at location n, since when $J_{in} = 0$ then p_{in} disappears from the expression, and if we observed category D at location n, then $M_n = 1$, but $J_{jn} = 0$, for $j = 1, 2, \ldots, D-1$, and thus we keep only p_{Dn}. Given that the realizations at each location are considered independent, the joint probability is simply the product of the observed marginal probabilities. □

Proposition 7.3 (Posterior distribution) *With the introduced notation and prior (7.30), the posterior is*

$$f(\boldsymbol{p}_N|\underline{\theta}_N, \underline{\underline{\phi}}_N, \underline{J}, \underline{M}) \propto \exp\left((\underline{\theta}_N + \underline{J})^t \cdot \underline{\pi}_N + \underline{\pi}_N^t \cdot \underline{\underline{\phi}}_N \cdot \underline{\pi}_N + \underline{M}^t \cdot \log \underline{p}_D\right), \quad (7.31)$$

where $\log \underline{p}_D$ is the vector of logarithms of the probabilities of the D-th category at all locations.

Proof: First we rearrange the likelihood,

$$L(\mathbf{p}_N|\underline{J},\underline{M}) = \prod_{n=1}^{N} \frac{p_{Dn}^{M_n}}{p_{Dn}^{\sum_{j=1}^{D-1} J_{jn}}} \prod_{i=1}^{D-1} p_{in}^{J_{in}} = \prod_{n=1}^{N} p_{Dn}^{M_n} \prod_{i=1}^{D-1} \left(\frac{p_{in}}{p_{Dn}}\right)^{J_{in}},$$

and take logarithms of it

$$\log L(\mathbf{p}_N|\underline{J},\underline{M}) = \sum_{n=1}^{N} M_n \log p_{Dn} + \sum_{i=1}^{D-1} J_{in} \log \frac{p_{in}}{p_{Dn}} = \underline{M}^t \cdot \log \underline{p}_D + \underline{J}^t \cdot \underline{\pi}_N.$$

Then, the logarithms of the posterior is, due to Bayes Theorem

$$\log f(\mathbf{p}_N|\underline{\theta}_N, \underline{\underline{\phi}}_N, \underline{J}, \underline{M}) = \kappa + \log f^0(\mathbf{p}_N|\underline{\theta}_N, \underline{\underline{\phi}}_N) + \log L(\mathbf{p}_N|\underline{J},\underline{M}),$$

which directly yields the desired result, given that κ only normalizes the density to sum up to one. □

Note that the posterior density follows an N-block Aitchison's $\mathcal{A}$ distribution (definition 2.28). Once the posterior distribution has been obtained, we look for a $\mathbf{p}_N^*$, representative of the whole density, by choosing any of the loss criteria explained in section 7.7. We will consider here the logistic mean and the maximum.

The logistic mean is not available through any analytical expression, and one must rely on Monte Carlo estimation methods. The process consists on:

1. the simulation of a large sample of size K from the posterior Aitchison's $\mathcal{A}$ distribution; due to property 2.6 about the decomposition of $\mathcal{A}$ distributions as products of a normal and Dirichlet distributions, simulations can be efficiently generated by drawing samples from the prior normal distribution, and applying an acceptance-rejection technique [Lantuéjoul, 2002] to them using the closed likelihood, which is of Dirichlet type;

2. by direct averaging of the coordinates of the simulated set we will obtain an estimate of
$$\underline{\pi}_N^* = \frac{1}{K}\sum_{k=1}^{K} \underline{\pi}_N^{(k)};$$
alternatively, instead of rejecting-accepting each sample $\underline{\pi}^{(k)}$ as a function of its Dirichlet density, this very density may be taken as the weight to use in a weighted arithmetic mean of the simulated values;

3. this will yield an estimate of $\mathbf{p}_N^*$ by application of the inverse alr transformation (6.4) to each block of $\mathbf{p}_n^*$.

Note that, although the simulation should be done on the whole domain $(\mathbb{S}^D)^N$, the averaging and back-transformation steps may be done by blocks. Techniques for simulating large joint normal distributions may be borrowed from geostatistics, in particular, sequential Gaussian simulation.

Property 7.6 (Maximum of the posterior distribution) *The mode of the posterior distribution (7.31) is attained at $\boldsymbol{p}_N$ solving the non-linear system of equations*

$$\underline{0} = \frac{d\log f(\boldsymbol{p})}{d\pi} = \underline{\theta}_N + \underline{J} + 2\underline{\underline{\phi}}_N \cdot \underline{\pi}_N + \underline{\underline{\Delta}} \cdot \underline{M}, \tag{7.32}$$

with $\underline{\underline{\Delta}}$ a $(D-1)N \times N$-identity-block matrix, where each block in the diagonal is a $(D-1)$-element column vector containing the derivative $\frac{d\log p_D}{d\pi} = ((\Delta_i)_{nm})$, and outside the diagonal containing identically-zero column vectors,

$$(\Delta_i)_{nm} = \frac{d\log(p_D)_m}{d(\pi_i)_n} = \begin{cases} 0, & n \neq m \\ (\boldsymbol{p}_{-D})_n, & n = m \end{cases},$$

denoting by $(\boldsymbol{p}_{-D})_n$ the closed composition at location n, and without the last element.

System (7.32) can be simplified if we assume $\underline{\underline{\Delta}} \cdot \underline{\underline{M}} \approx \underline{\underline{0}}$. Following the re-parametrization in terms of means and variances of the original normal distribution on the Simplex, we attain a maximum density at

$$\underline{\pi}^* = \underline{\mu} + \underline{\underline{\Sigma}} \cdot \underline{J},$$

or in other words, the final (rather rawly simplified) maximum posterior estimate is the prior mean $\underline{\mu}$ displaced a unit following the variance-covariance vector $\underline{\Sigma}_{J_i}$ associated to each of the observed categories. Recall that the elements of $\underline{J}$ sum up at most to N. *Proof:* Taking derivatives of the log-density with respect to $\underline{\pi}$ and equating to zero we obtain directly expression (7.32), following the same steps as in property 2.7 regarding the mode of an $\mathcal{A}$ distribution. Note that system of equations (7.32) is a non-linear one, and can be seen as the concatenation of N systems like those appearing in this property 2.7, one for each location, with interactions between locations exclusively due to $\underline{\underline{\phi}}$, a function of the prior covariance model.

Assuming $\underline{\underline{\Delta}} \cdot \underline{\underline{M}} = \underline{\underline{0}}$, we get

$$\begin{aligned} 0 &= \underline{\theta}_N + \underline{J} + 2\underline{\underline{\phi}}_N \cdot \underline{\pi}_N^* \\ -2\underline{\underline{\phi}}_N \cdot \underline{\pi}_N^* &= \underline{\theta}_N + \underline{J} \\ \underline{\underline{\Sigma}}_N^{-1} \cdot \underline{\pi}_N^* &= \underline{\underline{\Sigma}}^{-1} \cdot \underline{\mu}_N + \underline{J} \\ \underline{\pi}_N^* &= \underline{\underline{\Sigma}}_N \cdot \underline{\underline{\Sigma}}_N^{-1} \cdot \underline{\mu}_N + \underline{\underline{\Sigma}}_N \cdot \underline{J} \end{aligned}$$

which finally yields the simplified version. □

In the simplified interpretation, observing a category (*e.g.*A_i) which log-ratio highly correlates with the log-ratio of another one (a high $\mathrm{Corr}\,[\log p_i/p_D, \log p_j/p_D]$) would imply a collateral increase on the certainty of the second (the j-th category); conversely, observing one of two categories with negatively-correlated log-ratios would increase the certainty of the observed one and decrease the certainty of the other; finally, observing one of the categories involved in two non-correlated log-ratios would not affect the other one. In the spatial context, we would have as a particular result that beyond the range (all covariances dropped to zero) any estimation would provide simply the unchanged prior mean. Although this simplified version allows us to interpret what is occurring in the system, it must be taken with caution. On the side of the flaws, note that if we observe category A_D, this information is lost in the simplification and we do not modify the maximum posterior estimate accordingly.

7.5 Case study: estimation of high conductivity hazard

7.5.1 Kriging in the Simplex for generalized indicators

Let us illustrate the techniques here presented with the conductivity data set introduced in section 1.4.1 and already treated in section 3.6. Recall that we considered the

measurements to be the addition of a periodic drift and a real RF. The drift was estimated with equation (3.36), and the residuals for the first 10 days of July 2003 are represented in the upper part of figure 7.3. Deciles were estimated using the whole residuals for July 2003 conductivity series, and they are also represented in the same figure. These deciles are summarized in table 7.1.

Table 7.1: Estimated deciles for the residual conductivity of July 2003, after substraction of the periodic drift (done in section 3.6).

i	probability	decile
1	10%	-191.49
2	20%	-108.03
3	30%	-55.87
4	40%	-2.16
5	50%	24.51
6	60%	44.41
7	70%	67.06
8	80%	100.93
9	90%	143.67

Using these deciles, a set of 10 disjunctive indicators was defined, by application of equation (7.2). We considered a preliminary value $\alpha = 1$ to use in the generalized indicator function (7.15), and the matrix $\underline{\underline{\varphi}}$ of definition of an ilr basis (eq. 6.5) was taken as

$$\underline{\underline{\varphi}} = \begin{pmatrix} 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{-1}{\sqrt{2}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & \frac{-1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-1}{\sqrt{2}} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{-1}{2} & \frac{-1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{-1}{\sqrt{12}} & \frac{2}{\sqrt{12}} & 0 & 0 \\ 0 & \frac{3}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{-1}{\sqrt{24}} & \frac{3}{\sqrt{24}} & 0 \\ \frac{4}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{-1}{\sqrt{40}} & \frac{4}{\sqrt{40}} \end{pmatrix},$$

Each row of this matrix defines an element of the basis of the Simplex used in this case, obtained by taking closed exponentials. Note that the first 5 vectors balance the left and right categories at each side of the median, and the last four vectors balance each one of these groups from the center to the tails of the distribution. This matrix is expected to particularly describe the symmetry of the distribution of the conductivity residuals, and the importance of the tail against the mode. It was chosen to minimize the impact of a bad estimation in any category, since each has no influence on at least 4 vectors.

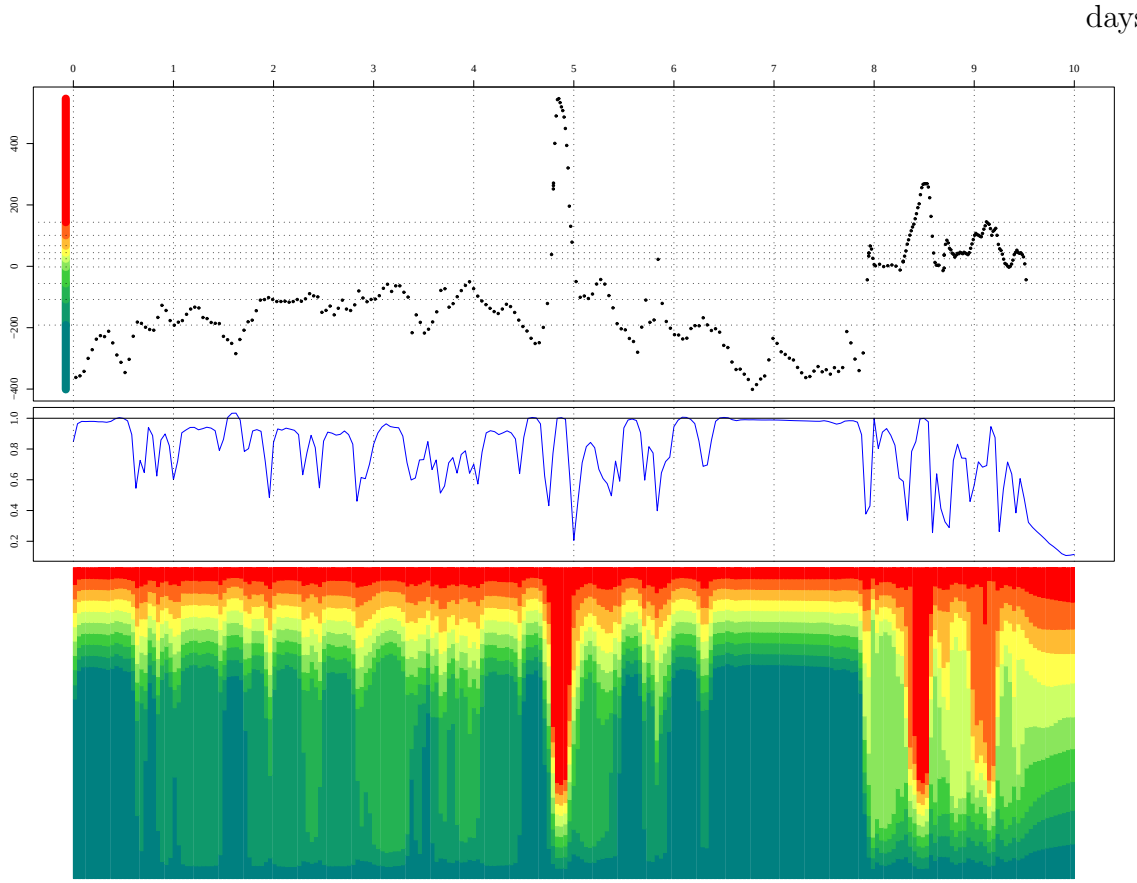

Figure 7.3: In the upper part, residual conductivity data set of the first 10 days of July 2003, with the estimated deciles. In the lower part, estimated distribution at each hour, discretized according to the deciles of the upper part. In the middle part, equivalent relative number of observations of the most favored category against the least favored one in the predicted locations.

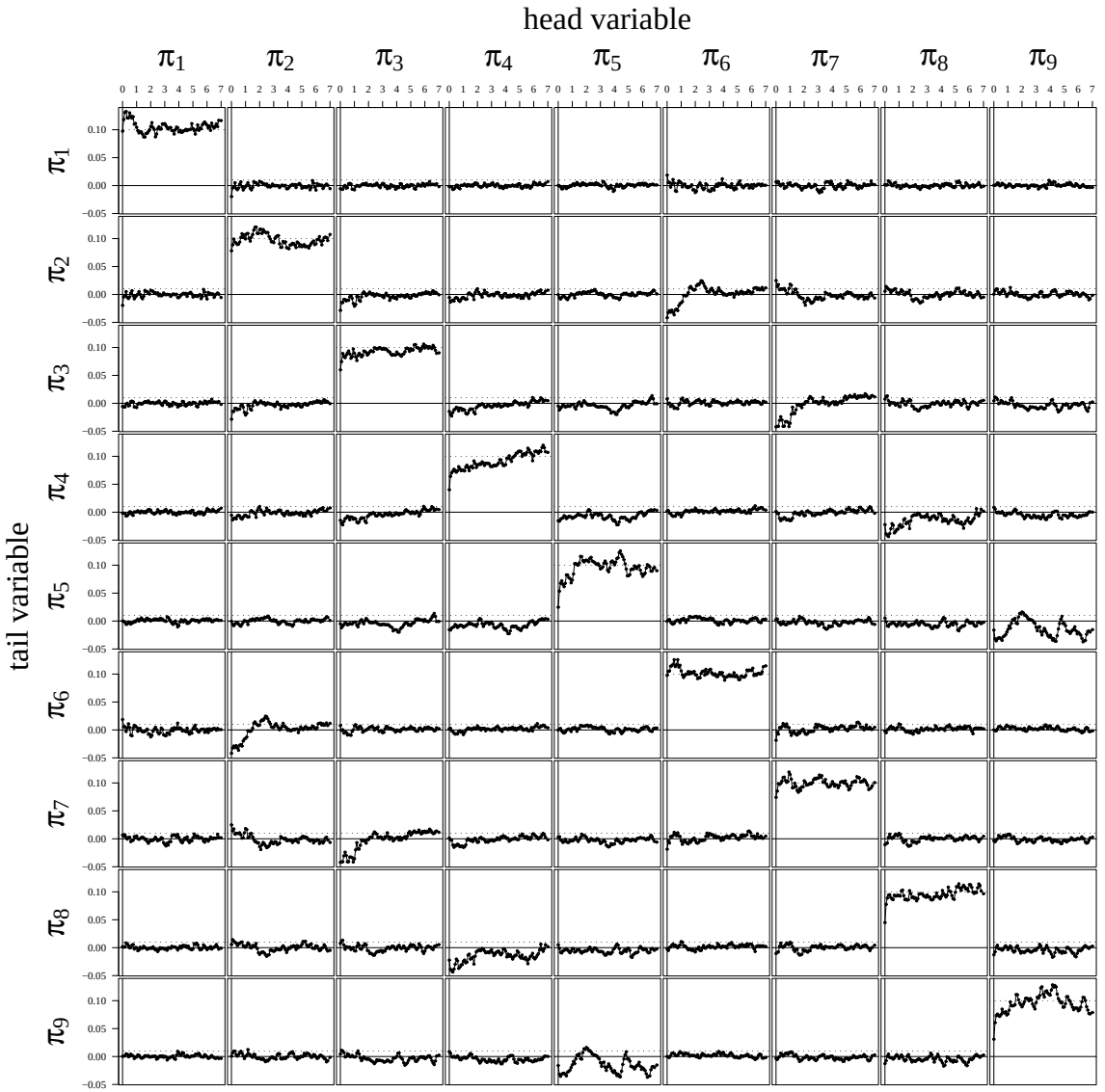

Figure 7.4: Matrix of generalized indicator auto- and cross-variogram plots. Note the symmetry of the figure, since $\gamma_{ij}(h) = \gamma_{ji}(h)$.

Table 7.2: Parameters for the models defining each auto-covariance function of the coordinates of the generalized indicator function. Short and long ranges are described through spherical models, whereas the hole effect is taken as a non-dampened cosine.

		short range		long range		hole effect	
variable	nugget	sill	range	sill	range	sill	period
π_1	0.02	0.070	0.25			0.02	1.5
π_2	0.07	0.023	0.23	0.011	2.29		
π_3	0.05	0.037	0.25	0.035	25.98		
π_4	0.03	0.028	0.09	0.036	2.017		
π_5		0.070	0.22	0.030	4.125		
π_6	0.02	0.070	0.25			0.02	1.4
π_7	0.05	0.050	0.50				
π_8	0.02	0.068	0.21	0.004	1.170		
π_9	0.02	0.031	0.12	0.051	1.848		

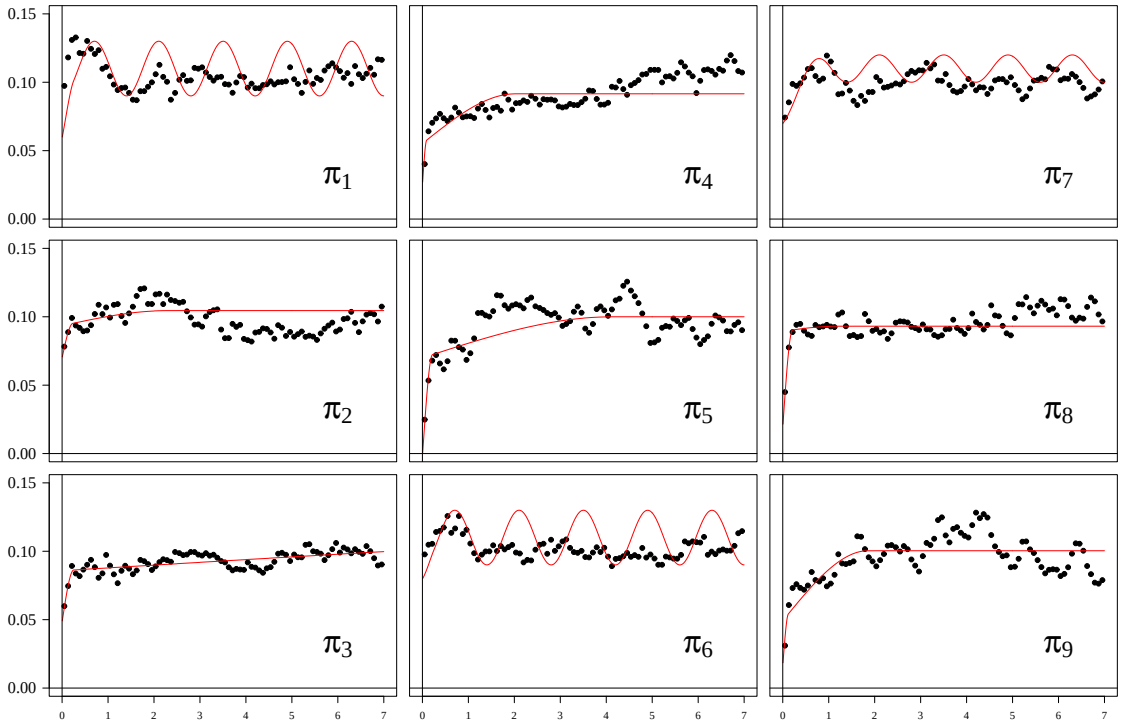

Figure 7.5: Generalized indicator variogram plots, with fitted models.

The resulting 9 coordinates of the generalized indicator function were treated through function `variogram` to compute auto- and cross-variograms, shown in figure 7.4. The low impact expected for the cross-variograms [Goovaerts, 1994] and the fact that modelling 36 functions will obviously introduce much more arbitrariness than modelling only 9 (the direct variograms) suggests to apply simple kriging to each coordinate independently. Thus, we modelled only these direct variograms, described in table 7.2 and plotted in figure 7.5.

Mean values for all these coordinates were taken as 0, given the fact that the original categories are equally probable: using equation (7.16), it is easy to show that

$$\underline{0} = \mathrm{E}\left[\underline{\pi}\right] = \alpha \cdot \underline{\underline{\varphi}} \cdot \mathbf{p} = \alpha \cdot \underline{\underline{\varphi}} \cdot \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10},\right).$$

Simple kriging was conducted using function `predict.gstat`. It yielded results plotted in figure 7.6, jointly with the generalized-indicator-transformed data set. Applying the transposed ilr matrix (introduced in section 6.5) to these kriged values we obtain the vector of exponents $\underline{\xi}$ of proposition 7.1. For each predicted location, the lengths of their ranges inform us of the relative number of observations in favor of the preferred category in front of the least favored one. This might be interpreted as an equivalent relative sample size, and is plotted in the middle part of figure 7.3. Multiplying them by a suitable value, *e.g.*$\alpha = 3.04$, and taking closed exponentials we obtain the final predicted distributions at each location, plotted in the lower part of figure 7.3. Note that this value of $\alpha = 3.04$ corresponds to a value of $a = 0.7$ in equation (7.14), which implies that after observing a given category, we will consider it to have a probability of 70% of occurrence, and the remaining 9 categories will be equally probable with a total probability of 30%. Figure 7.7 shows the final predictions obtained with other suitable α values, and the real influence of this α value: it finally conditions the certainty in the obtained predictions.

A closer look at the plots in figure 7.3 shows us some interesting properties of the logistic generalized indicator approach. The equivalent sample size, for instance, gives us an assessment of the reliability of the prediction itself: those areas were the residual conductivity falls constantly in the same category present higher equivalent sample sizes, which might be even slightly above one (see *e.g.*the green area around 1.5 days), whereas this sample size drops to near zero either at distances from the samples beyond the range of the variograms (*e.g.*the second half of the 10^{th} day), or when the residual conductivity suffers a sudden change (*e.g.*the end of the 5^{th} day). Note that the equivalent sample size rarely rises above one.

This method has been applied to the whole residual conductivity series of July 2002 and July 2003, using the same indicator levels of table 7.1, their covariance models of table 7.2 and figure 7.5, and a mean value of zero for all RFs. These residual conductivity series and the results of the method (using $\alpha = 3.04$) are plotted in figure 7.8. Some of these estimated distributions for the residual conductivity have been represented in figure 7.9. Note that in most of the distributions, a category is preferred above the others, but the others are more or less equally probable. The exceptions are:

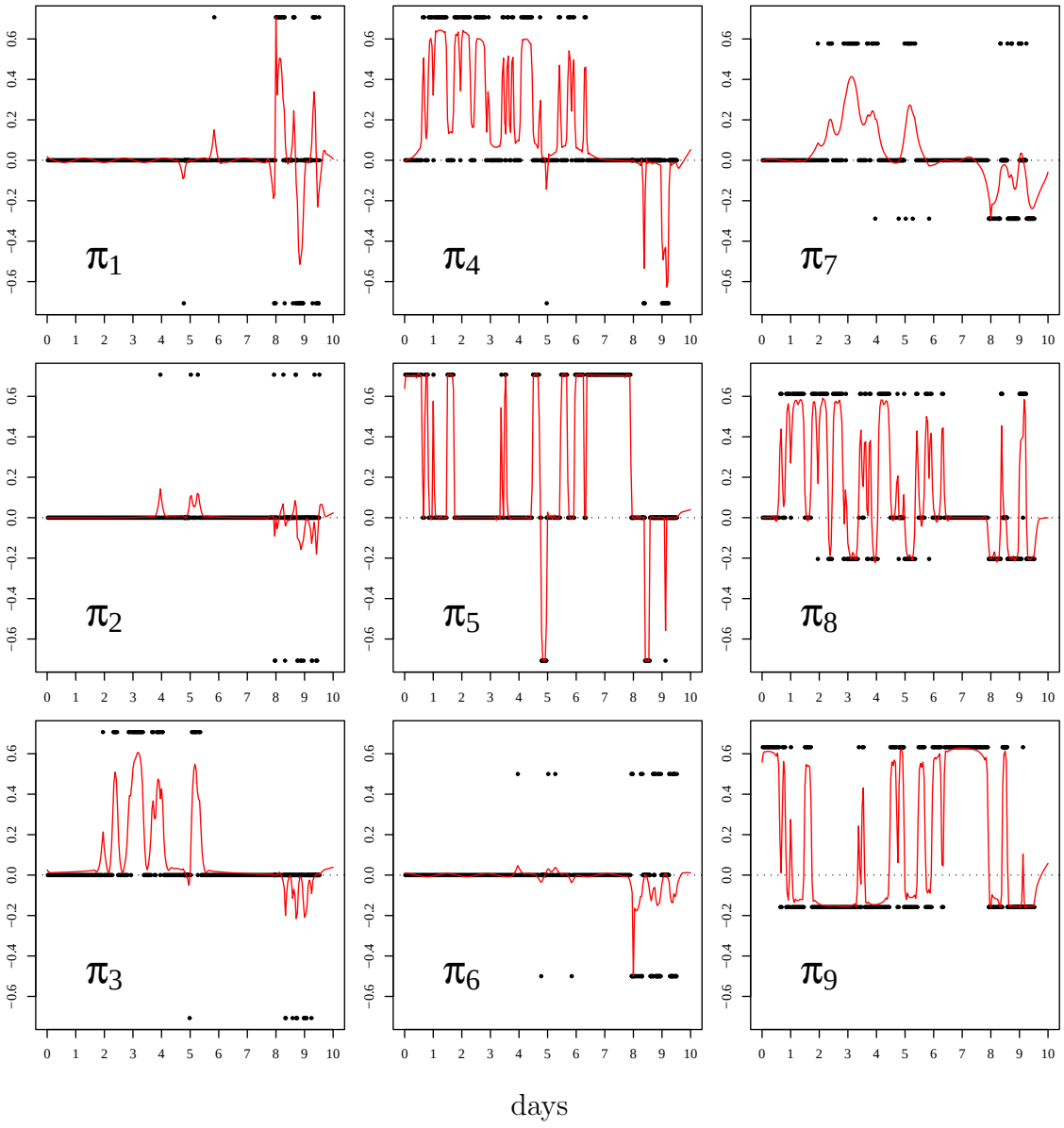

Figure 7.6: Generalized indicator values (dots) and simple kriging predictions (line).

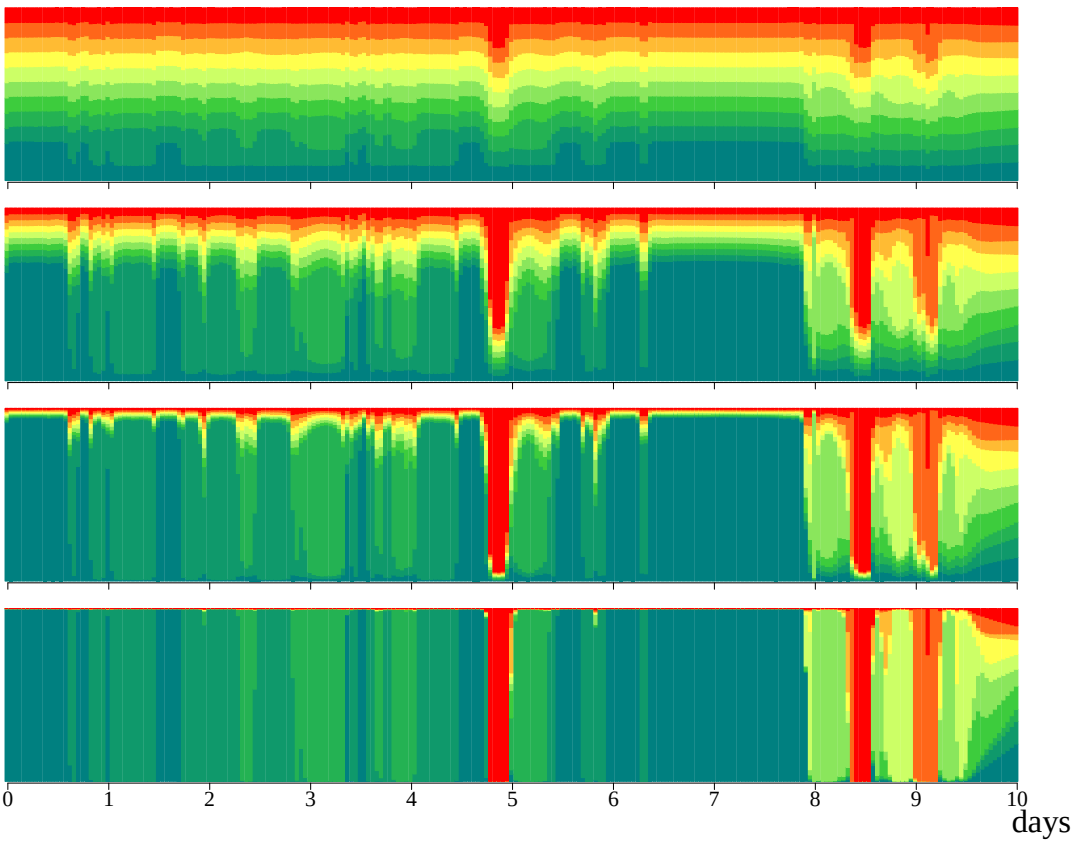

Figure 7.7: Final distribution predictions obtained (from top to bottom) with values of α equal to 1, 3.04, 5.14 and 13.71, corresponding to a values of 23%, 70%, 95% and 99.999%.

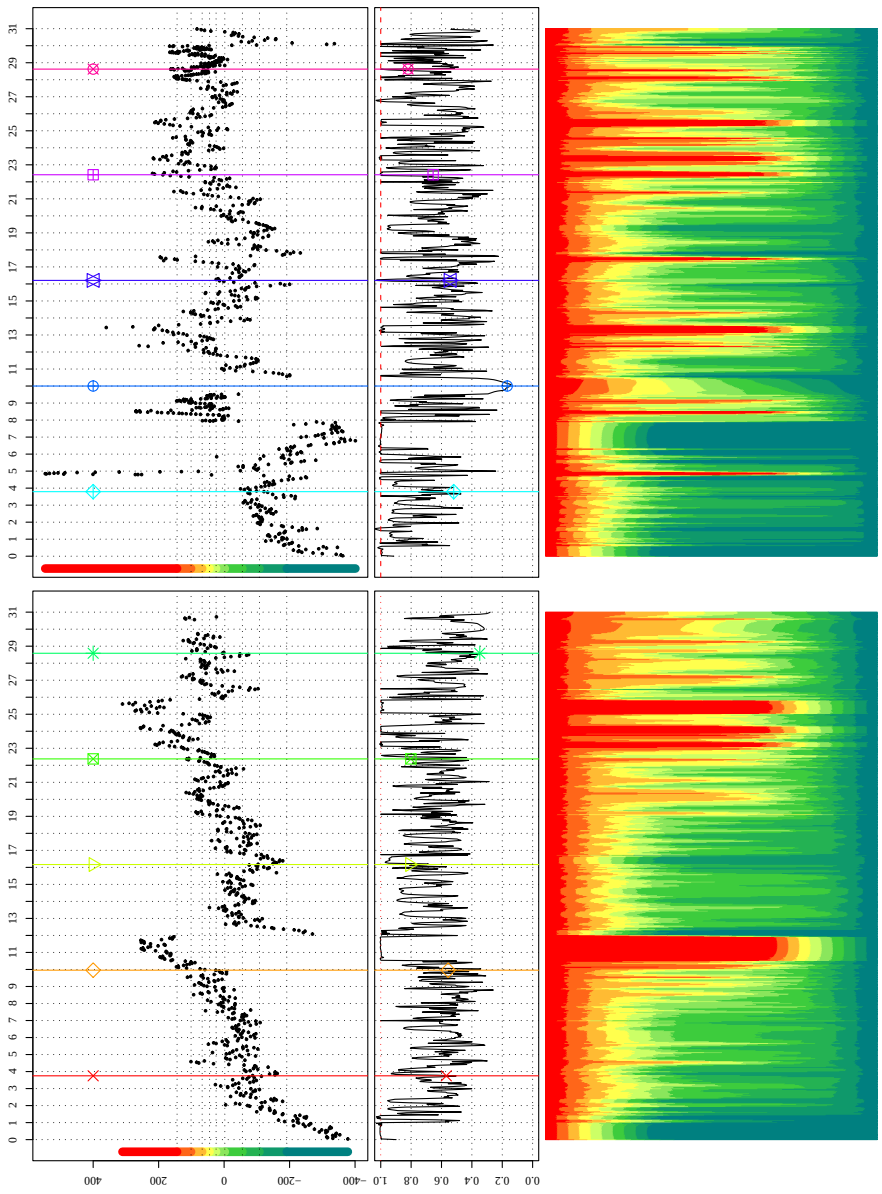

Figure 7.8: In the upper part, residual conductivity data set of the whole months of July 2002 (left) and 2003 (right), with the estimated deciles. Recall that these deciles were obtained using only the July 2003 series. In the lower part, estimated discrete distribution at each hour. In the middle part, equivalent relative number of observations of the most favored category against the least favored one in the predicted locations. Vertical color lines in the upper and middle plots mark some time moments, which estimated probability distribution is shown in figure 7.9; note that in the middle plots, the symbols are placed in the value of the equivalent sample size of each of these estimated distributions.

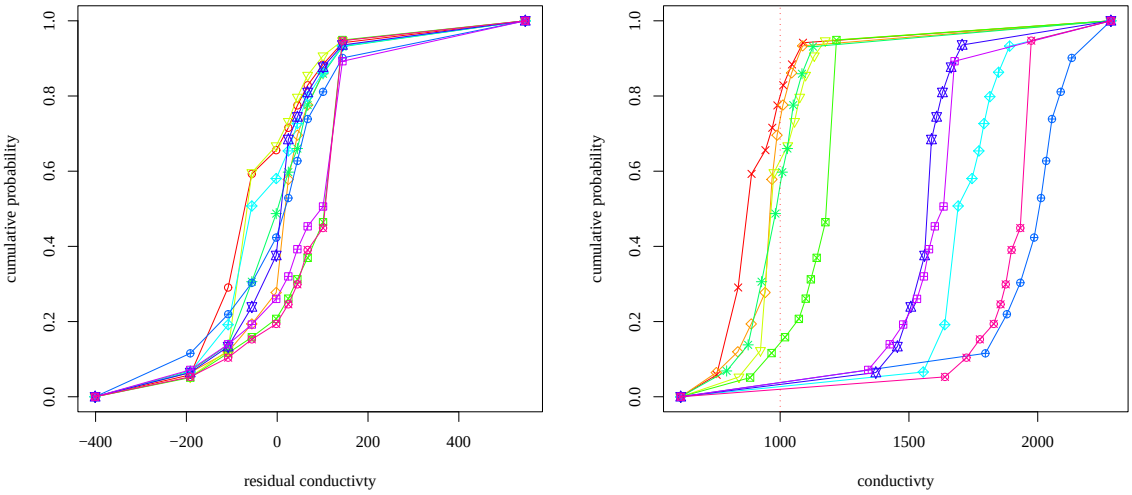

Figure 7.9: Estimated distributions of some selected prediction moments (see figure 7.8 for the corresponding symbols), for the residual conductivity (left) and the re-trended one (right), using the estimated regression trend of table 3.2. In the second plot, a vertical dotted line shows the reference value of $1000\mu S/cm$.

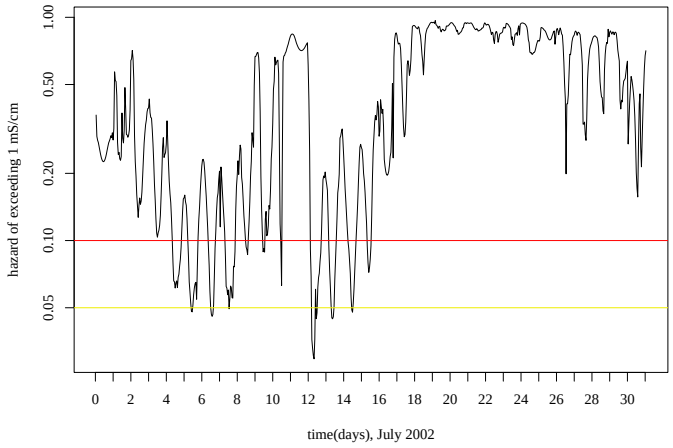

Figure 7.10: Estimated hazard of exceeding $1000\mu S/cm$ of conductivity, obtained using a linear spline to interpolate the discrete version of the probability distribution provided by figure 7.8.

a curve marked with $\oplus$ symbols (electric blue), which is almost 1/10 for all categories (the mean of the vector RF), and two with symbols $\times$ (red) and $*$ (green-blue), showing two equally favored categories.

Finally, these re-trended distributions were computed at each predicted location, and we obtained from them the hazard of being above 1000 $\mu S/cm$, using the same linear spline represented in figure 7.9 (right). This was done only for July 2002, because conductivity was always above this threshold during July 2003. This hazard is plotted in figure 7.10.

7.5.2 Kriging in the Simplex for a single indicator

Given that the interest lies on a single category of the original (non-detrended) conductivity series, it seems also reasonable to use the presented methodology with a single threshold of $1000\mu S/cm$, only for July 2002, although this is not a stationary series. The generalized indicator obtained using expression (7.19) is in this case a simple

$$p_2^* = \left\{ \begin{array}{ll} 0.95, & j_2(\vec{x}) = 1 \\ 0.05, & j_2(\vec{x}) = 0 \end{array} \right. ,$$

where p_2 represents the estimated probability of being above a threshold at an observed location, and $j_2 = 1$ when the observation was actually above the threshold, or zero otherwise. The coordinate of such a generalized indicator becomes

$$\pi^* = \log \frac{0.95}{0.05} \left\{ \begin{array}{ll} +\frac{1}{\sqrt{2}}, & j_2(\vec{x}) = 1 \\ -\frac{1}{\sqrt{2}}, & j_2(\vec{x}) = 0 \end{array} \right\} = \alpha \frac{1}{\sqrt{2}} \left\{ \begin{array}{ll} +1, & j_2(\vec{x}) = 1 \\ -1, & j_2(\vec{x}) = 0 \end{array} \right. , \tag{7.33}$$

with $\alpha = 2.94 = \log \frac{0.95}{0.05}$, the log-odd of the probabilities p_2 and p_1, of success against failure, *which we subjectively assign* to an actually-observed success. This value of 0.95 was suggested by figure 7.2.

Considering $\alpha = 1$, we obtained the so-called canonical generalized indicator, which was used in the actual computations. An experimental variogram was computed with function `variogram`, and a model was fitted to it. Both are plotted in figure 7.11, and the model is described by

$$\gamma(h) = 0.02 + 0.365 \cdot \mathrm{Exp}(a = 2.5) + 0.035 \cdot \mathrm{Hol}(a_t = 1), \tag{7.34}$$

with h, a and a_t measured in days.

The threshold was exceeded 335 times of 725 observation, which would yield a direct estimation of the probability of exceedance of 0.46. Given that the data set is not an independent one and conventional statistics may be biased, it seemed reasonable to parsimoniously assume a probability of 0.5 for both events of being below and above the threshold. Then, equation (7.20) yielded a mean value of zero, which was used in simple kriging. Using function `predict.gstat`, a prediction for π (the coordinate of the generalized indicator function) was computed, one at each full hour during the

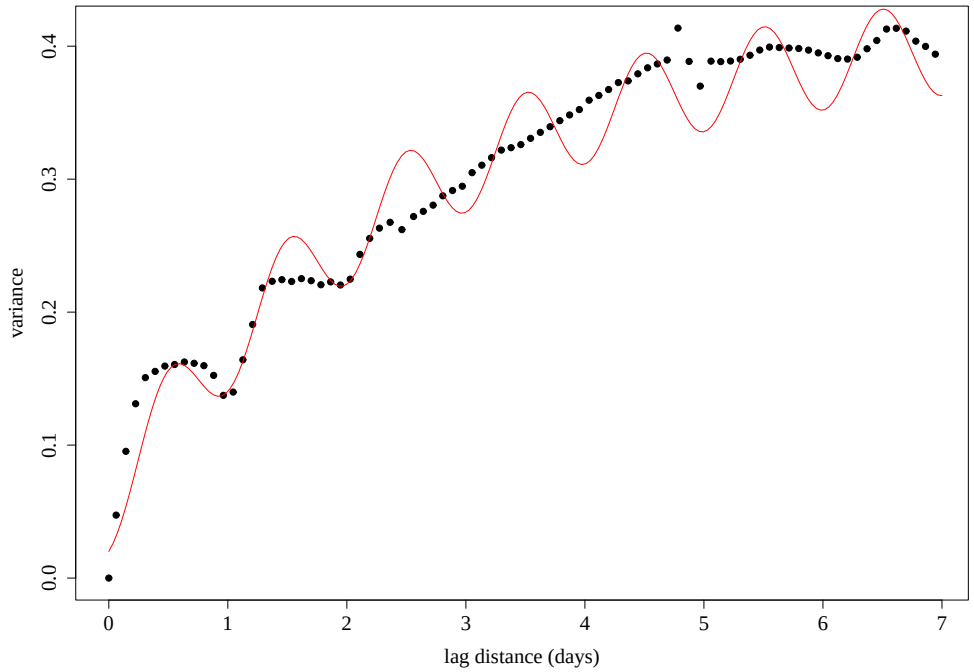

Figure 7.11: Experimental variogram (dots) and fitted model (line) of the coordinate of the generalized indicator function of conductivity above/below the threshold $1000\mu S/cm$. Note that a better fit would be obtained if the hole effect was dampened, something not allowed by the software used.

month of July 2002. Results were then scaled by the value $\alpha = 2.94$ and multiplied by the matrix of coordinates $\underline{\underline{\varphi}}$. A prediction for the pair of probabilities (p_1, p_2) was obtained by

$$(p_1, p_2)^* = \mathcal{C}\left(\exp(\alpha\varphi_{11}\pi^*), \exp(\alpha\varphi_{12}\pi^*)\right). \tag{7.35}$$

Results are plotted in the lower part of figure 7.12. Note that the conductivity data set is plotted in the upper part of this figure, whereas the equivalent number of observations in favor of the preferred category is plotted in the middle. This equivalent number may be computed in this simple case as

$$\xi' = |\pi^* \cdot (\varphi_{11} - \varphi_{12})| = \left|\sqrt{2}\pi^*\right|.$$

The final prediction of these probabilities of being below or above the threshold (figure 7.12) should be compared with the estimation of the hazard of exceeding the threshold of 1000 $\mu S/cm$ of conductivity computed with an estimated discrete version of the probability distribution (figure 7.10), and with a joint Gaussian assumption (figure 3.12).

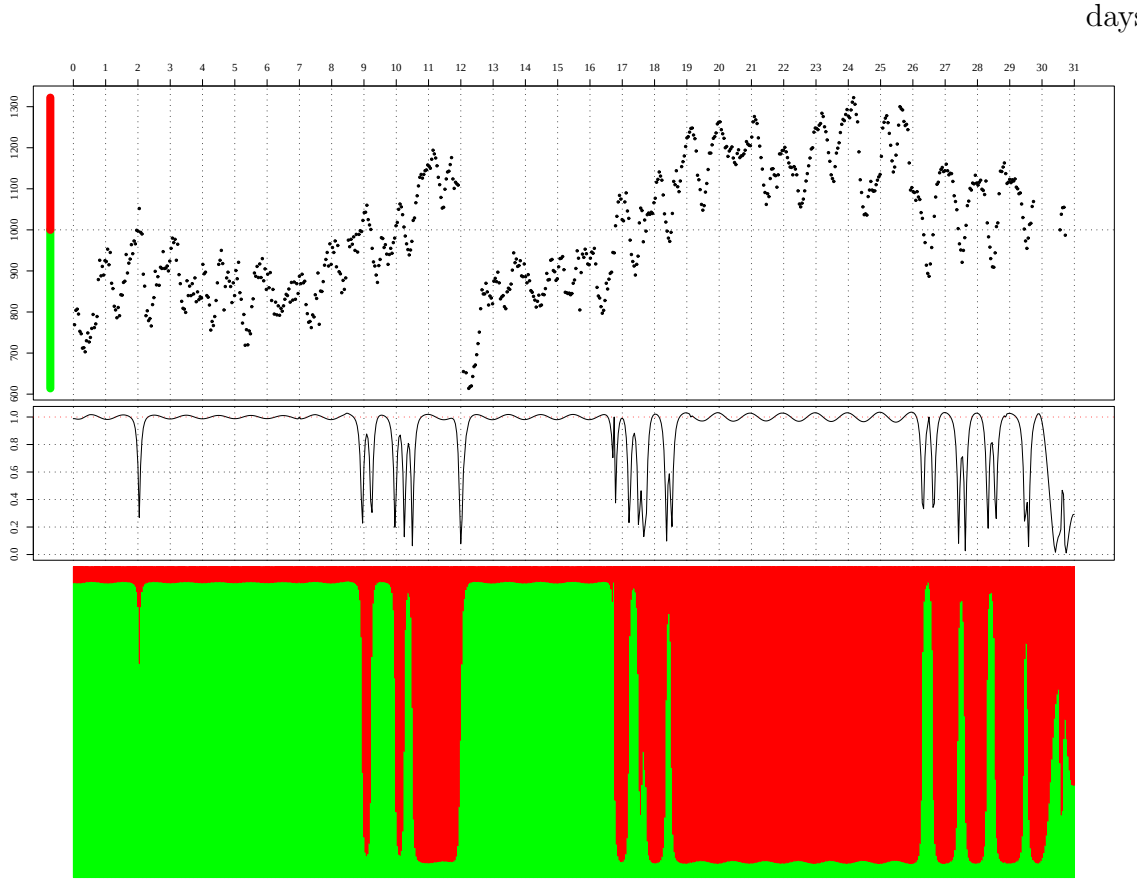

Figure 7.12: In the upper part, conductivity series of July 2002, with a single cutoff at 1000 $\mu S/cm$. In the middle, equivalent number of observations in favor of the preferred category; note that this drops to almost zero in the regions were observations fluctuate around the threshold. In the lower part, final predicted (discrete binomial) distribution, where the probabilities of being above and below the threshold are marked respectively in red and green.

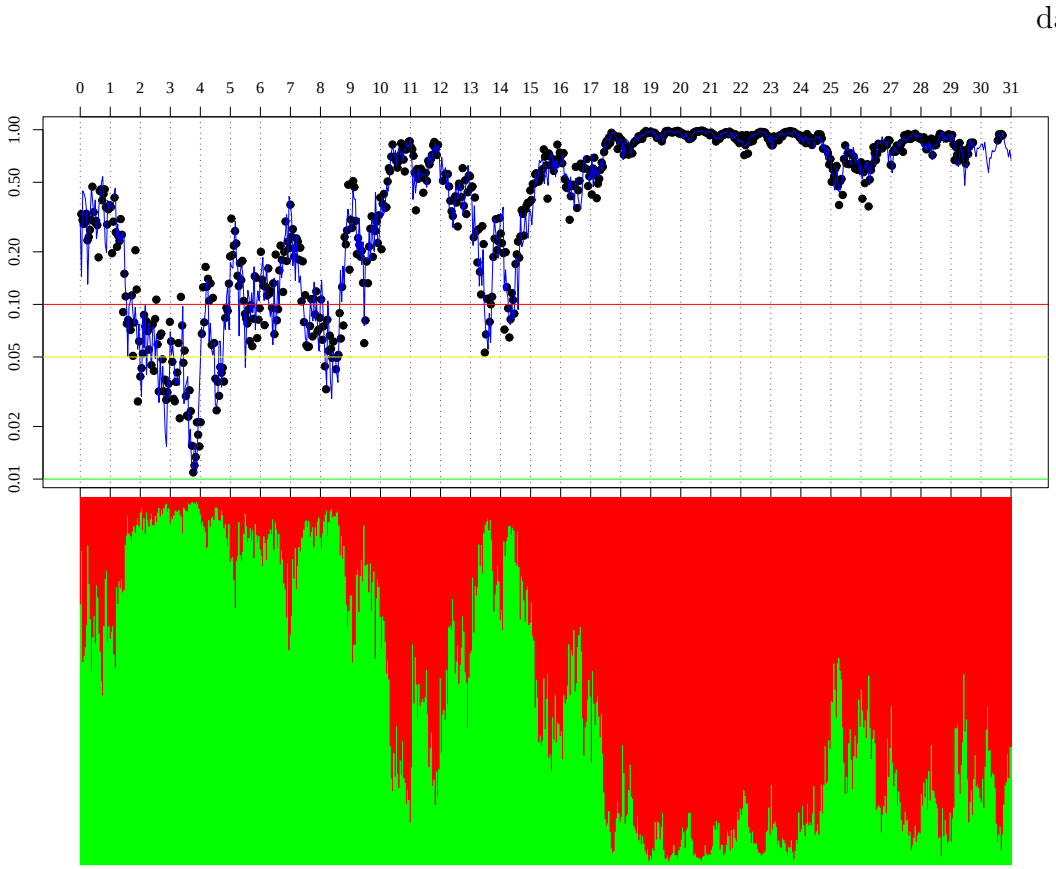

Figure 7.13: In the lower plot, probability estimates of being above (red) and below (green) the threshold of 1000 $\mu S/cm$ for the predicted nodes, one each hour. Note the slight periodicity of high probabilities of exceedance (red spikes) of approximately 24 hours in areas which are mainly below it (green areas). In the upper plot, estimated hazard of being above the threshold, for both the observed (black dots) and the predicted nodes (blue line). Note the log-scale of this upper plot, and the three horizontal reference lines.)

7.5.3 Bayesian estimation in the Simplex for a single indicator

With the structural analysis of the last example (the variogram of figure 7.11 and a mean value of zero), we may assume that the probability vector $\mathbf{p}(t) = (p_1(t), p_2(t))$ follows a Gaussian model in the two-part Simplex, which means that $\pi = \frac{1}{\sqrt{2}} \log \frac{p_2}{p_1}$ follows a joint normal model with zero mean, and a covariance described by the variogram of equation (7.34). We assume that the vector $\underline{\pi} = (\pi(t_n))$, where $n = 1, 2, \ldots N_o + N_p$, with $N_o = 725$ (the number of observed locations) and $N_p = 745$ (the number of predicted locations), *a priori* follows a multivariate normal distribution with mean value a vector of zeroes, and covariance matrix $\underline{\underline{\Sigma}}$ defined by the variogram (7.34).

Given an observation of the true conductivity $Z(t_n)$ at $t_n, n = 1, 2, \ldots N_o$, we may easily define the indicator of being above the threshold

$$j_2(t_n) = \begin{cases} 1, & Z(t_n) \geq 1000 \mu S/cm \\ 0, & Z(t_n) < 1000 \mu S/cm \end{cases} ,$$

which forms a vector $\underline{J} = (j_2(t_n))$ of observations of the indicator RF, with a probability parameter *a priori* distributed according to the normal model of the last paragraph. Using then proposition 7.3 we get an expression for the posterior distribution of this indicator RF. Recall that this distribution is of Aitchison type (definition 2.28).

From this distribution, we may choose as representative value (our final estimate) either its maximum or its mean. The mean value may be estimated by a Monte Carlo procedure, *e.g.*that already outlined in page 158. Using the sequential Gaussian simulation available in function `predict.gstat`, we simulated a large sample (with 10000 samples) of the prior distribution. For a given simulation, the Dirichlet density (properly a beta distribution) of the obtained value was computed at each observed node, using as parameters the observations j_2 and j_1 at that node. The product of all these densities defines the joint likelihood of an observation like that simulation: this likelihood gives us the weight of such a particular simulation. The weighted arithmetic average of all of the simulations of π is our mean estimate. The final probabilities of being above and below the threshold will be estimated by application of equation (7.35).

Results are plotted in figure 7.13, where we have represented the predicted nodes in two different ways. From one side, a log-hazard plot of being above the threshold shows the final estimated hazard for both the observed and the predicted locations. This plot should be compared with the left one in figure 3.12 (obtained under assumption of joint Gaussianity of the conductivity measures) and with figure 7.10 (obtained after estimating the probability function by approximating it through its deciles). From the other side, a barplot shows the final probability of being above and below the threshold only for the predicted nodes, since they are regularly spaced. This second plot should be compared with figure 7.12.

These comparisons show us that kriging in the Simplex applied to a single indicator mainly captures information on the observation of the indicators themselves, being a kriging technique, and there is no accumulation of information in the oversampled

areas. Contrarily, the bayesian estimation on the Simplex of a single indicator yields more extreme probabilities in highly sampled areas, and it includes more information (*e.g.*24h-periodicity) coming from the variogram, which forms the *a priori* model assumption.

7.6 Remarks

In this final chapter we applied kriging in the Simplex to estimate *finite discrete distributions*, since they could be viewed as compositional vectors. We took profit of some old and (relatively) new comparisons between the structure of the Simplex and classical operations for probabilities, like bayesian updating or information measures, and kriging was showed to be applicable in a straightforward fashion. It kept all its properties of unbiasedness, minimal variance and conditional expectation.

However, we pretended to offer an alternative to indicator kriging, which is usually applied to estimate a *finite discrete* version of the cumulative probability function of a RF at an unsampled location. We devised a two-step stochastic model, where we observe realizations of independent multinomial variables which proportion parameter form a compositional RF. Adapting kriging in the Simplex to this formalism required the introduction of design matrices (or a bayesian approach developed in the addendum). Finally we obtained a method which combines kriging and bayesian updating, so that a prior distribution for the unknown probability vector was updated by a multinomial likelihood where the *equivalent relative number of observation* was obtained from kriging. The method always yields admissible probabilities, with an account of the uncertainty attached to them, due to the Bayesian interpretation.

A second method was derived, which can be viewed as a bayesian refinement: this essentially is the joint updating of a general prior by the whole set of observations. The prior may be derived from the same hypothesis sustaining the first method. Results do not differ significantly between them, although the bayesian method seems to better capture information coming from the structural analysis.

The uncertainty of both methods has shown to be generally high, due to the small sample size derived from the predictions. Uncertainty is complementary of information: when only one sample at each location is available, information is sparse, and we can only increase it by assuming models: the bayesian method tried so, by assuming a model for the log-ratios of probabilities, yet keeping the model of the original variable general enough. Not too much should be expected of model-free techniques in geostatistical applications: they are highly uncertain by nature.

7.7 Addendum: Bayesian estimation of probability vectors

Here we present an alternative procedure to the ad-hoc generalized indicator approach used from section 7.3 on. This approach is based on Bayesian estimation (section 2.4.2), given that this technique allows us to complement the scarcity of the sample, *a single realization*, with prior information on the plausible values of $\mathbf{P}(\vec{x})$. In this section, we do not use the spatial dependence in any sense.

Binomial estimation

In the most simple situation, we are interested in determining the probability that $Z(\vec{x}_n)$ is above or below a single cutoff z_1, or in other words, in the probability of success ($p = p(\vec{x}_n)$) of a binomial random variable given the observation of that binomial variable exactly once: $j(\vec{x}_n)$. In a bayesian framework, one has to define a prior distribution for this probability of success, $f_P^0(p)$, update it by the likelihood of the sample to obtain a posterior distribution

$$f_P(p) \propto f_P^0(p) \cdot p^{j(\vec{x}_n)}(1-p)^{1-j(\vec{x}_n)}, \tag{7.36}$$

and finally select from this posterior distribution an estimate of p^* accordingly to a previously specified loss criterion. There are then two decisions in the hands of the analyst before an estimate can be obtained: the prior and the loss criterion to use.

Leonard and Hsu [1999, p. 134-146] suggest that scientists should never consider that they have no prior information on a phenomenon. Even when nothing can be really said, they can choose between a range of prior distributions, considered as non-informative:

- a classical decision in the face of no information is assuming the parameter p to have a prior uniform distribution in its range of variation, thus $p \sim \mathcal{U}(0,1)$;
- Jeffreys suggested to use a prior distribution of the beta family [Abramovitz and Stegun, 1965] with parameters $a_1 = a_2 = \frac{1}{2}$, due to the fact that Fisher information
$$F(a) = -\mathrm{E}\left[\frac{\partial^2 \log f(a)}{\partial a^2}\right],$$
provided by this prior distribution is unaltered for most regular transformations of p; also, Leonard and Hsu [1999] notice that the posterior (7.36) is quite appealing when $j(\vec{x}_n)$ is either zero or one, and they deem as excellent the frequency properties of the confidence intervals computed with the posterior distributions;
- these two prior distributions are in fact special cases of the beta distribution $\beta(a_1, a_2)$, with parameters a_1, a_2 positive; another classical option is to take a so-called *vague prior*, a beta with $a_1, a_2 \to 0$; however, this vague prior may not

be adequate for this case, because the posterior obtained with a single datum (7.36) will not be a proper distribution;

- one can choose the parameter to be logistic-normally distributed with a real mean and a positive variance, $p \sim \mathcal{LN}(\mu, \sigma)$; again, a vague prior could be obtained by making $\sigma \to \infty$, although the posterior obtained would be in this case also an improper distribution

- finally, given this last option, the parameter can be also a priori distributed as a normal on the 2-part Simplex $\mathbf{p} = (p, 1-p) \sim \mathcal{N}_{\mathbb{S}^2}(\mu, \sigma)$, which implies that its coordinate with respect to the Aitchison geometry (7.12) is normally distributed $\pi \sim \mathcal{N}(\mu, \sigma)$ where

$$\pi = \langle \mathbf{p}, \mathbf{e} \rangle_A = \frac{1}{\sqrt{2}} \log \frac{p_1}{p_2} = \frac{1}{\sqrt{2}} \log \frac{p}{1-p}.$$

  Note that these two last cases are equivalent, except for the factor $\frac{1}{\sqrt{2}}$ and the change of the Lebesgue measure [Mateu-Figueras et al., 2003].

When the prior distribution follows a beta model $\beta(\alpha_1, \alpha_2)$, then the posterior distribution (7.36) follows also a beta model $\beta\left(\alpha_1 + j(\vec{x}), \alpha_2 + (1 - j(\vec{x}))\right)$. This good analytic property is not satisfied by the logistic-normal or the normal in the Simplex prior distributions, since their posteriors do not follow any normal distribution, which call for numerical methods to handle them.

Proposition 7.4 *The selection of a value of p^* from this posterior depends on both the chosen loss criterion and the measure used with p:*

1. *the maximum is chosen when the loss function is a Dirac delta,*

2. *the median minimizes the absolute value of the error;*

3. *the arithmetic mean minimizes the squared substraction (Euclidean) error, and corresponds to the expectation of the posterior under a classical Lebesgue measure*

4. *the logistic mean minimizes the squared Aitchison error, and corresponds to the expectation of the posterior under a logistic measure (2.3); given the general context of this alternative to indicator kriging, this criterion should be the preferred one.*

The final estimator of p^* as a function of the prior and the loss function is of the form

$$p^* = \begin{cases} a, & j(\vec{x}) = 1 \\ 1-a, & j(\vec{x}) = 0 \end{cases};$$

where a is detailed in table 7.3.

Proof: Except for the last one, the loss function criteria are standard, and proofs of these propositions will not be included here. Regarding the logistic one, we look for the value $\mathbf{p}^*$ such that $\mathrm{E}\left[d_A^2(\mathbf{p}^*, \mathbf{P})\right]$ is minimal:

$$\mathrm{E}\left[d_A^2(\mathbf{p}, \mathbf{P})\right] = \int d_A^2(\mathbf{p}, \mathbf{P}) dF(\mathbf{P}),$$

which is equivalent to the concept of metric variance around $\mathbf{p}$ introduced by Pawlowsky-Glahn and Egozcue [2001], or to the trace of the variance matrix of $\mathbf{P}$ around $\mathbf{p}$. Then, these authors proof that the minimum is achieved by the metric center,

$$\mathbf{p}^* = \mathrm{E}_A\left[\mathbf{P}\right].$$

The values of table 7.3 are obtained directly using property 2.3 (mode and mean of a Dirichlet-distributed variable under a classical real geometry), property 2.4 (mode of a Dirichlet-distributed variable under a logistic geometry), or applying numerical integration methods (to compute the median with a Jeffreys prior, and the logistic mean). □

Table 7.3: Estimators p^* for several loss criteria and prior distributions. The table contains the estimator when $j(\vec{x}) = 1$; for $j(\vec{x}) = 0$, it is obtained taking $1 - p^*$

	uniform prior	Jeffreys prior
maximum posterior (classical)	1	1
maximum posterior (logistic)	2/3	3/4
median	$1/\sqrt{2}$	0.836806
arithmetic mean	2/3	3/4
logistic mean	~ 0.73	~ 0.88

Multinomial estimation

Following the same approach in the case of multinomial vectors, we are interested in the probability that $Z(\vec{x}_n)$ takes each one of its possible outcomes $\{A_1, A_2, \ldots, A_D\}$. Thus we want to estimate the components of a vector $\mathbf{p}(\vec{x}_n) = (p_1, p_2, \ldots, p_D)$ from a single observation, which may be encoded in the vector of disjunctive indicators $\mathbf{j}(\vec{x}_n)$ (7.2). Given a prior distribution of this vector, $f_{\mathbf{P}}^0(\mathbf{p})$, the posterior distribution is

$$f_{\mathbf{P}}(\mathbf{p}) \propto f_{\mathbf{P}}^0(\mathbf{p}) \cdot \prod_{i=1}^{D} p_i^{j_i(\vec{x}_n)}.$$

Possible choices for the prior distribution are again either

- the Dirichlet $\mathcal{D}(a_1, a_2, \ldots a_D)$, the multidimensional version of the beta distribution, including the case of uniformly-distributed values on the Simplex $\mathbb{S}^D$ when $a_i = 1$; in this case, the posterior distribution is from the same class $\mathcal{D}(a_i + j_i(\vec{x}))$;

- the additive logistic normal $\mathcal{ALN}(\underline{\mu}, \underline{\underline{\Sigma}})$[Aitchison, 1982], or its counterpart normal on the D-part Simplex [Mateu-Figueras et al., 2003], which give an Aitchison's $\mathcal{A}$-distributed posterior (definition 2.28), that can be handled only with numerical methods.

The last step is selecting from this posterior distribution an estimate of $\mathbf{p}^*$, following the loss criteria already introduced. For a uniform prior distribution, for instance, the resulting posterior distribution is $\mathcal{D}(1+j_1, 1+j_2, \dots 1+j_D)$-distributed (by direct application of property 2.3). The posterior moments have no closed expression from the point of view of the geometry of the Simplex, which does not allow us to use the logistic loss criterion from a practical point of view. Moreover, the median is undefined in multidimensional problems, so the absolute value loss criterion is also useless. The criterion of the maximum yields again the same prediction as indicator kriging in its classical form (property 2.4). The remaining possibility is taking the arithmetic mean, which yields an estimator

$$\mathbf{p}_E^* = \mathcal{C}(\mathbf{1}+\mathbf{j}) \rightarrow p_i^* = \begin{cases} \frac{2}{D+1}, & j_i(\vec{x}) = 1 \\ \frac{1}{D+1}, & j_i(\vec{x}) = 0 \end{cases},$$

and, according to 2.4, coincides with the criterion of the maximum under an Aitchison geometry of the Simplex.

Note that the multinomial likelihood assumed for $Z(\vec{x})$ as a categorical variable yields the same estimated probability for all non-observed categories, independently of any order or distance relationship between them. This is always the case, for all the loss criteria and for any prior distribution which does not favor any category. In other words, this Bayesian approach yields in the end a generalized indicator function (7.14) with an a value dependent on the prior and the loss function.

Chapter 8

Conclusions

8.1 Discussion of case studies

8.1.1 Water pollution

Series characterization

From the Gualba series, we have studied five parameters: water conductivity, pH, ammonium content, ammonia content, and water temperature, either directly or transformed to $p\mathrm{K}_a$, the potential acidity constant governing the equilibrium between ammonium and ammonia. For each of these parameters, a sample space was chosen, an Euclidean space structure was built on them, and coordinates were taken with respect to arbitrary bases of these spaces. This information is summarized in table 8.1.

Table 8.1: Summary of geometries considered for each parameter of interest in the Gualba station.

chapter	parameter	variable	space	coordinate
3	conductivity	$\mathbf{C} = (C)$	$\mathbb{R}_+ \subset \mathbb{R}$	C
3	water temperature	$\mathbf{T} = (T)$	$\mathbb{R}$	T
5	hydrogen ion	$\mathbf{Z} = [\mathrm{H_3O^+}]$	$\mathbb{R}_+$	$-\log_{10}[\mathrm{H_3O^+}] = p\mathrm{H}$
5	ammonium	$\mathbf{Z} = \left[\mathrm{NH_4^+}\right]$	$\mathbb{R}_+$	$-\log_{10}\left[\mathrm{NH_4^+}\right] = p\mathrm{NH_4}$
5	ammonia	$\mathbf{Z} = [\mathrm{NH_3}]$	$\mathbb{R}_+$	$-\log_{10}[\mathrm{NH_3}] = p\mathrm{NH_3}$
5	acidity constant	$\mathbf{Z} = [\mathrm{K}_a]$	$\mathbb{R}_+$	$-\log_{10}\mathrm{K}_a = p\mathrm{K}_a$

The coordinates were then analyzed by taking into account their link to solar radiation, which is known to change periodically in time. Since direct information about this solar radiation is not available, we took water temperature as a proxy for it, and analyzed its decomposition in periodic waves, using Fourier analysis. Its energy spectrum (figure 3.6) gave some important waves, from which we highlight those with periods of 1, 2.5, 10 and 25 days, and 1 year. Regression techniques were then applied to compute the amplitudes and phases of each of these waves for all the variables of interest, independently for each month of July 2002 and July 2003. The coefficients can be found

in table 3.2 (for conductivity) and in table 5.1 (for the ammonia system parameters). We point out that, accordingly with these regression results, maximal $p\mathrm{K}_a$ (minimal water temperature) occurs before 1 a.m. (approximately 0.03 days after midnight), pH is maximal between 2 and 3 p.m., ammonium reaches its minimum (maximum $p\mathrm{NH}_4$) between midnight and 2:30 a.m., and conductivity is maximal between 2:30 and 4:30 p.m., depending on the year and the method of estimation. Such fluctuations are clearly displayed in figures 3.5 and 5.3, portraying the data sets, and figures 3.11 and 5.7, containing the final predictions.

Residuals obtained from regression (jointly considering both July 2002 and 2003) were regarded as RFs, characterized by a zero mean and covariance structures dominated by hole effects (see figures 3.9 and 5.4-5.6). From these covariance functions, it is interesting to note a $\sim 80\%$ positive cross-correlation between pH and ammonium (figure 5.4).

It is worth recalling that, in a strict sense, if the regression residuals are self-correlated, classical regression is not applicable, and we should use instead kriging of the mean; however, since we found the differences between both techniques to be small in the case of conductivity (table 3.2), we consider for the rest of variables the simpler regression as a good approximation to the theoretically-better kriging of the mean.

Hazard and water quality

The characterization of the series explained in the last section allowed us to obtain an estimation of the mean and the variance of the values of all those variables at all full hours, during the months of July 2002 and 2003. Explicitly assuming the random functions to be Gaussian on their respective sample spaces (accordingly with table 8.1), we used these values to compute the probability of exceeding each of the hazardous thresholds shown in table 1.1 and defining water quality categories. The hazard of conductivity above 1000 $\mu S/cm$ is displayed in figure 3.12, whereas figure 5.8 shows the hazard associated to pH, ammonia and ammonium. These two figures display a very different picture. In the case of ammonia system, probability of exceeding each threshold presents sharp clear-cut pulses from almost zero to almost one, indicating that the thresholds are almost-daily exceeded. In contrast, the probability of conductivity above 1000 $\mu S/cm$ is more constantly below or above the threshold, and the only periodicity to be seen has an approximate 2.5-day period. However, we regard this last periodic picture as an artifact derived from the sampling strategy and the properties of kriging variance, since this periodicity is not so strongly seen in conductivity series.

If we do not want to assume a Gaussian model for the random function, we have also the option of using indicator techniques to directly estimate probabilities. We tried to approximate the distribution of conductivity through a discretization of its domain by deciles, and use it to compute the hazard of being above 1000 $\mu S/cm$, and also to directly estimate the distribution of the auxiliary variable "*being above that threshold*" (or "*having a water quality index of 4 due to high conductivity*"). Both sets of variograms (those used in the decile case and that used in the single-threshold case)

presented 1-day hole effects and relatively short-range structures. It is not surprising then that the estimated probabilities of being above and below the threshold (see figures 7.10, 7.12 and 7.13) present always 1-day fluctuations, which were not displayed by the Gaussian-related technique of the first paragraph, dominated by the 2.5-period explained there.

Using any of these techniques, it is evident that the probability of exceeding each threshold varies along time, mainly with a daily basis. The presence of these strong variations implies that a single measurement of any chemical parameter may be meaningless. We have seen in the previous section that this effect should be particularly born in mind with reactive pollutants, since it might be possible to reach highly toxic moments passing unnoticed: *e.g.*we could measure a low ammonia concentration at night, but the day after produce pollution conditions. Then online control stations are far more important in these situations, since they allow us to continuously control pollution parameters, but also to prematurely estimate the probability of hazardous future situations.

8.1.2 Air pollution

The system $\{\mathrm{Fe}, \mathrm{Pb}, \mathrm{Hg}\}$ of moss sample data set obtained from the ukrainian carpathian range was regarded as compositional, since its components are measures of the relative amount of each one of these elements in the whole. Furthermore, knowledge of the pollution processes dominant in the area suggested to interpret the relative amount of Fe as a measure of the influence of corrosion pollution, the relative Pb amount as a proxy for combustion, and Hg as a relative indicator of regional pollution by industrial emissions. The combined amount of these three pollutants could be a measure of total pollution impact. However, we decided to disregard it, and focus only on relative influences of these three processes, because this total amount can also be conceptually related to the time of exposure, thus to the plant age, which is not known.

The sample space of the data is then a 3-part Simplex. We chose a basis which contains two elements: the first balances Fe against the other two pollutants, and the second balances Pb against Hg. The first coordinate is then an indicator of the influence of big particles against gases, whereas the second coordinate indicates whether the gases come mainly from combustion local sources or from regional industrial emissions. Unsurprisingly, the variogram of the first coordinate has a clearly shorter range than the variogram of the second one (figure 6.2).

Prediction of these two coordinates across the studied region (figures 6.3 and 6.4) shows that regional pollution is most important on the northern face of the mountains (comparing the pollution maps with figure 1.4), corrosion is predominant near the cities of Drogóbich and Chernovtsí, whereas combustion dominates the southwestern basin east of Uzhgorod.

Results of kriging could also be used in this case to compute probabilities of presenting certain hazardous amounts of each of these metals, if these hazardous levels were known. We would follow the same approach as in the water pollution case, but

using bi-variate normal distributions in the Simplex. However, the fact that here we were only interested in *relative* changes reduces the interest of this computation: no environmental quality index could be obtained from this problem, but only a balance between different sources of pollution.

8.2 Discussion of methods

The concepts of *sample space* and *scale* of a data set are well-known in statistics. However, its practical importance in applied data analysis has passed mainly unnoticed, in particular in the environmental and geological sciences. During the sixties, a concern in the geological community arose about these concepts regarding compositional data. Its sample space, the Simplex, could be given a *meaningful* (=compatible with the desired scale) Euclidean space structure. This led to the introduction of algebraic concepts (basis, coordinates, projections) *before* statistics were used, and *after* results were obtained.

These ideas were summarized by Pawlowsky-Glahn [2003] in the operative *principle of working on coordinates*: as a zero step of any statistical analysis, we must identify the sample space and a scale for our data, and when this space has a meaningful Euclidean structure, take coordinates with respect to any of its orthonormal basis; statistics may be applied then to the coefficients with respect to that basis, and the obtained results might be applied again to the basis to recover objects from the original sample space.

Geostatistical techniques represent no exception to this principle. The very concept of random function and its essential properties (mean value, covariance structure, stationarity) are well-defined on the coordinates with respect to any basis of an Euclidean space. Furthermore, we have shown that these essential properties may be understood as object or operations in the Euclidean space itself. In this way, the mean value is a vector minimizing the spread around it, the covariance structure is an endomorphism describing this spread, etc.

The case of covariance/variogram functions is particularly important. Covariance structures defined on the coordinates with respect to two different basis of the same space present the same properties of symmetry, or validity by Fourier transforms, and their global range of independence are equal. These are intrinsic properties of the covariance structures (seen as endomorphisms), and not artifacts of the basis.

Estimation and prediction (kriging) present also no problem of definition: either using coordinates (and applying the results to the basis in use) or defining kriging as a linear transformation, the final predictions are exactly the same, as are confidence regions built around them. From a theoretical point of view, *kriging in an Euclidean space* is thus independent of the chosen basis. Among the interesting optimal properties, this kriging predictor minimizes the expected distance between the prediction and the true value.

Even the strongest geostatistical result, the conditional expectation character of simple kriging in a Gaussian random function, is reproduced for vector-valued random

functions. The Gaussian assumption allows then to go a step beyond the simple estimation: it makes *normal kriging in an Euclidean space* yield the probability distribution on the Euclidean space.

These considerations have shown their capital importance in two common sample spaces: the positive real line (and its multivariate version, the positive orthant of the real space), and the Simplex. These spaces can be given an Euclidean space structure based on simple operations (product and powering), and respecting a meaningful log-relative metric.

The case of a positive variable is better known, since it usually appears in applications. Ignoring its sample space structure gives the classical estimator—known as lognormal kriging—a set of well-known problems: unclear character of conditional expectation estimator when simple kriging is not applicable, non-optimal confidence intervals, local bias in block kriging, and generally bad properties of local conditional expectation estimation. Taking into account the sample space structure, we obtain *kriging in the positive real line*, which has none of these problems. Furthermore, change-of-support models get clear and valid definitions in such a framework.

The generalization of these results to compositional data (considering the Simplex as an Euclidean space) is straightforward, and *kriging in the Simplex* is then the best unbiased linear predictor of a compositional vector in the geometry of the Simplex. It yields valid compositions as results, positive and closed, something which is not always true for other existing interpolation techniques applied to compositions.

It is well established that there is a link between the structure of the Simplex and classical operations for multinomial probability vectors, like Bayesian updating or information measures. Based on this link, we apply *kriging in the Simplex* to estimate multinomial probability vectors—or finite discrete probability distributions of variables in any set, not necessarily an Euclidean space—. In the case that we truly observed them, *kriging in the Simplex applied to multinomial probability vectors* keeps all its properties of unbiasedness, minimal variance and conditional expectation character, and always yields valid probability vectors, positive and summing up to one.

But these probability vectors are in practice seldom observed. Instead, one observes a sample of the regionalized variable from which the conditional probability distribution is wanted at any location. The classical approach here—indicator kriging—is the application of indicator functions defined at some cutoffs and the geostatistical treatment of the indicator-transformed data set. This method often yields impossible probability estimates.

To deal with this framework, we devise a two-step stochastic model, where at each sampled location we observe a single-trial realization of an independent multinomial variable, which probability vector forms a random function. Using this formalism, the multinomial probability vectors may be estimated in a Bayesian framework at each sampled location updating a non-informative prior by the indicator-transformed data set. Afterwards, these bayesian estimations are introduced in the geostatistical techniques to interpolate them. The resulting *kriging in the Simplex applied to generalized indicators* also yields admissible probabilities, positive and summing up to one.

The preliminary estimation procedure does not change the shape or the range of the covariance structure of the estimated multinomial vectors, and it only scales this structure. This implies that existing indicator variography software is perfectly useful. Moreover, it is proven that such a scaling has no effect in the kriging procedure.

We put also forward a refined method, which essentially tries the joint updating of a general prior covering the whole domain by all the observations. This prior is derived from the same hypothesis underlying *kriging in the Simplex applied to generalized indicators*. Results do not differ significantly between them, although this method is able to modify the estimated probability distribution at a sampled location using information coming from nearby samples. On the side of the disadvantages, this joint Bayesian method needs extensive computation.

In all these non-parametric cases, uncertainty of results is far higher than those obtained with Gaussian assumptions. Not too much should be expected of model-free techniques in geostatistical applications. When a single sample is available—the case of geostatistics—models become a necessary complement. The last Bayesian method assumed a model for the log-ratios of probabilities, yet trying to let the model for the original variable be the most general possible.

8.3 Future work

This work covers a relatively wide range of statistical techniques, theoretical issues and practical cases, which have not been totally explored. This is not only a matter of time, but of coherence. Here we used basic concepts of Algebra, Probability and Measure Theory, jointly with fundamental Geostatistics, to show how a general technique of statistical analysis, the principle of working on coordinates, applies to regionalized variables and, doing so, it solves many of their classical inference problems. Some environmental case studies have been used to illustrate these concepts. We left for further work avenues which diverge too far from this picture.

- Selectivity curves in the context of positive variables are briefly explained here, and they still lack practical studies to assess their usefulness beyond its theoretical correctness. The same can be said of change-of-support models: although we presented here a theoretical justification based on Cartier's relation, it would be interesting to apply it to real and simulated cases.

- Compositional data sets have a particular characteristic: their coordinates are not one-to-one-related to their components. This clearly hinders the interpretation of results. An interesting preprocessing would be the selection of an optimal basis, not necessarily an orthonormal one, which keeps to a minimum the influence of every component and, if possible, minimizes the cross-covariances between coordinates. Bi-plot representations of the covariance at some selected lag distances might be useful.

- One of the main theoretical problems of indicator kriging is the need of indicator covariance systems which yield valid probability estimates. Unfortunately, the conditions defining this validity are not fully known. kriging in the Simplex applied to multinomial probability vectors is independent of this limitation, because it yields valid probabilities in any case. However, it is unclear whether this independence is transferred to kriging in the Simplex applied to generalized indicators. It would be interesting to conduct simulation or theoretical studies of the properties of covariance systems computed from known probability distributions.

- The bayesian technique applied to indicators was here presented as a way to integrate into the estimated probability distribution of a sampled location information from nearby samples. A full characterization of this technique still lacks the exploration of, at least, three different avenues.

  - The posterior is fully characterized, but obtaining its maximum implies the solution of a huge multivariate non-linear system of equations. Preliminary studies suggest that the size of the system might be more problematic than the non-linear part.
  - Instead of the maximum of the posterior, its mean can also be computed—using Monte Carlo estimation methods—. The simulation of large random vectors is known to be problematic, because it is difficult to adequately sample the full multivariate distribution. It would be interesting then to study the marginal properties of the N-block Aitchison distribution, and try to draw first an optimal estimate for the sampled locations, and use it afterwards for the unsampled ones. This makes sense, because the unsampled locations do not carry any information, and they should not affect the estimation on the sampled ones.
  - Assuming a hierarchical model is another way to drastically reduce the dimension of the random vector to sample in the Monte Carlo method. Following Tjelmeland and Lund [2003], we may assume the covariance function between the unknown probability vectors to be of a particular type (*e.g.*spherical) with a random range and a random covariance matrix at the origin. Models for these parameters might be respectively an exponential distribution and a Wishart distribution. Then MCMC (Markov Chain Monte Carlo) methods should be applied to estimate the posterior distribution of all these parameters.

- We have briefly mentioned a generalization of indicator kriging to continuous distributions, disjunctive kriging. This technique yields as predictor of the distribution a linear combination of univariate functions of the data set which best approximates the true conditional probability in an L^2 sense (page 7.1.5). But we also said that Egozcue and Díaz-Barrero [2003] and van den Boogaart [2004] suggest that this distance is flawed to deal with probability distributions, and

they introduce alternative Hilbert space structures to the classical L^2 space of functions. The development of a disjunctive kriging-like technique based in these alternative Hilbert spaces is a powerful idea.

- Finally, the Gualba data set is only superficially explored. Using the whole two-year series of all the available chemical parameters implies no theoretical problem, but practical ones: the size of the sample, more than 70000 measurements, and the fact that most of them are not simultaneous, make the treatment of this data set a problem in itself.

Chapter 9

Notation summary

Algebra and geometry		page
$\mathbb{E}$,$\mathbb{A}$	generic vector of Euclidean space, set of elements	15
$\mathbb{F}$	generic vector subspace	15
$\mathbb{R}^D$	D-dimensional real space	26
$\mathbb{R}^D_+$	D-dimensional real positive space	29
$\mathbb{S}^D$	D-part Simplex	33
$\oplus$	Abelian group operation, inner sum	14
$\ominus$	inverse inner sum	
$\odot$	external product	
$\langle \cdot, \cdot \rangle_{\mathbb{E}}$	scalar product	15
$\| \cdot \|_{\mathbb{E}}$	norm	16
$d(\cdot, \cdot)$	distance	16
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{g}$	vectors (boldface lowercase Latin character)	14
$\mathbf{n}$	neutral element vector	
$\mathbf{e}_i$	i-th basis element	
λ, μ	scalars	14
$\mathbf{E}, \mathbf{F}$	vector sets (boldface uppercase Latin character)	
$\mathbf{E}$	generating system	15
$\mathbf{E}$	basis	15
$\mathbf{E}$	orthogonal basis	16
$\mathbf{E}$	orthonormal basis	17
$\underline{\alpha}, \underline{\beta}, \underline{\gamma}$	vectors of coordinates (underlined lowercase Greek character)	15
$\underline{\underline{T}}, \underline{\underline{\varphi}}$	matrices of scalar values (double-underlined character)	81
$\underline{\underline{\varphi}}$	matrix of change of basis	
$\underline{\underline{I}}_D$	identity matrix in a D-dimensional space	
$\underline{\underline{1}}_D$	$D \times D$ one matrix	122
T, Σ,C,γ	operators (endomorphisms)	81
$\Sigma\mathbf{x}$	operator acting on a vector $\mathbf{x}$	81

Geometry of the Simplex		page
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{g}$	compositions (boldface lowercase Latin character)	33
$\oplus$	perturbation	35
$\odot$	power operation	35
$\langle \cdot, \cdot \rangle_A$	compositional scalar product	-
$\| \cdot \|_A$	compositional norm	-
$d_A(\cdot, \cdot)$	compositional distance	35
$\mathcal{C}(\cdot)$	closure operation	35
$\mathrm{alr}(\cdot)$	additive log-ratio transform	120
$\mathrm{clr}(\cdot)$	centered log-ratio transform	120
$\mathrm{ilr}(\cdot)$	isometric log-ratio transform	122
φ	rectangular matrix relating clr and ilr transforms	
$\mathrm{agl}(\cdot)$	inverse additive log-ratio transform	121
Measure and probability		
$\lambda_{\mathbb{E}}$	Lebesgue measure in the space $\mathbb{E}$	19
$\lambda_{\mathbb{R}}$	classical real Lebesgue measure	19
$P(\cdot)$	probability measure	20
$F(\cdot)$	probability law	20
$f(\cdot)$	density function	20
$\Pr[\cdot]$	probability of an event	20
Distribution models		
$\underline{\mu}$	position parameter vector	39
$\underline{\underline{\Sigma}}$	scale-dispersion parameter matrix	39
$\underline{\theta}$	alternative position parameter vector	37,39
$\underline{\underline{\phi}}$	interaction parameter matrix	39
$\mathcal{D}(\underline{\theta})$	Dirichlet distribution	37
$\mathcal{N}_{\mathbb{S}^D}\left(\underline{\mu}, \underline{\underline{\Sigma}}\right)$	normal distribution on the Simplex	39
$\mathcal{A}\left(\underline{\theta}, \underline{\underline{\phi}}\right)$	Aitchison distribution	40
Random variables		
Z	random variable (uppercase character)	-
z	outcome of a random variable (lowercase character)	-
$\mathbf{Z}$	random vector in a generic Euclidean space	20
$\mathbf{z}$	outcome of a random vector in a generic Euclidean space	
$\underline{\zeta}$	random vector in the real space	20
Moments and inference		
$\mathrm{E}_{\mathbb{E}}[\cdot]$	vector expectation in $\mathbb{E}$	20
$\mathrm{Var}_{\mathbb{E}}[\cdot]$	covariance matrix in $\mathbb{E}$ (of the components of a vector)	20
$\mathrm{Cov}_{\mathbb{E}}[\cdot, \cdot]$	covariance matrix in $\mathbb{E}$ (of two vectors)	20
$z_1, z_2, \ldots z_N$	sample of size N	24
$\bar{z}$	arithmetic mean of the sample	-
θ	generic parameter	-
$\hat{\theta}$	estimator of θ	-
$\tilde{\theta}$	estimation of θ	-

Random functions (RF)		page
$\mathbb{R}^p$	physical space	47
$\mathcal{D} \subset \mathbb{R}^p$	domain of the RF	
$\vec{x} \in \mathcal{D}$	physical location	
$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N$	location set	-
$\vec{x}_0$	predicted location	58
$n, m \in (0,)1, 2, \ldots, N$	location index	-
$\vec{h}$	lag distance (difference vector)	48
h	lag distance (difference vector norm)	-
$\mathbb{E}$	sample space (space image, support)	81
$\mathbf{Z}(\vec{x})$	vector RF	48
$\{\mathbf{z}(\vec{x}_n)\}$	observed regionalized sample	
$\zeta_i(\vec{x})$	i-th coordinate RF of $\mathbf{Z}(\vec{x})$	
$i, j \in 1, \ldots, D$	coordinate index	-
$\underline{\underline{C}}_{nm} = \underline{\underline{C}}(\vec{x}_n, \vec{x}_m)$	(coordinate) covariance between locations	49
$\underline{\underline{\gamma}}_{nm} = \underline{\underline{\gamma}}(\vec{x}_n, \vec{x}_m)$	(coordinate) variogram between locations	49
$C_{ij}(\vec{h})$	covariance between coordinates	50
$K_{ij}(\vec{h})$	non-centered covariance between coordinates	135
$\hat{C}_{ij}(\vec{h})$	experimental covariance	
$f_a(\vec{x})$	drift function	58
$a \in 1, 2, \ldots, A$	drift function index	
ν_a	Lagrange multiplier of the a-th drift function	
$\mathbf{z}_0^*$	kriging predictor	59
$\vec{x}_0$	predicted location	58
λ_{ni}	kriging weight	59
σ^2_{XK}	X-kriging variance (simple, universal, drift)	
$\underline{\underline{\Lambda}}$	kriging weight matrix	85
σ_{ij}	kriging covariance between coordinates	
$v, V, W \subset \mathcal{D}$	physical blocks	
$Z_v(\vec{x})$	sampling function	66
$Z_v(\vec{x})$	regularized RF	
$C_v(\vec{h})$	regularized covariance	
$\sigma(v\|V)$	dispersion variance	67
$\phi(\cdot)$	transformation (to Gaussian marginals)	70
$T(\cdot), Q(\cdot), \tilde{m}(\cdot), B(\cdot)$	selectivity functions	102
$I_i(\cdot)$	i-th cutoff (cumulative) indicator transform	134
$J_i(\cdot)$	i-th category (disjunctive) indicator transform	134
$\mathbf{I}, \mathbf{J}$	full indicator RFs	
$\mathbf{i}(\vec{x}_n), \mathbf{j}(\vec{x}_n)$	fully-observed indicator vector	
$\mathbf{P}$	multinomial probability vector RF	140
$\mathbf{p}, \mathbf{q}$	multinomial probability vectors	-

Note: references with blank page number are related to the preceding one.

Bibliography

Abramovitz, M. and I. Stegun (1965). *Handbook of mathematical functions*. Dover, New Yowk.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44(2): 139–177.

Aitchison, J. (1984). Reducing the dimensionality of compositional data sets. *Mathematical Geology*, 16(6):617–636.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). ISBN 0-412-28060-4. 416 p.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In Vera Pawlowsky-Glahn, editor, *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology*, volume I, II and addendum, pages 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p. ISBN ISBN: 84-87867-97-9.

Aitchison, J. and J. A. C. Brown (1957). *The Lognormal Distribution*. Cambridge University Press, Cambridge (UK). 176 p.

Aitchison, J. and C. Barceló-Vidal, J.J. Egozcue, and V. Pawlowsky-Glahn. A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In Ulf Bayer, Heinz Burger, and Wolfdietrich Skala, editors, *Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology*, volume I and II, pages 387–392. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Armstrong, M. and G. Matheron (1986a). Disjunctive kriging revisited: part i. *Mathematical Geology*, 18(8):711–728.

Armstrong, M. and G. Matheron (1986b). Disjunctive kriging revisited: part ii. *Mathematical Geology*, 18(8):729–742.

Berberian, S.K. (1961) *Introduction to Hilbert Space*. University Press, New York. Translation: 1971, Ed. Teide, Barcelona, Spain.

Besag, J., L. York, and A. Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43:1–59.

Billheimer, D., P. Guttorp, and W.F. Fagan. (2001) Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456):1205–1214.

Bogaert, P. (2002). Spatial prediction of categorical variables: the bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment*, 16: 425–448. doi: 10.1007/S00477-002-0114-4.

Bogaert, P. (1999). On the optimal estimation of the cumulative distribution function in presence of spatial dependence. *Mathematical Geology*, 3(2):213–239.

v.d. Boogaart, K.G. (2004). personal communication. boogaart@uni-greifswald.de.

v.d. Boogaart, K.G. (2004). Statistics structured by the aitchison space. internal report, December 2004.

v.d. Boogaart, K. G. and A. Brenning (2001). Why is universal kriging better than IRFk-kriging: estimation of variograms in the presence of trend. In Ross, G., Ed., *Proceedings of IAMG'01 — The seventh annual conference of the International Association for Mathematical Geology*, CD-ROM. Cancún (Mexico)

Carle, C.F. and G.E. Fogg (1996). Transition probability-based indicator geostatistics. *Mathematical Geology*, 28(4):453–476.

Carr, J.R. (1994). Order relation correction experiments for probability kriging. *Mathematical Geology*, 26(5):605–621.

Carr J.R. and N.H. Mao (1993). A general-form of probability kriging for estimation of the indicator and uniform transforms. *Mathematical Geology*, 25(4):425–438.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research*, 65(12):4185–4193.

Chayes, F. (1971). *Ratio Correlation*. University of Chicago Press, Chicago, IL (USA). 99 p.

Chilès, J.P. and P. Delfiner (1999). *Geostatistics — modeling spatial uncertainty*. Series in Probability and Statistics. John Wiley and Sons, Inc., New York, NY (USA). ISBN 0-471-08315-1. 695 p.

Christakos, G. (2000). *Modern spatio-temporal geostatistics*. Number 6 in Studies on Mathematical Geology. Oxford University Press, New York.

Christakos, G. (1990). A bayesian/maximum entropy view to the spatial estimation problem. *Mathematical Geology*, 22(7):763–777.

Clark, I. (1979). *Practical Geostatistics*. Applied Science Publishers, London (UK). 129 p.

Clark, I. and W.V. Harper (2000). *Practical Geostatistics 2000*. Ecosse North America Llc, Columbus Ohio (USA). 342 p.

Clifford, P. (1998) Discussion of model-based geostatistics. In Diggle et al. [1998], *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, pages 299–350.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley and Sons, New York, NY (USA). 900 p.

Daunis-i-Estadella, J., J.J Egozcue and V. Pawlowsky-Glahn (2002). Least squares regression in the Simplex. In Ulf Bayer, Heinz Burger, and Wolfdietrich Skala, editors, *Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology*, volume I and II, pages 411–416. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

David, M. (1977). *Geostatistical Ore Reserve Estimation*, volume 2 of *Series on Developments in Geomathematics*. Elsevier, New York, NY (USA). 364 p.

Deutsch C. and A. Journel (1992). *GSLIB - Geostatistical Software Library and User's Guide*. Oxford University Press, New York, NY (USA). 340 p. and 2 diskettes.

Diggle, P.J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 47(3):299–350.

Doob, J.L.(1992). *Stochastic Processes*. Wiley, New York, NY (USA). (reprinted, 1990).

Dowd, P.A. (1982). Lognormal kriging–the general case. *Mathematical Geology*, 14(5): 475–499.

Eaton, M. L. (1983) *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

Egozcue, J.J. (2002). La información es una composición (*Information is a composition*). Seminar reports of the thematic network on CoDa, Dept. Informàtica i Matemàtica Aplicada- Universitat de Girona.

Egozcue, J.J and J.L. Díaz–Barrero (2003) Hilbert space on probability density functions with aitchison geometry. In Thió-Henestrosa and Martín-Fernández [2003].

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300. ISSN 0882-8121.

Egozcue, J.J., J.L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica*, 22(1):(in press).

Emery, X. (2004). On the consistency of the indirect lognormal correction. *Stoch Envir Res and Risk Ass*, 18:258–264.

von Eynatten,H., V. Pawlowsky-Glahn, and J.J. Egozcue (2002). Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Mathematical Geology*, 34(3):249–257. ISSN 0882-8121.

Fahrmeir, L and A. Hamerle, editors (1984). *Multivariate Statistische Verfahren*. Walter de Gruyter, Berlin (D), 796 p.

Goovaerts, P. (1994). Comparative performance of indicator algorithms for modelling conditional probability distribution functions. *Mathematical Geology*, 26:389–411.

Goovaerts, P., R. Webster, and J.P. Dubois (1997). Assessing the risk of soil contamination in the swiss jura using indicator geostatistics. *Environ. Ecol. Statist.*, 2: 331–344.

Haas, A. and Ph. Formery (2002). Uncertainties in facies proportion estimation i. theoretical framework: the Dirichlet distribution. *Mathematical Geology*, 34(6):679–702.

Handcock, M.S. and M.I. Stein (1993). A bayesian analysis of kriging. *Technometrics*, 35:403–410.

Hewitt, K. (1997). *Regions of risk : a geographical introduction to disasters*. Longman, Essex (UK). 389 p.

Idescat (2005). Institut d'estadística de catalunya. official web page, 2005. URL `http://www.idescat.net`.

Isaaks, E.H. and R. M. Srivastava (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York, NY (USA). 561 p.

Journel, A.G. and D. Posa (1990). Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, 22(8):1011–1025.

Journel, A.G. (1980). The lognormal approach to predicting local distributions of selective mining unit grades. *Mathematical Geology*, 12(4):285–303.

Journel, A.G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology*, 15(3):445–468.

Journel, A.G. and C.J. Huijbregts (1978). *Mining Geostatistics*. Academic Press, London (UK). 600 p.

Juang, K.W. , D.Y. Lee, and C.K. Hhsiao (1998). Kriging with cumulative distribution function of order statistics for delineation of heavy-metal contaminated soils. *Soil Science*, 163(10):797–804.

Kullback, S. (1997). *Information Theory and Statistics, an unabridged republication of the Dover 1968 edition*. Dover publications, Minetola.

Lantuéjoul, C. (2002). *Geostatistical simulation: models and algorithms*. Springer.

Lasky, S.G. (1950) How tonnage and grade relations help predict ore reserves. *Engineering and Mining Journal*, 151(4):81–85.

Le, N.D. and J.V. Zidek (1992). Interpolation with uncertain spatial covariances: a bayesian alternative to kriging. *Journal of Multivariate Analysis*, 43:351–374.

Leonard, T. and J.S.J. Hsu (1999). *Bayesian Methods: an analysis for statisticians and interdisciplinary researchers*. Series in statistical and probabilistic methods. Cambridge University Press. ISBN 0-521-59417-0.

Mapfre (2002). *Manual de contaminación ambiental*. MAPFRE, Madrid (Spain), 2nd edition.

Marcotte, D. and P Groleau (1997). A simple and robust lognormal estimator. *Mathematical Geology*, pages 993–1009.

Martí-Roca, E. (2004). personal communication. eugenia@ceab.csic.es

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J.A. Martín-Fernández (2002). Normal in $\Re^+$ vs lognormal in $\Re$. In Ulf Bayer, Heinz Burger, and Wolfdietrich Skala, editors, *Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology*, volume I and II, pages 305–310. Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2003). Distributions on the simplex. In Thió-Henestrosa and Martín-Fernández [2003].

Matheron, G. (1976). A simple substitute for the conditional expectation: the disjunctive kriging. In Massimo Guarascio, Michel David, and C. Huijbregts, editors, *Advanced Geostatistics in the Mining Industry*, volume 24 of *NATO Advances Study Institute Series; Series C: Mathematical and Physical Sciences,* pages 221–236. D. Reidel Publishing Company, Dordrecht (NL), 461 p.

Matheron, G. (1965) *Les variables régionalisées et leur estimation—une application de la théorie des fonctions aléatoires aux sciences de la nature*. Masson et Cie., Paris (F). 305 p.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of the Royal Society of London*, 29:367–376.

Myers, D.E. (1982). Matrix formulation of co-kriging. *Mathematical Geology*, 14(3): 249–257.

Nelder, J.A. and R.W.M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, series A*, 135:370–384.

Nielsen, O.A. (1997). *An Introduction to Integration and Measure Theory*. Canadian Mathematical Society series of monographs and advanced texts. Wiley.

Omre, H. (1987). Bayesian kriging–merging observations and qualified guesses in kriging. *Mathematical Geology*, 19(1):25–39.

Pardo-Igúzquiza, E. and P.A. Dowd (2005). Multiple indicator cokriging with application to optimal sampling for environmental monitoring. *Computers and Geosciences*, 31(1):1–13.

Pawlowsky, V. (1986). *Räumliche Strukturanalyse und Schätzung ortsabhängiger Kompositionen mit Anwendungsbeispielen aus der Geologie*. PhD thesis, Fachbereich Geowissenschaften, Freie Universität Berlin, Berlin (D). 170 p.

Pawlowsky, V. R.A. Olea, and J.C. Davis (1994). Additive logratio estimation of regionalized compositional data: an application to calculation of oil reserves. In Roussos Dimitrakopoulos, editor, *Geostatistics for the Next Century*, volume 6 of *Series on Quantitative Geology and Geostatistics*, pages 371–382. Kluwer Academic Publishers, Dordrecht (NL), 497 p. ISBN 0-7923-2650-4.

Pawlowsky, V. R.A. Olea, and C. Barceló (1996). Estimation of regionalized compositions using different backtransformations. In International Geological Congress IGC, editor, *Documents of the 30th IGC*, page 4808. Geological Publishing House, Beijing (PRC), CD-ROM. ISBN 7–900001–00–X.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In Thió-Henestrosa and Martín-Fernández [2003].

Pawlowsky-Glahn (1984). On spurious spatial covariance between variables of constant sum. *Science de la Terre, Sér. Informatique*, 21:107–113. ISSN 0335-9255.

Pawlowsky-Glahn, V. and J.J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15(5):384–398.

Pawlowsky-Glahn, V. and R.A. Olea (2004). *Geostatistical Analysis of Compositional Data*. Number 7 in Studies in Mathematical Geology. Oxford University Press. ISBN 0-19-517166-7.

Pebesma, E.J. and C.G. Wesseling (1998). Gstat: A program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, 24(1):17–31.

Poch, M. (1999) *Les qualitats de l'aigua*. Departament de Medi Ambient, Generalitat de Catalunya.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

Rendu, J.M. (1979). Normal and lognormal estimation. *Mathematical Geology*, 11(4): 407–422.

Rényi, A.(1997). *Cálculo de probabilidades*. Editorial Reverté, 641 p.

Rojo, J. (1986). *Álgebra lineal*. AC, 2 edition.

Roth, C. (1998). Is lognormal kriging suitable for local estimation? *Mathematical Geology*, 30(8):999–1009.

Samper-Calvete, F.J. and J. Carrera-Ramírez. (1990) *Geoestadística – Aplicaciones a la hidrología subterránea (Geostatistics–Applications to hydrogeology)*. Centro Internacional de Métodos Numéricos en Ingeniería (CIMNE), Barcelona (E). ISBN 84-404-6045-7. 484 p.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656.

Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. *Third Sym. Mathem. Statist. Pretoria South Africa. Laubscher, N.F. (Ed)*, pages 51–97.

Sullivan, J. (1984). Conditional recovery estimation through probability kriging—theory and practice. In *Geostatistics for Natural Resources Characterization*, 2nd NATO-ASI. Stanford, CA (USA), 2 Vols., 1092 p.

Suro-Perez, V. and A. Journel (1991). Indicator principal component kriging. *Mathematical Geology*, 23(5):759–788.

Thió-Henestrosa, S. and J.A. Martín-Fernández, editors (2003). *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings*. Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Tjelmeland, H. and K.V. Lund (2003). Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, 30(1):87–100.

Tolosana-Delgado, R. (2004). Daily variations in stream water chemistry and their implications for ammonia generation. Master thesis, Institut de Medi Ambient - Universitat de Girona.

Tolosana-Delgado, R. and V. Pawlowsky-Glahn (2003). A new approach to kriging of positive variables. In John Cubitt, editor, *Proceedings of IAMG'03 — The ninth annual conference of the International Association for Mathematical Geology*. University of Portsmouth, Portsmouth (UK).

Tyutyunnik, Y.G. (2004). personal communication. carme@univ.kiev.ua

Vargas-Guzman, J.A. and R. Dimitrakopoulos (2003). Successive nonparametric estimation of conditional distributions. *Mathematical Geology*, 35(1):39–52.

Verly, G. (1983). The multigaussian approach and its applications to the estimation of local reserves. *Journal Of The International Association For Mathematical Geology*, 15(2):259–286.

Wackernagel, H. (1998) *Multivariate Geostatistics, An Introduction With Applications (2nd edition)*. Springer Verlag, Berlin (D). ISBN 3-540-64721-X. 291 p.

Walwoort D.J.J. and J.J. de Gruijter (2001). Compositional kriging: a spatial interpolation method for compositional data. *Mathematical Geology*, 33(8):951–966.

Wismer, D.A. and R. Chattergy (1978). Introduction to nonlinear optimization. A problem solving approach. Elsevier, Amsterdam (Netherlands). 395 p. (cited in Walwoort and de Gruijter, 2001)

Yao, T. and A.G. Journel (1998). Automatic modeling of (cross) covariance tables using Fast Fourier Transform. *Mathematical Geology*, 30(6):589–615.