58

Towards a Soil Information System with quantified accuracy

Three approaches for stochastic simulation of soil maps

D.J. Brus
G.B.M. Heuvelink

rapporten

WOt

Wettelijke Onderzoekstaken Natuur & Milieu

WAGENINGEN UR
For quality of life

Towards a Soil Information System with quantified accuracy

This report was produced in accordance with the Quality Manual of the Statutory Research Tasks Unit for Nature & the Environment.

The ‘WOt Reports’ series presents the findings of research projects implemented for the Statutory Research Tasks Unit for Nature & the Environment by various centres of expertise.

Towards a Soil Information System with quantified accuracy

Three approaches for stochastic simulation of soil maps

D.J. Brus

G.B.M. Heuvelink

Rapport 58

Wettelijke Onderzoekstaken Natuur & Milieu

Wageningen, december 2007

Abstract
Brus, D.J. & G.B.M. Heuvelink, 2007. *Towards a Soil Information System with quantified accuracy; Three approaches for stochastic simulation of soil maps*. Wageningen, Statutory Research Tasks Unit for Nature and the Environment. WOt-rapport 58. 96 p. 15 Fig.; 1 Tab.; 51 Ref.

This report describes (geo)statistical methods for mapping soil type and soil properties from soil observations and explanatory information. These methods do not only produce a map, but also quantify the associated uncertainty in terms of probability distributions. Stochastic simulation is used to generate possible realities from the probability distribution of the soil property, which are useful to communicate uncertainty about the soil and can also be used in Monte Carlo uncertainty propagation studies. Three approaches for deriving the complex probability distributions are described: Kriging, Bayesian Maximum Entropy, and Markov random fields. All three methods apply to quantitative (continuous) and qualitative (categorical) soil properties. The theory is illustrated with simulation of soil properties used as input by the soil acidification model SMART2. It is concluded that Kriging methods are well-developed and have proven their use on numerous occasions, but the Bayesian Maximum Entropy and Markov random field approach are valuable alternatives that deserve attention,
particularly for statistical modelling of qualitative soil properties.

Key words: Soil map, Uncertainty, Kriging, Geostatistics, Bayesian Maximum Entropy, Markov random field

Referaat
Brus, D.J. & G.B.M. Heuvelink, 2007. *Naar een bodeminformatiesysteem met gekwantificeerde nauwkeurigheid: drie benaderingen voor het stochastisch simuleren van bodemkaarten*. Wageningen, Wettelijke Onderzoekstaken Natuur & Milieu, WOt-rapport 58. 96 blz. 15 fig.; 1 tab.; .51 ref.

In dit rapport worden (geo)statistische methoden beschreven voor het in kaart brengen van bodemtypen en bodemeigenschappen op basis van veldwaarnemingen en verklarende informatie. Deze methoden leveren niet alleen een kaart op, maar kwantificeren tevens de bijbehorende mate van onzekerheid in termen van kansverdelingen. Met behulp van stochastische simulatie worden ‘mogelijke werkelijkheden’ gegenereerd uit de kansverdeling van de bodemeigenschap, waarmee de mate van onzekerheid over de bodem inzichtelijk kan worden gemaakt en die tevens kunnen worden toegepast in Monte Carlo studies over onzekerheidsvoortplanting. Drie benaderingen voor het verkrijgen van de complexe kansverdelingen worden beschreven: kriging, Bayesian Maximum Entropy en Markov random fields. Alle drie de methoden kunnen worden toegepast zowel op kwantitatieve (continue) als kwalitatieve (categorische) bodemeigenschappen. De theorie wordt geïllustreerd met een simulatie van bodemeigenschappen die worden gebruikt als invoer voor het bodemverzuringsmodel SMART2. Geconcludeerd wordt dat de kriging-methoden goed ontwikkeld zijn en hun nut vele malen hebben bewezen, maar dat Bayesian Maximum Entropy en Markov random fields waardevolle alternatieven vormen die het overwegen waard zijn, met name als het gaat om het gebruik van statistische modellen voor kwalitatieve ruimtelijke variabelen.

Trefwoorden: Bodemkaart, Onzekerheid, Kriging, Geostatistiek, Bayesian Maximum Entropy, Markov random fields

ISSN 1871-028X

Postbus 47, 6700 AA Wageningen.
Tel: (0317) 47 47 00; fax: (0317) 41 90 00; e-mail: info.alterra@wur.nl

The WOt Reports series is published by the Statutory Research Tasks Unit for Nature and the Environment, which is part of Wageningen University & Research Centre. This report can also be obtained from the Unit’s secretarial office**. The report can also be downloaded from www.wotnatuurenmilieu.wur.nl.**

Statutory Research Tasks Unit for Nature & the Environment, P.O. Box 47, NL-6700 AA Wageningen, The Netherlands
Phone: +31 317 47 78 44; Fax: +31 317 42 49 88; e-mail: info.wnm@wur.nl;
Internet: www.wotnatuurenmilieu.wur.nl

F-0010 [2007] Project WOT-04-002 – 5231083-02 [Rapport 58 – december 2007]

Contents

Preface 3

Summary 5

Samenvatting 9

1 Introduction 13

1.1 Defining the problem . . . 13
1.2 Aim of project . . . 14
1.3 Modelling uncertain soil properties with probability distributions . . 15
1.3.1 Uncertain soil properties at a single location . . . 16
1.3.2 Spatially distributed uncertain soil properties . . . 17
1.3.3 Multiple uncertain soil properties . . . 17
1.4 Prediction versus simulation . . . 18
1.4.1 Prediction . . . 19
1.4.2 Simulation . . . 19
1.5 Contents of this report . . . 21

I QUANTITATIVE SOIL MAPS 23

2 Kriging methods 25

2.1 Observations at points only . . . 26
2.1.1 Simple kriging . . . 26
2.1.2 Ordinary kriging . . . 27
2.1.3 Co-kriging . . . 28
2.2 Observations at points and map of covariable . . . 29
2.2.1 Universal kriging . . . 29
2.2.2 Stratified kriging . . . 30
2.3 Simulation . . . 30

3 Bayesian Maximum Entropy 33

3.1 Observations at points only . . . 35
3.1.1 Computing the unconditional multi-point pdf . . . 35
3.1.2 Conditioning the multi-point pdf . . . 36
3.2 Observations at points and map of target variable . . . 37
3.3 Simulation . . . 37

4 Markov Random Fields 39

4.1 The Markov property . . . 40
4.2 Estimation of the conditional density . . . 41
4.3 Simulation . . . 42

4.3.1 Conditioning to point observations 43
4.3.2 Conditioning to maps of covariables 43
4.4 Prediction 44

II QUALITATIVE SOIL MAPS 45

5 Kriging methods 47
5.1 Observations at points only 48
5.1.1 Simple indicator kriging 48
5.1.2 Ordinary indicator kriging 49
5.1.3 Indicator-cokriging 50
5.1.4 Soft-indicator kriging 51
5.2 Observations at points and map of categories 51
5.2.1 Simple indicator kriging with local prior means 51
5.3 Indicator simulation of soil categories 52

6 Bayesian Maximum Entropy 53
6.1 Observations at points only 53
6.1.1 Computing the unconditional multi-point pdf 54
6.1.2 Conditioning the multi-point pdf 56
6.2 Observations at points and map of categories 57
6.2.1 Computing the unconditional multi-point pdf 58
6.2.2 Conditioning the multi-point pdf 59
6.3 Simulation of soil categories with Bayesian Maximum Entropy . . . 59

7 Markov Random Fields 61
7.1 Estimation of the conditional probability 61
7.2 Simulation and Prediction 62

III EXAMPLES: MAPPING SOIL PROPERTIES USED BY THE SMART - SUMO - MOVE MODEL CHAIN 65

8 Introduction 67

9 BME simulation of soil type 69

10 Sequential Gaussian simulation of continuous soil properties 75
10.1 Exploratory data analysis 75
10.2 Variography 77
10.3 Spatial prediction and simulation 78
10.3.1 Selectivity constant for H - (Ca,Mg) exchange (KHBC2) . . . 78
10.3.2 Selectivity constant for Al - (Ca,Mg) exchange (KAlBC2) . . 79
10.3.3 Dissolution constant for Al-hydroxide ($\mathrm{KAl_{ox}}$) 81

11 Discussion and conclusions 85

Bibliography 87

Preface

Alterra has a long tradition in making soil maps. These soil maps, together with the observations at points (soil profile descriptions) are stored in a unique Soil Information System covering the Netherlands. The information in this system is widely used for a large variety of applications. Decisions in e.g. land use planning and environmental policy making are based on this information. Triggered by the discussion on the quality of environmental models, initiated in 1999 by Hans de Kwaadsteniet of the National Institute for Public Health and the Environment (RIVM) through an interview in the newspaper 'Trouw', Alterra also started research on the quality of their models and databases. For the Soil Map of the Netherlands 1 : 50 000, already in 1988 a project started, under the supervision of Jaap de Gruijter, to quantify the quality of this map in terms of purity and spatial variation of soil properties within mapping units. This project resulted in valuable information on the quality of soil maps. For some applications, this information is not enough, and information is required on the spatial distribution of errors in the soil map. This knowledge gap was the main motivation of the research reported here.

The study was part of the program Basic Research for the Netherlands Environmental assessment Agency (MNP). This program is funded by the Dutch Ministry of agriculture, Nature and Food Quality (LNV). We would like to thank Harm Houweling of MNP for his efforts to keep this important topic on the research agenda. Also we would like to thank Dimitri D'Or of FSS International. Large parts of this report, mainly those on Bayesian Maximum Entropy, are based on his PhD thesis.

Summary

Soil maps are usually constructed such that they provide the best estimate of the spatial distribution of the soil. However, the estimates contained in the soil map are not perfect because they are typically based on limited knowledge and limited information. To acknowledge that soil maps are not free of errors, the uncertainty in the estimates may therefore be represented with error bounds that characterize the accuracy of the map. This report uses (geo)statistical methods to derive maps of soil type and soil properties from soil observations and explanatory information. A principal property of these methods is that they do not only produce a map but also quantify the associated uncertainty.

In statistics, the value of an uncertain variable, such as that of a soil property at a given geographic location, is characterized by a probability distribution. This signifies that there is not one possible outcome but instead that there is a range of possible outcomes, each with a probability of occurrence. Wide probability distributions mean large uncertainty, narrow distributions refer to small uncertainty. For soil properties that are spatially distributed, a complex probability distribution must be defined, with parameters that may vary with location and that includes spatial correlation. Once the full probability distribution of a spatially distributed soil property is derived it may be used for prediction and stochastic simulation. Prediction refers to taking at each location a central value from the probability distribution, such as the mean, median or mode. This results in a single deterministic map that provides the best estimate of the soil property at each location. Stochastic simulation is used to generate a possible reality from the probability distribution of the soil property, using a pseudo-random number generator. Stochastic simulation produces an infinite number of possible realities, which is useful to communicate the uncertainty about the soil and can also be used in Monte Carlo uncertainty propagation studies. These and some other introductional issues are discussed in the first chapter of this report.

The report is then split in three main parts. Parts I and II present the theory of three approaches to deriving the complex probability distribution of a soil property. The three approaches are Kriging methods, Bayesian Maximum Entropy, and Markov random fields. Part I (chapters 2 to 4) addresses quantitative soil properties that are measured on a continuous numerical scale, part II (chapters 5 to 7) deals with qualitative soil properties that are measured on a categorical or discrete numerical scale. Part III (chapters 8 to 10) applies several of the presented methods to map soil type and selected soil chemical properties for natural areas in the Netherlands.

Chapter 2 reviews existing Kriging methods for quantitative properties. The basic idea of kriging is to exploit the fact that observations close to one another are more alike than observations further apart. The spatial correlation is quantified by means

of a (auto)covariance function or semivariogram, which is subsequently used to make predictions at unobserved locations. The chapter discusses simple, ordinary and cokriging and extends these to universal and stratified kriging. The latter two not only use point observations for prediction and simulation, but also use spatially exhaustive explanatory information, either to define a trend function or to divide the area into relatively homogeneous strata.

The Bayesian Maximum Entropy (BME) method presented in chapter 3 also characterizes uncertain spatial variables with probability distribution functions. However, it takes a somewhat different approach than kriging and can therefore result in different probability distributions characterizing the same soil property. The basic idea is to choose probability distribution functions that have maximum entropy, while satisfying a number of constraints. The entropy of a random variable is a measure of its uncertainty. The larger the entropy, the larger the uncertainty. Common constraints are that the mean, variance and spatial correlation of the resulting probability distribution are known and fixed. Under these three constraints, BME yields a normal distribution and the results are identical to those of kriging. However, BME can also handle different and additional constraints. In this respect it is more flexible than kriging, but the price paid for this is an increased numerical complexity. The two main steps in BME are first to compute the unconditional probability distribution of the variable using a numerical technique known as iterative scaling, and second to condition the distribution on the available observations.

The goal of the Markov random field (MRF) approach discussed in chapter 4 is similar to that of Kriging and Bayesian Maximum Entropy. It aims to characterize uncertain spatially distributed variables with probability distribution functions. In the case of MRF, the starting point is defined in terms of *conditional* probability distributions, in which case the probability of occurrence of the variable of interest at some location is defined conditional to the value of the variable at neighbouring locations. Key property of the MRF approach is that it assumes that the local neighbourhood contains all information necessary to characterize the probability distribution of the variable. In addition, it uses numerical approaches, in particular Markov Chain Monte Carlo simulation, to compute predictions and simulations of the uncertain variable. MRF is not as well developed as BME and kriging, but it appears a promising technique that has recently drawn increased attention, simultaneously with the development of Markov Chain Monte Carlo methods.

Chapters 5 to 7 present the theory of Kriging, BME and MRF for qualitative spatial variables. For Kiging, this leads to indicator kriging techniques, which are well-developed in geostatistics but have been criticized for their weak statistical basis. BME is a good alternative, although the numerical complexity is quite demanding, particularly when the number of classes is large. MRF can handle a large number of classes more easily, but it is not easy to estimate the many parameters required by the method and to verify that the chosen parameter combination is statistically valid.

The theory presented in parts I and II is illustrated with examples in part III. Chapter 8 gives an introduction, in which background information about the data used and the necessity of stochastic simulation of the soil properties considered is explained. The data are soil type and chemical soil properties of the Netherlands, as used by the soil acidification model SMART2. An uncertainty propagation analysis of SMART2 requires that the uncertainty in soil type and soil properties is characterized using probability distribution functions and that 'possible realities' of these variables are generated using spatial stochastic simulation.

Chapter 9 uses the Bayesian Maximum Entropy method to simulate the Dutch soil type for all natural areas in the Netherlands. Seven different soil types are considered. Information used to build the statistical model and condition the simulated maps are provided by the 1 : 50 000 Dutch soil map and observations of soil type at over 8000 points. The point observations are considered error-free, whereas the soil map information is not. Therefore, the simulated soil type at some location may differ from that predicted by the 1 : 50 000 soil map. Indeed the simulated maps show that deviations occur, but overall the simulated maps do not differ much from the 1 : 50 000 soil map and among each other. Apparently, the information contained in the 1 : 50 000 soil map and in the 8000 observations is that high that little uncertainty about the spatial distribution of soil type remains.

Three quantitative soil properties were interpolated and simulated using the Kriging approach in chapter 10. These were the selectivity constant for H - (Ca,Mg) exchange, the selectivity constant for Al - (Ca,Mg) exchange, and the dissolution constant for Al-hydroxide. Exploratory data analysis on 317 observations from forest soils in the Netherlands showed that the latter of the three was severely skewed. It was therefore log-transformed prior to the statistical modelling. The analysis also showed that the selectivity constant for Al - (Ca,Mg) exchange was significantly correlated with soil type. The mean of this variable was therefore modelled as soil type-dependent and a universal kriging approach was used. The semivariograms of all three variables showed clear spatial dependence, with spatial correlation lengths ranging from 60 to 100 km. Consequently, the prediction maps showed clear spatial patterns, some of which could not easily be explained. As expected, the simulated maps of the soil properties are more noisy than the prediction maps, but the overall spatial pattern is still present. Differences between the simulated maps were meaningful and indicate that the uncertainty about the spatial distribution of these properties is far from negligible.

The final chapter discusses the advantages and disavantages of the three approaches described in this report. Kriging methods are well-developed and have proven their use on numerous occasions, but the Bayesian Maximum Entropy and Markov random field approach are valuable alternatives that deserve attention, particularly for statistical modelling of qualitative spatial variables.

Samenvatting

Bodemkaarten worden gewoonlijk dusdanig samengesteld dat ze een zo goed mogelijk beeld bieden van de ruimtelijke verdeling van de bodem. De bodemkaart is echter niet volmaakt, omdat ze gebaseerd is op beperkte kennis en informatie. Om aan te geven dat bodemkaarten niet vrij zijn van fouten, zou de onzekerheid in de schattingen kunnen worden weergegeven door middel van foutenmarges, die de mate van nauwkeurigheid van de kaart weergeven. In het huidige rapport worden met behulp van (geo)statistische methoden kaarten van bodemtypen en bodemeigenschappen afgeleid uit waarnemingen aan de bodem en informatie over omgevingskenmerken die gerelateerd zijn aan de bodem. Een belangrijk kenmerk van deze methoden is dat ze niet alleen een kaart opleveren maar tevens de bijbehorende mate van onzekerheid kwantificeren.

In de statistiek wordt de waarde van een onzekere variabele, zoals die van een bodemeigenschap op een gegeven geografische locatie, gekenmerkt door een bepaalde kansverdeling. Dit houdt in dat er niet maar n bepaalde uitkomst mogelijk is, maar een zekere range aan mogelijke uitkomsten, die elk met een bepaalde mate van waarschijnlijkheid kunnen optreden. Brede kansverdelingen betekenen een hoge mate van onzekerheid, terwijl smalle verdelingen overeenkomen met een lage onzekerheid. Voor ruimtelijk verdeelde bodemeigenschappen moet een complexe kansverdeling worden gedefinieerd, met parameters die kunnen variren met de locatie, en die voorzien is van een ruimtelijke correlatie. Wanneer de volledige kansverdeling van een ruimtelijk verdeelde bodemeigenschap eenmaal is afgeleid, kan deze worden gebruikt voor voorspellingen en stochastische simulatie. Voorspelling verwijst hierbij naar het voor elke locatie nemen van een centrale waarde uit de kansverdeling, zoals het gemiddelde, de mediaan of de modus. Dit resulteert in n enkele deterministische kaart met voor elke locatie de beste schatting van de bodemeigenschap. Stochastische simulatie wordt gebruikt om een mogelijke werkelijkheid te genereren uit de kansverdeling van de bodemeigenschap, met behulp van een generator van pseudo-toevalsgetallen. Stochastische simulatie leidt tot een oneindig aantal mogelijke werkelijkheden, wat kan worden gebruikt om de mate van onzekerheid over de bodem inzichtelijk te maken, evenals in Monte Carlo-studies over onzekerheidsvoortplanting. Deze en enkele andere achtergrondaspecten worden besproken in Hoofdstuk 1.

De rest van het rapport is onderverdeeld in drie gedeelten. In Deel I en II wordt de theorie van drie benaderingen voor het afleiden van de complexe kansverdeling van een bodemeigenschap beschreven. Bij deze drie benaderingen gaat het respectievelijk om kriging-methoden, Bayesian Maximum Entropy en Markov random fields. In deel I (hoofdstuk 2 t/m 4) worden kwantitatieve bodemeigenschappen behandeld gemeten op een continue numerieke schaal, terwijl het in deel II (hoofdstuk 5 t/m 7) gaat om kwalitatieve bodemeigenschappen gemeten op een categorische of discrete numerieke schaal. In deel III (hoofdstuk 8 t/m 10) worden verschillende van de beschreven methoden toegepast voor het in kaart brengen van bodemtypen en enkele bodemchemische eigenschappen in natuurgebieden in Nederland.

In hoofdstuk 2 wordt een overzicht gegeven van bestaande kriging-methoden voor kwantitatieve eigenschappen. Het uitgangspunt bij kriging is dat gebruik wordt gemaakt van het gegeven dat waarnemingen die dicht bij elkaar zijn gedaan meer op elkaar lijken dan waarnemingen die op grotere afstand van elkaar zijn gedaan. De ruimtelijke correlatie wordt gekwantificeerd door middel van een (auto)covariantiefunctie of semivariogram, dat dan vervolgens wordt gebruikt om voorspellingen te doen over locaties waar geen waarnemingen zijn gedaan. In dit hoofdstuk bespreken we simple, ordinary en cokriging, en breiden dit uit naar universal en stratified kriging. Bij deze laatste twee worden voor voorspellingen en simulaties niet alleen puntwaarnemingen gebruikt, maar ook gebiedsdekkende verklarende informatie, hetzij om een trendfunctie te definiren hetzij om het onderzoeksgebied in relatief homogene strata onder te verdelen.

Bij de in het derde hoofdstuk besproken Bayesian Maximum Entropy (BME) methode worden ook onzekere ruimtelijke variabelen gekarakteriseerd met behulp van kansverdelingsfuncties. Hierbij wordt echter een enigszins andere benadering gebruikt dan bij kriging, wat kan leiden tot een andere kansverdeling voor dezelfde bodemeigenschap. Het uitgangspunt is het kiezen van de kansverdelingsfuncties met de hoogste entropie, waarbij moet worden voldaan aan een aantal randvoorwaarden. De entropie van een stochastische variabele is een maat voor de onzekerheid ervan. Hoe hoger de entropie, hoe groter de onzekerheid. Gebruikelijke randvoorwaarden zijn onder andere dat het gemiddelde, de variantie en de ruimtelijke correlatie van de resulterende kansverdeling bekend en overal hetzelfde moeten zijn. Onder deze drie randvoorwaarden levert BME een normale verdeling op, met resultaten die identiek zijn aan die van kriging. BME kan echter ook omgaan met andere en bijkomende randvoorwaarden. In dit opzicht is deze methode meer flexibel dan kriging, maar de prijs die hiervoor wordt betaald is een grotere numerieke complexiteit. De twee voornaamste stappen bij BME zijn ten eerste het berekenen van de onvoorwaardelijke kansverdeling van de variabele met behulp van een numerieke techniek die iterative scaling wordt genoemd, en ten tweede het conditioneren van de verdeling op de beschikbare waarnemeringen

Het doel van de in hoofdstuk 4 besproken Markov random fields (MRF) benadering lijkt op dat van kriging en van BME. Doel is het karakteriseren van onzekere ruimtelijk verdeelde variabelen met behulp van kansverdelingsfuncties. In het geval van MRF wordt begonnen met het definiren van voorwaardelijke kansverdelingen, waarbij de kans dat de bestudeerde variabele op een locatie een bepaalde waarde aanneemt wordt gedefinieerd in termen van de waarde van deze variabele op naburige locaties. Een essentile eigenschap van de MRF-benadering is dat hierbij wordt aangenomen dat de plaatselijke omgeving alle informatie bevat die nodig is voor het karakteriseren van de kansverdeling van de variabele. Daarnaast wordt gebruik gemaakt van numerieke benaderingen, met name Markov Chain Monte Carlo simulatie, voor het berekenen van voorspellingen en simulaties van de onzekere variabele. MRF is nog niet zo ver ontwikkeld als BME en kriging, maar lijkt toch een veelbelovende techniek, die de laatste tijd steeds meer aandacht krijgt, naarmate de Markov Chain Monte Carlo methoden verder worden ontwikkeld.

In hoofdstuk 5 tot en met 7 wordt de theorie achter kriging, BME en MRF voor kwalitatieve ruimtelijke variabelen besproken. Voor kriging leidt dit tot indicator-kriging technieken, die in de geostatistiek goed ontwikkeld zijn, maar waarop kritiek is geleverd vanwege hun zwakke theoretische onderbouwing. BME is een goed alternatief, ondanks de nogal veeleisende numerieke complexiteit, vooral wanneer het aantal klassen groot is. MRF heeft minder moeite met grote aantallen klassen, maar het is niet gemakkelijk het grote aantal parameters dat de methode vereist te schatten en te verifiren dat de gekozen combinatie van parameters statistisch valide is.

De in deel I en II gepresenteerde theorie wordt in deel III gellustreerd met voorbeelden. In Hoofdstuk 8 wordt ter inleiding achtergrondinformatie gepresenteerd over de gebruikte gegevens en wordt de noodzaak van het gebruik van stochastische simulatie van de bestudeerde bodemeigenschappen toegelicht. De gegevens betreffen bodemtypen en chemische bodemeigenschappen voor Nederland, zoals die worden gebruikt in het bodemverzuringsmodel SMART2. Voor een analyse van de manier waarop onzekerheden zich voortplanten in SMART2 is het nodig om de onzekerheid in bodemtype en bodemeigenschappen te karakteriseren met behulp van kansverdelingsfuncties en om de mogelijke werkelijkheden van deze tien in WOt-rapport 58 genoemde variabelen te genereren met behulp van ruimtelijke stochastische simulatie.

In hoofdstuk 9 wordt met behulp van de Bayesian Maximum Entropy methode voor alle natuurgebieden in Nederland het bodemtype gesimuleerd. Hierbij worden zeven verschillende bodemtypen meegenomen. De informatie die nodig is om het statistische model te construeren en de gesimuleerde kaarten te conditioneren wordt verkregen uit de bodemkaart van Nederland (1:50.000) en uit waarnemingen van het bodemtype afkomstig van meer dan 8000 punten. Van de puntwaarnemingen wordt aangenomen dat deze foutloos zijn, in tegenstelling tot de informatie uit de kaart. Het kan dus voorkomen dat het gesimuleerde bodemtype voor een bepaalde locatie verschilt van het type dat door de 1:50.000 kaart wordt aangegeven. Uit de gesimuleerde kaarten blijkt inderdaad dat er afwijkingen optreden, maar over het algemeen verschillen de gesimuleerde kaarten niet sterk van de 1:50.000 kaart, en ook vertonen ze geen grote onderlinge verschillen. Kennelijk bevatten de 1:50.000 bodemkaart en de 8000 waarnemingen zo veel informatie dat er wat betreft de ruimtelijke verdeling van het bodemtype weinig onzekerheid overblijft.

In hoofdstuk 10 worden met behulp van kriging drie kwantitatieve bodemeigenschappen genterpoleerd en gesimuleerd. Hiervoor werden gekozen de selectiviteitsconstante voor H (Ca, Mg) uitwisseling, de selectiviteitsconstante voor Al (Ca, Mg) uitwisseling en de oplossingsconstante voor Al-hydroxide. Uit een orinterende gegevensanalyse van 317 waarnemingen afkomstig van Nederlandse bosbodems bleek dat de derde van deze eigenschappen een zeer scheve verdeling had. Daarom werden deze gegevens voorafgaand aan het gebruik in het statistische model log-getransformeerd. Uit de analyse bleek ook dat de selectiviteitsconstante voor Al (Ca, Mg) uitwisseling significant gecorreleerd was met het bodemtype. Daarom werd de gemiddelde waarde van deze variabele gemodelleerd als bodemtype-afhankelijk, en werd universal kriging toegepast. Uit het semivariogram van alle drie de variabelen bleek een duidelijke ruimtelijke afhankelijkheid, waarbij de ruimtelijke correlatielengte uiteenliep van 60 tot 100 km. Als gevolg hiervan vertoonden de voorspellingskaarten duidelijke ruimtelijke patronen, waarvan sommige niet eenvoudig te verklaren waren. Zoals verwacht bevatten de gesimuleerde kaarten van de bodemeigenschappen meer ruis dan de voorspellingskaarten, maar het algehele ruimtelijke patroon blijft aanwezig. De verschillen tussen de gesimuleerde kaarten waren betekenisvol en geven aan dat de onzekerheid in de ruimtelijke verdeling van deze eigenschappen beslist niet verwaarloosbaar is.

In het laatste hoofdstuk worden de voor- en nadelen van de drie in dit rapport beschreven benaderingen besproken. De kriging-methoden zijn goed ontwikkeld en hebben hun nut talloze malen bewezen, maar Bayesian Maximum Entropy en Markov random fields vormen waardevolle alternatieven die het overwegen waard zijn, met name als het gaat om het gebruik van statistische modellen voor kwalitatieve ruimtelijke variabelen.

Chapter 1

Introduction

1.1 Defining the problem

Soil maps are used for many purposes. To name a few examples, they are used in planning to evaluate land allocation scenarios, in agronomy to assess the suitability of the land for growing crops or assess the faith of pollutants such as pesticides, in ecology to develop nature conservation plans, and in hydrology and climatology to describe the role of the soil in the hydrological cycle. Since many years, national governments and international organisations have therefore put much effort in mapping the soil. Soil maps are also increasingly used to derive spatially distributed soil inputs to environmental and ecological process models. For instance, soil maps provide important information about physical, chemical and biological soil properties needed by acidification, groundwater flow and greenhouse gas emission models. However, it is well known that soil maps are not perfect and contain errors. Soil map units are usually impure, meaning that not all locations in a soil map unit have the soil type associated with the unit. Also, continuous soil properties such as pH, clay and organic carbon content always show spatial variation within soil map units, even within units that are perfectly pure. Hence thematic soil maps that are derived from the general soil map by assigning constant values to map units using pedo-transfer functions will contain errors, even if the general soil map would be free of errors.

Besides soil maps, many countries and organisations also have soil information systems, which are digital data bases that store measurements of the soil at point locations. At these points, the soil profile is described and classified. Basic soil properties such as the organic carbon and clay content of the soil horizons are estimated in the field by hand-estimates, or measured in the laboratory, and stored in the data base. Much of the data collected in this way can be considered as hard information, i.e. relatively certain, when compared to the information derived from the soil map. Nonetheless, these data are also not error-free and are affected by measurement and interpretation errors. Maps of specific soil properties obtained by spatial interpolation in between the observation points are in addition affected by interpolation error.

In the Netherlands the situation is as described above. The soil map at scale 1 : 50 000, which was completed in the 1990s (Stichting voor Bodemkartering, 1995), is used in many agricultural, environmental and ecological studies. Many of these studies are commissioned by the Statutory Research Tasks Unit for Nature and the Environment. The soil map of the Netherlands 1 : 50 000 is also used to generate soil-related input for process models such as the soil acidification model SMART, using pedo-transfer functions. Stimulated by a national discussion on the accuracy of outputs of process models used by the National Institute for Public Health and the Environment (RIVM), the Board of Directors of Wageningen University and Research Centre and the Directors of Alterra installed a task force for the quality control of datasets and models used for the Statutory Research Tasks. In their final report, the task force presents several recommendations how to improve the quality control of the datasets and models (Jansen (ed), 2004). In the meantime, several studies were executed to analyze the uncertainty of the output of individual process models and of chains of models (Schouwenberg et al., 2000; van Dobben et al., 2002a; Wamelink et al., 2003). However, in these studies only the contribution of uncertain non-spatial model parameters was analyzed, whereas the contribution of spatially varying input variables such as soil properties was ignored. The main reason for this was the lack of information about the uncertainties associated with the soil map and properties derived from it. An exception is the study by Brus and Jansen (2004) on uncertainty of heavy metal concentrations in arable crops. They studied the combined effect of uncertainty about: 1) basic soil properties derived from the soil map and heavy metal concenrttaions in soil, both used as model input; 2) model parameters and, 3) residual variance, on the model output.

Several validation studies have been done to assess the quality of the soil map 1 : 50 000 (Marsman and de Gruijter, 1986). Although these studies reveal the overall quality of the soil map 1 : 50 000, expressed in terms of map purity and spatial variance of quantitative soil properties within mapping units (Visschers et al., 2007), in many cases this is insufficient information for a spatial uncertainty analysis. The accuracy obtained is not location-specific but merely an overall measure. If the model includes spatial interactions or if one wants to compute spatial aggregates of the model output, then one must also characterize the uncertainty about the spatial distribution of the soil property, including how the uncertainty at one location is related to that at neighbouring locations and beyond.

Rather than validation studies that quantify the accuracy of an existing map afterwards in a global measure such as purity or mean squared error, there is a need for methods that yield maps of soil types and soil properties that simultaneously quantify the accuracy of the map produced. In order to be used in uncertainty propagation studies, the accuracy must be fully specified and include location-specific accuracy measures as well as quantify the spatial dependence of uncertainties. This calls for a geostatistical approach, in which case maps of soil types and soil properties are represented by probability distributions.

1.2 Aim of project

In this report we describe and illustrate statistical methods for mapping soil properties measured on a qualitative scale (categorical and ordinal data, such as soil type, soil colour and soil erodibility class) and a quantitative scale (interval and ratio data, such as soil pH, organic matter content and soil thickness). Qualitative soil maps are also referred to as categorical soil maps. Quantitative variables can be discrete (finite or countably infinite number of possible outcomes) or continuous (infinite number of possible outcomes). In this report we consider continuous quantitative data only. An important aspect in the evaluation is the appropriateness of the method for handling different sources of data and information, such as certain (effectively error-free, 'hard') and uncertain ('soft') point observations of the target soil property, and a map of the soil property. In the case of a map one has spatially exhaustive information, but this information is usually uncertain, i.e. the values depicted on the map must be interpreted as merely an estimate of the target property. Also the possibility of exploiting ancillary information, i.e. information on covariates related to the target soil property, observed at points or contained in maps, is evaluated. Examples of ancillary information on maps are digital terrain models, land use and geological maps and remote sensing images.

Key property of all statistical methods presented in this work is that they not only produce a map of the soil property, but that they also quantify the uncertainty associated with the map produced. This is what is lacking in most of the existing soil maps and soil information systems. A convenient and powerful way to describe uncertainty is to make use of probability distribution functions. Before presenting an outline of this report, we therefore first introduce the basic concepts of probability distributions and describe how probability theory and statistics may be used to describe the uncertainty in mapped soil properties.

1.3 Modelling uncertain soil properties with probability distributions

In the previous section we observed that soil maps are rarely, if ever, free of errors. The problem is that although we may know that a soil map contains errors, we do not in fact know these errors. After all, if we would know them then we could simply eliminate the errors by adjusting the map. For example, if the soil map specifies that at some location the clay content of the topsoil equals 32 per cent, while we know that in reality it is 27 per cent, then we could simply eliminate the error by subtracting 5 per cent from the mapped value. If we would measure the clay content everywhere with negligible measurement error, then we could easily adjust the soil map everywhere and obtain an error-free map of soil clay content. However, in reality we are not able to do so, because we do not have the resources to exhaustively and accurately measure soil clay content with sufficient spatial resolution.

Although we may not know the error, in many cases we do have some idea about the distribution of values that the error is likely to take. For example, we may know or have sufficient confidence that the chances are equal that the error is positive or negative (no bias or systematic error), or we may know that in only one out of ten cases the absolute error is greater than a given number (known width of confidence interval). Such knowledge may be based on validation data, expert judgement or, under certain assumptions, be derived from the spatial variation of the soil property and the method that was used to create the soil map. A major part of this report is dedicated to how the error in maps of soil properties can be characterized in this way. Where it boils down to in the end is that, although we may not be able to determine the actual error, we are able to quantify the error in statistical terms using probability distribution functions. In this section we introduce the theory and basic concepts of probability distribution theory applicable to spatially distributed soil properties.

1.3.1 Uncertain soil properties at a single location

For a continuous soil property at a single location, we represent its unknown (because of uncertainty) value by a *random variable*. A random variable has no fixed value but has many (often infinite) possible values, each with a certain probability of occurrence. A random variable is completely characterized by its *probability distribution function* (pdf). If we denote the continuous soil property by $\hat{Z}$ then its cumulative pdf $F_{\hat{Z}}$ at some location $\mathbf{s}$ is defined as:

$$F_{\hat{Z}}(\hat{z}, \mathbf{s}) = P[\hat{Z}(\mathbf{s}) \leq \hat{z}] \tag{1.1}$$

where P represents probability. Note that the hat-symbol ($\hat{}$) is used to underline that the random variable is an *estimator* or *predictor* of the true soil property. In fact, this notation will also be useful in chapter 2, where a distinction between Z and $\hat{Z}$ will be made. For all $\mathbf{s}$, the function $F_{\hat{Z}}$ must be a non-decreasing continuously differentiable function of $\hat{z}$ with limit values $F_{\hat{Z}}(-\infty, \mathbf{s}) = 0$ and $F_{\hat{Z}}(\infty, \mathbf{s}) = 1$. There are no other restrictions. The shape of $F_{\hat{Z}}$ can take many forms, although in practice one often parametrizes the pdf to a common shape, such as the normal, lognormal, exponential or uniform distribution. Important parameters of the pdf are its mean $\mu_{\hat{Z}}(\mathbf{s})$, which represents the expected or average value of $\hat{Z}(\mathbf{s})$, and its variance $V_{\hat{Z}}(\mathbf{s})$, which characterizes the variation or spread of $\hat{Z}(\mathbf{s})$ around its central value. The mean and variance are defined as:

$$\mu_{\hat{Z}}(\mathbf{s}) = E[\hat{Z}(\mathbf{s})] = \int_{-\infty}^{\infty} \hat{z} \cdot f_{\hat{Z}}(\hat{z}, \mathbf{s}) \, \mathrm{d}\hat{z} \tag{1.2}$$

$$V_{\hat{Z}}(\mathbf{s}) = E[(\hat{Z}(\mathbf{s}) - \mu_{\hat{Z}}(\mathbf{s}))^2] = \int_{-\infty}^{\infty} (\hat{z} - \mu_{\hat{Z}}(\mathbf{s}))^2 \cdot f_{\hat{Z}}(\hat{z}, \mathbf{s}) \, \mathrm{d}\hat{z} \tag{1.3}$$

where E means mathematical expectation and $f_{\hat{Z}}$ is the mathematical derivative of $F_{\hat{Z}}$, commonly known as the probability density function. The variance $V_{\hat{Z}}(\mathbf{s})$, or alternatively its square root the standard deviation $\sigma_{\hat{Z}}(\mathbf{s})$, is a measure for the uncertainty in the soil property. For example, when the soil property is assumed normally distributed then there is a 95% chance that the true value of the soil property lies within two standard deviations from its expected value.

Uncertain categorical soil properties are also characterized with pdfs. However, the difference is that here discrete pdfs are used, since the number of outcomes of the soil property is typically finite. The discrete pdf π_C characterising the uncertain categorical soil property C at $\mathbf{s}$ is given by

$$\pi_C(c_i, \mathbf{s}) = P[C(\mathbf{s}) = c_i] \tag{1.4}$$

where $i = 1, \ldots, n_c$ with n_c the number of soil categories. The uncertain soil property is fully characterized at $\mathbf{s}$ when $\pi_C(c_i, \mathbf{s})$ is known for all c_i. Each of the individual probabilities must be greater or equal to zero and their sum must equal one. Common shapes for discrete pdfs are the uniform, binomial and multinomial distribution. In soil science applications non-parametric discrete pdfs are also often used, which means that one simply needs to specify the probability for all soil categories. The degree of uncertainty about the categorical soil property is contained in these probabilities. An uncertain categorical soil property will have multiple outcomes with non-zero probability. A certain categorical soil property will have only one outcome with probability one, while all other outcomes have probability zero.

1.3.2 Spatially distributed uncertain soil properties

We can be uncertain about the soil at multiple or all locations in an area of interest. In this case we do not have a single random variable, but a set of random variables, one for each location in the study area. This set of random variables is referred to as a Spatial Random Field (SRF). The (one-point) pdfs defined above can then be specified for each location and used to describe the uncertainty at all locations within the area, yielding pdfs and pdf parameters for all locations $\mathbf{s}$ in the domain of interest D. Thus, neither the mean $\mu_{\hat{Z}}(\mathbf{s})$ and variance $V_{\hat{Z}}(\mathbf{s})$ of an uncertain continuous soil property nor the probability $\pi_C(\mathbf{s})$ of an uncertain categorical soil property need be constant in space but will typically vary with $\mathbf{s}$.

In addition, spatially distributed uncertain soil properties will usually be spatially dependent. In order to represent spatial dependence, it is necessary to specify the *multi-point* probability distribution function:

$$F_{\hat{Z}\ldots\hat{Z}}(\hat{z}_1,\ldots,\hat{z}_n,\mathbf{s}_1,\ldots,\mathbf{s}_n) = P[\hat{Z}(\mathbf{s}_1)\leq\hat{z}_1,\ldots,\hat{Z}(\mathbf{s}_n)\leq\hat{z}_n] \tag{1.5}$$

$$\pi_{C\ldots C}(c_i,\ldots,c_j,\mathbf{s}_1,\ldots,\mathbf{s}_n) = P[C(\mathbf{s}_1)=c_i,\ldots,C(\mathbf{s}_n)=c_j] \tag{1.6}$$

For the continuous case, important parameters of the multi-point pdf are the vector of means $\mu_{\hat{Z}} = [\mu_{\hat{Z}}(\mathbf{s}_1)\ \mu_{\hat{Z}}(\mathbf{s}_2)\ldots\mu_{\hat{Z}}(\mathbf{s}_n)]^T$ and the $n\times n$ variance-covariance matrix $\mathbf{C}_{\hat{Z}}$, whose ij-th element stores the covariance of $\hat{Z}(\mathbf{s}_i)$ and $\hat{Z}(\mathbf{s}_j)$. The number of parameters of these multi-point pdfs can be extremely large, particularly in the discrete case. For instance, even for the relatively simple case of $n_c = 12$ categories and $n = 6$ locations, nearly three million (12^6) probabilities need to be specified. Clearly, simplifying assumptions are needed to reduce the number of parameters in practical applications. One important simplifying assumption that is frequently made is *stationarity*. Stationarity entails that the probability distributions, Eq. (1.5) and Eq. (1.6), are invariant to spatial translation, meaning that

$$\begin{aligned} &P[\hat{Z}(\mathbf{s}_1+\mathbf{h})\leq\hat{z}_1,\ldots,\hat{Z}(\mathbf{s}_n+\mathbf{h})\leq\hat{z}_n] \\ &\qquad = P[\hat{Z}(\mathbf{s}_1)\leq\hat{z}_1,\ldots,\hat{Z}(\mathbf{s}_n)\leq\hat{z}_n] \end{aligned} \tag{1.7}$$

$$\begin{aligned} &P[C(\mathbf{s}_1+\mathbf{h})=c_i,\ldots,C(\mathbf{s}_n+\mathbf{h})=c_j] \\ &\qquad = P[C(\mathbf{s}_1)=c_i,\ldots,C(\mathbf{s}_n)=c_j] \end{aligned} \tag{1.8}$$

for all $\mathbf{h}$. A slightly less rigid assumption for the continuous case is that of second-order stationarity, which states that $\mu_{\hat{Z}}$ does not depend on $\mathbf{s}$ and that the covariance between $\hat{Z}(\mathbf{s})$ and $\hat{Z}(\mathbf{s}+\mathbf{h})$ only depends on the separation vector $\mathbf{h}$. Under second-order stationarity, a useful parameter characterising the spatial dependence in $\hat{Z}$ is the autocovariance function $C_{\hat{Z}}(\mathbf{h})$:

$$C_{\hat{Z}}(\mathbf{h}) = Cov[\hat{Z}(\mathbf{s}),\hat{Z}(\mathbf{s}+\mathbf{h})] = E[(\hat{Z}(\mathbf{s})-\mu_{\hat{Z}}(\mathbf{s}))\cdot(\hat{Z}(\mathbf{s}+\mathbf{h})-\mu_{\hat{Z}}(\mathbf{s}+\mathbf{h}))] \tag{1.9}$$

1.3.3 Multiple uncertain soil properties

The soil has many properties. In the previous section we presented a generic statistical model for representing these properties in situations where we are uncertain about them and where these properties are spatially dependent. However, often the properties are also mutually dependent or cross-correlated. This may have an important impact on analyses that use multiple soil properties as inputs, such as in crop growth and soil erosion models. In order to take these mutual dependencies into account in a statistical framework, the joint distribution of multiple soil properties must be considered. In case of two continuous soil properties $\hat{Z}_1$ and $\hat{Z}_2$ or two categorical soil properties C and D, the joint distribution of the pair of variables is given by:

$$F_{\hat{Z}_1\hat{Z}_2}(\hat{z}_1, \hat{z}_2, \mathbf{s}, \mathbf{s}') = P[\hat{Z}_1(\mathbf{s}) \leq \hat{z}_1, \hat{Z}_2(\mathbf{s}') \leq \hat{z}_2] \tag{1.10}$$

$$\pi_{CD}(c_i, d_j, \mathbf{s}, \mathbf{s}') = P[C(\mathbf{s}) = c_i, D(\mathbf{s}') = d_j] \tag{1.11}$$

Extension to dependencies between more than two properties is straightforward. Under the second-order stationarity assumption, the statistical dependence between $\hat{Z}_1$ and $\hat{Z}_2$ as a function of distance is characterized by the cross-covariance function:

$$C_{\hat{Z}_1\hat{Z}_2}(\mathbf{h}) = Cov[\hat{Z}_1(\mathbf{s}), \hat{Z}_2(\mathbf{s} + \mathbf{h})] \tag{1.12}$$

The multivariate extension makes parametrisation even more complex.

Much of the theory developed in geostatistics and spatial statistics has been devoted to simplification of the generic models introduced in this section, in order to reduce the number of parameters such that they can reliably be estimated from the amount of information and observations that are typically available in practice. The challenge is to make assumptions that leave an estimable model that still captures the essentials of the studied soil properties and their uncertainties. The stationarity assumption is an important one, but particularly for the discrete case additional assumptions are required. Also, in some cases the stationarity assumption is too rigid and needs to be relaxed by specifying trend functions or splitting up the domain in strata that each have different probability models. These and other issues will be addressed in subsequent chapters, where state of the art approaches for modelling uncertain soil properties are described and illustrated.

1.4 Prediction versus simulation

The probability distributions defined in the previous section fully characterize the spatial and multivariate distribution of soil properties, including the uncertainties associated with them. However, in many practical circumstances users will not want to work with a complex probability distribution of a given soil property but will want to use a single map of that soil property, which must somehow be derived from the probability distribution. In such cases, it seems sensible to choose from the probability distribution that value which is most likely, where we obviously need to specify more precisely what we mean by 'most likely'. In statistics, this procedure is referred to as (spatial) *prediction*. Alternatively, users may want to generate multiple possible realities by sampling from the probability distributions, in analogy to throwing a die or tossing a coin. This is useful to visualize and communicate the uncertainty about the soil property and necessary for so-called Monte Carlo uncertainty propagation analyses (Heuvelink et al., 2007). This procedure is referred to as (stochastic) *simulation*. Both procedures are briefly described below.

1.4.1 Prediction

Let us consider the probability distribution of a continuous soil property at some location, as given in Eq. (1.1). The distribution conveys that the soil property can take an infinite number of possible values, each with a certain probability, but from these we wish to choose one, representative value, in the centre of the probability distribution. One possibility is to use the most likely value, now meaning the value which has the largest probability density. This value is known as the mode of the distribution, which, in fact, need not be unique. Another possibility, more common and also the one that we will use throughout this report, is to use the mean or expected value of the distribution, given in Eq. (1.2). This gives a unique value which in addition has the attractive property that it is *unbiased*, meaning that the expected value of the difference between the soil property and its predicted value equals zero:

$$E[(\mu_{\hat{Z}}(\mathbf{s}) - \hat{Z}(\mathbf{s}))] = \mu_{\hat{Z}}(\mathbf{s}) - E[\hat{Z}(\mathbf{s})] = 0 \tag{1.13}$$

The expected value of $\hat{Z}$, given by $\mu_{\hat{Z}}$, can be computed for all $\mathbf{s}$, yielding a prediction map of the continuous soil property. The prediction map may be interpreted as our 'best guess' of the true soil map, whereby unbiasedness guarantees that on average over- and underestimations cancel out. Unbiasedness means that there is no systematic error, but random errors cannot be eliminated. The magnitude of the random errors is quantified by the prediction error variance as given by Eq. (1.3).

Prediction maps of continuous soil variables tend to be much smoother than reality. Because it is unknown whether the soil property is greater or smaller than the mean of the probability distribution, it is best to use the mean as prediction, thus pulling the prediction towards the centre of the distribution. As a result, extremes in the probability distribution of the soil property are not represented in the prediction map.

For categorical soil properties we cannot compute an expected value. Instead, the obvious alternative is to use the category with the maximum probability for prediction, i.e. the mode. For example, when the probability associated with soil types A, B and C at some location equals 0.30, 0.55 and 0.15, respectively, then the predicted soil type would be soil type B. In rare cases where there are more categories that have equal and maximum probability, one of these categories may be chosen at random. The probability of making a wrong prediction is 1 minus the mode. Computing the prediction at all locations in the geographical domain of interest (in practice grid cells or polygons), yields a prediction map of the categorical soil variable. Note that using the mode for prediction means that soil types with small probabilities are not selected. Thus, soil types with large probabilities will be over-represented on the prediction map. For example, if at each location soil type A has probability 0.30 and type B has probability 0.70, then the prediction map will assign type B to all locations, whereas we would expect that in reality about 30 per cent of the area is occupied by soil type A.

1.4.2 Simulation

In stochastic simulation, it is not the most likely or expected value of the soil property that one aims for, but instead the goal is to generate a possible reality from the probability distribution of the uncertain soil property. As explained in section (1.3), we imagine that the true reality is one of an infinite number of possible realities that can be generated. Because of uncertainty we do not know which of the simulated realities is the true one. Therefore simulation is often used repeatedly (say 100 or 500 times), because the purpose is to create a sample of realizations that reproduces the original probability distribution sufficiently accurately. Realizations are typically generated using a pseudo random number generator, which is fairly easily accomplished for one-point (marginal) distributions, but becomes more tricky for multi-point (joint) distributions and for distributions of spatially dependent soil variables. Several algorithms are described in detail in later chapters.

For continuous soil properties, a histogram of a sufficiently large sample of realizations generated by the simulation will closely match the imposed probability distribution. This implies that extremes are not under-represented, as was the case in prediction. Thus, smoothing does not occur and the spatial pattern in the simulated maps more realistically represents the true spatial variation of the soil property. However, the price paid for this is that the difference between a single realization and the true soil map is greater than that between the prediction map and the true soil map. In fact, the variance of the difference for simulation is twice as large as that for prediction with the mean:

$$\begin{aligned} E[(\hat{Z}(\mathbf{s}) - \tilde{Z}(\mathbf{s}))^2] &= E[(\hat{Z}(\mathbf{s}) - \mu_{\hat{Z}}(\mathbf{s}) - (\tilde{Z}(\mathbf{s}) - \mu_{\hat{Z}}(\mathbf{s})))^2] \\ &= E[(\hat{Z}(\mathbf{s}) - \mu_{\hat{Z}}(\mathbf{s}))^2] + E[(\tilde{Z}(\mathbf{s}) - \mu_{\hat{Z}}(\mathbf{s}))^2] = 2V_{\hat{Z}}(\mathbf{s}) \end{aligned} \qquad (1.14)$$

where $\tilde{Z}(\mathbf{s})$ represents the simulated continuous soil property at $\mathbf{s}$. It has the same probability distribution as $\hat{Z}(\mathbf{s})$ and is statistically independent from it. It is therefore not wise to use a simulated reality as a prediction. However, simulation is useful for several other reasons. First, as noted above, the spatial pattern of a simulation reproduces the true spatial variation of the soil property. Second, presenting and comparing multiple simulated realities (such as in animations or simply a display of several realizations next to each other) is an attractive way to communicate uncertainty. Third, simulated realities are required for a Monte Carlo uncertainty propagation analysis, whereby it is analysed how uncertainty in the soil property propagates through a model (i.e. a soil acidification model, crop growth model or greenhouse gas emission model) that uses the soil property as input.

It is worthwhile to note that the average of a large number of realizations is equal to the prediction and that the sample variance computed from a large number of simulations will be equal to the prediction variance $V_{\hat{Z}}$. This is because for large sample sizes the sample mean and variance closely approximate the true mean $\mu_{\hat{Z}}$ and variance $V_{\hat{Z}}$.

Simulation from an uncertain categorical soil property means drawing realizations from a discrete probability distribution. Here, the probability of drawing a specific value is simply equal to the probability associated with that value. Thus, if soil type A has probability 0.30 at some location then it will be simulated in about 30 of 100 cases. Again, simulation is easy for marginal distributions but becomes more difficult for multivariate distributions or when the uncertain categorical soil property is spatially dependent. Simulated maps of categorical soil properties will typically be more noisy than the corresponding prediction maps. Unlike for prediction, categories with relatively small probabilities are not under-represented and will be simulated in proportion to their probability of occurrence. For example, if at each location soil type A has probability 0.30 and B has probability 0.70, then about 30 per cent of the area of each simulated map will be occupied by soil type A and about 30 per cent by type B.

1.5 Contents of this report

The remainder of this report is divided in three parts. The first part describes three specific approaches to derive the probability distribution of a spatially distributed quantitative soil property. The second part applies these approaches to spatially distributed qualitative soil properties. The three approaches are:

1. Kriging
2. Bayesian Maximum Entropy
3. Markov Random Fields

The common goal of all three approaches is to obtain a probability distribution of a spatially distributed soil property from available information, which may be observations at point locations, (e.g. measurements of the target soil property or correlated environmental variables, such as other soil properties, geological and hydrological properties), spatially exhaustive information (e.g. a soil type map, land use map or digital terrain model), or both. The probability distribution must reflect the knowledge gained from the available information and must realistically describe the remaining uncertainty about the target soil properties. Once the probability distributions are derived, prediction and simulation can be based on them using the approach presented in section (1.4), although in practice prediction and simulation are firmly integrated in the three approaches.

Part three illustrates the approaches with examples. All examples use the Netherlands as the study area and focus on soil properties that are crucial to predicting the soil status for biodiversity and biomass accumulation in semi-natural ecosystems, such as can be predicted with the SMART - SUMO - MOVE model chain (Schouwenberg et al., 2000; van Dobben et al., 2002b).

Conclusions on the appropriateness of the three approaches are presented in a final chapter.

It is useful to note that the only maps considered in this work are raster maps, which are defined at a suitable spatial resolution (trade-off between sufficient spatial detail and computational and storage requirements). Polygon maps must therefore be rasterized before they enter the analysis. Output cannot be produced as a polygon map because this would presume that the area of interest can be divided in a limited number of map units for which all locations have the same probability distribution, which is not realistic. Raster maps are also often referred to as continuous maps, as opposed to discontinuous or choropleth maps. However, in this report we will not use this term because it can be confused with maps depicting a continuous variable, i.e. a variable measured on interval- or ratio scale.

Part I

QUANTITATIVE SOIL MAPS

Chapter 2

Kriging methods

Kriging refers to a set of techniques in which the value of the target variable at a non-observed location $\mathbf{s}_0$ is predicted as a linear combination of the observed values in the neighbourhood of that location. The kriging predictor is the Best Linear Unbiased Predictor (BLUP), which means that it has the smallest mean squared error among all linear unbiased predictors. The kriging predictor can be formulated as a predictor from a generalized linear regression model (Christensen, 1991). Generalized in this case refers to the relaxation of the assumption of independent observations made in classical linear regression.

In kriging a continuous random variable $Z(\mathbf{s})$ is decomposed as the sum of a trend component $\mu(\mathbf{s})$ (model mean) and a residual component $R(\mathbf{s})$:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + R(\mathbf{s}) \tag{2.1}$$

The aim of kriging is to predict $Z(\mathbf{s}_0)$ at an unobserved location $\mathbf{s}_0$ from point observations and additional information, by *conditioning* $Z(\mathbf{s}_0)$ to the observations and additional information. The prediction will be referred to as $\hat{Z}(\mathbf{s}_0)$, which is identical to the $\hat{Z}(\mathbf{s}_0)$ introduced in chapter (1).

As stated above there are several kriging predictors, and these predictors differ with respect to the model for the trend component $\mu(\mathbf{s})$. Distinction can be made between kriging predictors assuming that the model mean is constant (over the area as a whole or within a neighbourhood of the prediction location) and kriging predictors in which the model mean varies in space. In practice, the choice between these two types is based on the available data and information. We distinguish between the following situations

1. observations of the target variable and covariable at points only
2. observations at points and map of covariable

In the first situation one has no information about the model mean at the prediction location $\mathbf{s}_0$. In this situation, the assumption is made that this mean is constant over the area of interest, or within a the neighbourhood of the prediction location. In simple kriging it is assumed that the model mean is known, whereas in ordinary kriging the model mean (possibly within neighbourhoods) is estimated from the sample. The consequence is that in simple kriging the uncertainty about the model is not incorporated in the calculated prediction-error variance, whereas it is in the ordinary kriging variance. If one has, besides the observations of the target variable, observations of a covariable, multivariate kriging predictors can be used, of which we treat the co-kriging predictor.

In the second situation, besides the observations at points, one has a map of a covariable. In this situation the map may be used to model the spatial variation of the trend component (model mean), and possibly the model variance. The appropriate kriging predictor in this case is the universal kriging predictor. The covariable can be categorical or continuous.

2.1 Observations at points only

2.1.1 Simple kriging

In simple kriging it is assumed that the Spatial Random Function (SRF) Z is second-order stationary, i.e the model mean $\mu(\mathbf{s})$ is constant, the variance is constant and finite, and the (auto)-covariance is not dependent on the locations but only on the separation vector $\mathbf{h}$. Moreover, it is assumed that the model mean is known without error. The value at a non-observed location is predicted as the model mean plus a linear combination of the deviations of the values at neighbouring locations from the model mean:

$$\hat{Z}(\mathbf{s}_0) = \mu + \sum_{\alpha=1}^{n} \lambda_\alpha [Z(\mathbf{s}_\alpha) - \mu] \tag{2.2}$$

where $\hat{Z}(\mathbf{s}_0)$ is the predicted value for the continuous random variable $Z(\mathbf{s}_0)$, λ_α is the weight assigned to datum $Z(\mathbf{s}_\alpha)$, and n is the number of data considered in the neighbourhood of $\mathbf{s}_0$. The weights λ_α should be chosen such that the prediction-error variance is minimized. The prediction-error variance equals

$$\sigma^2_{\mathrm{SK}} = C(0) + \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(\mathbf{h}_{\alpha\beta}) - 2 \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{h}_{\alpha\mathbf{0}}) \tag{2.3}$$

This variance can be minimized by equating the partial derivatives with respect to the λ_α's to zero:

$$\frac{\partial \sigma^2_{\mathrm{SK}}}{\partial \lambda_{\alpha,i}} = 2 \sum_{\beta=1}^{n} \lambda_\beta \, C(\mathbf{h}_{\alpha\beta}) - 2C(\mathbf{h}_{\alpha 0}) = 0 \qquad \alpha = 1, \ldots, n \; . \tag{2.4}$$

This results in the following matrix equation, referred to as the simple kriging system

$$\begin{bmatrix} C(\mathbf{h}_{11}) & \cdots & C(\mathbf{h}_{1n}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{h}_{n1}) & \cdots & C(\mathbf{h}_{nn}) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C(\mathbf{h}_{10}) \\ \vdots \\ C(\mathbf{h}_{n0}) \end{bmatrix} \tag{2.5}$$

For the simple kriging predictor the prediction-error variance, Eq. (2.3), can be simplified to (de Marsily, 1986, p. 290)

$$\sigma^2_{\mathrm{SK}} = C(0) - \sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha 0}(\mathbf{h}). \tag{2.6}$$

Until now it is assumed that the observations (measurements) are certain, errorless. If all the observations are uncertain, and the uncertainty (variance) is more or less equal for all observations, then we can simply estimate the experimental variogram for these uncertain data, and proceed as described above. The observation error is automatically accounted for in the fitted variogram (nugget > 0), and in the computed kriging variance. If some observations are more uncertain (have larger variance) than the others, then this additional uncertainty can be accounted for by adding this extra variance to the elements on the diagonal of the covariance-matrix associated with these observations (Delhomme, 1978).

2.1.2 Ordinary kriging

The simple kriging variance, Eq. (2.6), does not include uncertainty about the model mean, which is in many situations unrealistic and undesirable. If we want to incorporate this uncertainty, then the value of the continuous variable at a non-observed location can be predicted with the ordinary kriging predictor. The unknown model mean is filtered from the simple kriging predictor (Eq. (2.2)) by forcing the kriging weights to sum to 1:

$$\sum_{\alpha=1}^{n} \lambda_\alpha = 1. \tag{2.7}$$

This leads to the ordinary kriging predictor:

$$\hat{Z}(\mathbf{s}_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(\mathbf{s}_\alpha). \tag{2.8}$$

Minimizing the prediction-error variance, Eq. (2.3), under the unbiasedness constraint (Eq. (2.7)) leads to the following matrix equation (ordinary kriging system):

$$\begin{bmatrix} C(\mathbf{h}_{11}) & \cdots & C(\mathbf{h}_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C(\mathbf{h}_{n1}) & \cdots & C(\mathbf{h}_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{bmatrix} = \begin{bmatrix} C(\mathbf{h}_{10}) \\ \vdots \\ C(\mathbf{h}_{n0}) \\ 1 \end{bmatrix}, \tag{2.9}$$

where ν is a Lagrange parameter.

For the ordinary kriging predictor, the prediction-error variance (Eq. (2.3)) can be simplified to

$$\sigma^2_{\mathrm{OK}} = C(0) - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{h}_{\alpha\mathbf{0}}) - \nu. \tag{2.10}$$

This ordinary kriging system in terms of (auto-)covariances holds for second-order stationary SRF's, with finite variance C(0). For intrinsic SRF's (no finite variance) the ordinary kriging system must be written in terms of semivariances:

$$\begin{bmatrix} \gamma(\mathbf{h}_{11}) & \cdots & \gamma(\mathbf{h}_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\mathbf{h}_{n1}) & \cdots & \gamma(\mathbf{h}_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{bmatrix} = \begin{bmatrix} \gamma(\mathbf{h}_{10}) \\ \vdots \\ \gamma(\mathbf{h}_{n0}) \\ 1 \end{bmatrix}. \tag{2.11}$$

The prediction-error variance, Eq. (2.3), can then be simplified to

$$\sigma^2_{\mathrm{OK}} = \sum_{\alpha=1}^{n} \lambda_\alpha \gamma(\mathbf{h}_{\alpha\mathbf{0}}) + \nu. \tag{2.12}$$

Similar to simple kriging, observation errors can be accounted for by adding the observation-error variance to the diagonal of the covariance-matrix (Delhomme, 1978). In this way any mixture of hard and soft data can be processed.

2.1.3 Co-kriging

We now consider the case where one has, besides observations on the target variable Z_1, observations at locations on a quantitative covariable Z_2. If the means of the target variable and covariable are unknown, the appropriate predictor is the ordinary co-kriging predictor:

$$\hat{Z}_1(\mathbf{s}_0) = \sum_{\alpha=1}^{n} \lambda_{1\alpha} Z_1(\mathbf{s}_\alpha) + \sum_{\beta=1}^{m} \lambda_{2\beta} Z_2(\mathbf{s}_\beta), \tag{2.13}$$

with

$$\sum_{\alpha=1}^{n} \lambda_{1\alpha} = 1, \tag{2.14}$$

and

$$\sum_{\beta=1}^{m} \lambda_{2\beta} = 0. \tag{2.15}$$

Minimizing the prediction-error variance under the unbiasedness constraints (Eqs (2.14) and (2.15)) leads to the following matrix equation (ordinary co-kriging system):

$$\begin{bmatrix} C_{11}(\mathbf{h}_{11}) & \cdots & C_{11}(\mathbf{h}_{1n}) & C_{12}(\mathbf{h}_{11}) & \cdots & C_{12}(\mathbf{h}_{1m}) & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ C_{11}(\mathbf{h}_{n1}) & \cdots & C_{11}(\mathbf{h}_{nn}) & C_{12}(\mathbf{h}_{n1}) & \cdots & C_{12}(\mathbf{h}_{nm}) & 1 & 0 \\ C_{21}(\mathbf{h}_{11}) & \cdots & C_{21}(\mathbf{h}_{1n}) & C_{22}(\mathbf{h}_{11}) & \cdots & C_{22}(\mathbf{h}_{1m}) & 1 & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ C_{21}(\mathbf{h}_{m1}) & \cdots & C_{21}(\mathbf{h}_{mn}) & C_{22}(\mathbf{h}_{m1}) & \cdots & C_{22}(\mathbf{h}_{mm}) & 1 & 0 \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{11} \\ \vdots \\ \lambda_{1n} \\ \lambda_{21} \\ \vdots \\ \lambda_{2m} \\ \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} C_{11}(\mathbf{h}_{10}) \\ \vdots \\ C_{11}(\mathbf{h}_{n0}) \\ C_{21}(\mathbf{h}_{10}) \\ \vdots \\ C_{21}(\mathbf{h}_{m0}) \\ 1 \\ 0 \end{bmatrix}. \tag{2.16}$$

The minimum prediction-error variance thus obtained can be calculated with

$$\sigma^2_{\mathrm{OCK}} = C_{11}(0) - \sum_{\alpha=1}^{n} \lambda_{1\alpha} C_{11}(\mathbf{h}_{\alpha\mathbf{0}}) - \nu_1 - \sum_{\beta=1}^{m} \lambda_{2\beta} C_{21}(\mathbf{h}_{\beta\mathbf{0}}). \tag{2.17}$$

2.2 Observations at points and map of covariable

2.2.1 Universal kriging

If one has a map of a covariable or maps of several covariables, then the universal kriging predictor is appropriate. In this predictor the model mean μ in Eq. (2.1) is modelled as a linear function of the covariables:

$$\mu(\mathbf{s}) = \sum_{j=0}^{p} X_j(\mathbf{s})\beta_j \tag{2.18}$$

where $X_j(\mathbf{s})$ is the jth covariable (predictor or regressor) at location $\mathbf{s}$ and β_j is the regression coefficient associated with this covariable. The first covariable X_0 has the constant value 1, which makes that the kriging weights sum to 1 (see hereafter). The universal kriging predictor can then be obtained by substituting this trend function in Eq. (2.2):

$$\hat{Z}(\mathbf{s}_0) = \sum_{j=0}^{p} X_j(\mathbf{s}_0)\beta_j + \sum_{\alpha=1}^{n} \lambda_\alpha [Z(\mathbf{s}_\alpha) - \sum_{j=0}^{p} X_j(\mathbf{s}_\alpha)\beta_j]. \tag{2.19}$$

Similar to the ordinary kriging predictor, the unknown coefficients are filtered from the predictor by imposing the following constraints:

$$\sum_{\alpha=1}^{n} \lambda_\alpha x_j(\mathbf{s}_\alpha) = x_j(\mathbf{s}_0) \qquad j = 0, \cdots, p. \tag{2.20}$$

Minimisation of the prediction-error variance leads to the following matrix equation:

$$\begin{bmatrix} C(\mathbf{h}_{11}) & \cdots & C(\mathbf{h}_{1n}) & 1 & x_1(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{h}_{n1}) & \cdots & C(\mathbf{h}_{nn}) & 1 & x_1(\mathbf{s}_n) & \cdots & x_p(\mathbf{s}_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ x_1(\mathbf{s}_1) & \cdots & x_1(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \\ x_p(\mathbf{s}_1) & \cdots & x_p(\mathbf{s}_n) & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu_0 \\ \nu_1 \\ \vdots \\ \nu_p \end{bmatrix} = \begin{bmatrix} C(\mathbf{h}_{10}) \\ \vdots \\ C(\mathbf{h}_{n0}) \\ 1 \\ x_1(\mathbf{s}_0) \\ \vdots \\ x_p(\mathbf{s}_0) \end{bmatrix}. \tag{2.21}$$

The minimum prediction-error variance thus obtained equals

$$\sigma^2_{\mathrm{UK}} = C(0) - \sum_{\alpha=1}^{n} \lambda_\alpha C(\mathbf{h}_{\alpha\mathbf{0}}) - \sum_{k=1}^{p} \nu_p x_p(\mathbf{s}_0). \tag{2.22}$$

Note that the (auto-)covariance terms in Eqs (2.21) and (2.22) are the (auto-)covariances for the residuals, i.e. the observations minus the trend. Estimating this residual covariance function (variogram) is not trivial because, to estimate the residual covariance function (variogram), one must know the trend, and for estimating the trend one must know the residual covariance function (variogram). A proper way of estimating the trend and the covariance function (variogram) of the residuals simultaneously is Residual Maximum Likelihood (Lark and Webster, 2006). A simple alternative, leading to biased estimates of the variogram, is iterative Generalized Least Squares fitting of the trend. In the first iteration the trend is estimated by Ordinary Least Squares, i.e. one assumes that the residuals are uncorrelated. The estimated trend is filtered from the data, and the covariance function (variogram) of the residuals is calculated. This covariance function (variogram) is used to re-estimate the trend by Generalized Least Squares, and to re-calculate the residuals and its variogram. This is repeated until the estimated trend coefficients do not change anymore.

The covariable can also be categorical. In this case the categorical variable is replaced by n_c indicator-variables (with n_c the number of categories) that are used as regressors, see chapter (5). Note that in universal kriging it is assumed that the residuals for all categories have common mean, variance and (auto-)covariance function. If this is too strong an assumption, residuals can be standardized before estimating the variogram and spatial interpolation of the residuals. This necessary leads to iterative GLS fitting of the trend (Finke et al., 2004).

2.2.2 Stratified kriging

In stratified kriging the study area is partitioned into a finite number of subareas, referred to as strata. Besides this, the data are partitioned into the same groups. After stratification, spatial interpolation is performed for each stratum separately, i.e. only observations within the stratum of the interpolation location are used (Stein et al., 1988; M. Voltz, 1990; Brus et al., 1996). The rationale for this is that, contrary to universal kriging with a categorical covariable, in stratified kriging the target variables $Z(\mathbf{s})$ at locations in different strata are assumed independent. To interpolate or simulate values within a given stratum, simple or ordinary kriging can be used. Simple kriging requires knowledge of the stratum means (model means per stratum), whereas ordinary kriging does not.

2.3 Simulation

Various techniques can be used for stochastic spatial simulation, perhaps the most attractive one being the **sequential Gaussian simulation** algorithm (Goovaerts, 1997). This method works will all kriging methods discussed so far but it does assume that the variable of interest is normally distributed. Briefly, this method works as follows.

Each location (grid cell) of the spatial domain is visited in a random sequence (in fact, random visiting is not strictly required). At each location, the conditional probability distribution of the variable is computed. For the first location, this is simply the normal distribution with a user-specified model mean and standard deviation. A value from this probability distribution is drawn using an appropriate pseudo-random number generator and assigned to the location. At the second location, the conditional probability distribution is computed by conditioning the variable at the location to the value that was sampled at the first location. This is usually done using simple kriging.

Simple kriging, like all kriging methods, not only produces an estimate of the attribute value but also quantifies the uncertainty attached to it, by means of a kriging standard deviation. The conditional probability distribution at the second location has as mean $\hat{Z}$ (Eq. (2.2)), and as standard deviation the square root of the simple kriging variance, σ^2_{SK} (Eq. (2.6)). From this distribution, a value is drawn. At the third location, the conditional probability distribution is calculated again, now using the two previous locations as conditioning data in simple kriging. This process is repeated until values for all locations have been drawn.

Note that in an early stage, when values have been drawn at only few locations, it is possible that the next random location is outside the range of influence of the other locations. In that case, the simple kriging estimate and standard deviation equal the pre-defined mean and standard deviation. At a later stage, previously drawn locations are more likely to be within the range and will influence the mean and standard deviation.

The sequential simulation method can also be used in cases where the variable is a priori known at some locations. This is referred to as conditional simulation. The prior data are used as conditioning data from the beginning. When one has prior data, one may use ordinary kriging, or even universal kriging, to estimate the mean and variance of the pdf at simulation nodes.

There is a choice on freely available software-packages for simulation with kriging techniques: GSLIB (Deutsch and Journel, 1992), GSTAT (Pebesma and G.Wesseling, 1998), and BMELIB contains MATLAB scripts for unconditional (simuseq.m) and conditional simulation (simuseqcond.m) (Christakos et al., 2002).

Chapter 3

Bayesian Maximum Entropy

Bayesian Maximum Entropy is a relatively new approach for spatial mapping that allows the user to incorporate a wide variety of data sources of various quality on a sound theoretical basis (Christakos, 1990, 2000; Christakos et al., 2002; Bogaert and D'Or, 2002). Broadly speaking, BME consists of two steps. In the first step the unconditional multi-point pdf at the prediction location and the neighbouring observation locations is computed. This is done by maximizing the entropy of the distribution under several constraints. In the second step this unconditional multi-point pdf is conditioned on the observations at the neighbouring locations.

The central concept in the BME approach is entropy. The entropy of a discrete random variable is defined as:

$$H_C = \sum_{i=1}^{n_c} \pi_C(c_i, \mathbf{s}) \log \frac{1}{\pi_C(c_i, \mathbf{s})} = -\sum_{i=1}^{n_c} \pi_C(c_i, \mathbf{s}) \log \pi_C(c_i, \mathbf{s}), \tag{3.1}$$

where $\pi_C(c_i, \mathbf{s})$ is the probability that the random variable C at location $\mathbf{s}$ takes the value c_i and n_c is the number of categories.

For continuous random variables, summation is replaced by integration, and the $\pi_C(c_i, \mathbf{s})$'s by the pdf values $f_Z(z, \mathbf{s})$:

$$H_Z = -\int\limits_{D_Z} f_Z(z, \mathbf{s}) \log f_Z(z, \mathbf{s}) \mathrm{d}z, \tag{3.2}$$

where D_Z is the domain of the random variable Z.

For discrete random variables the minimum value for the entropy is 0, which occurs when one of the possible outcomes has probability 1, and the maximum is $\log n_c$, occurring when all outcomes have equal probability. In other words, the maximum entropy distribution for discrete random variables is the uniform distribution. This shows that entropy is a measure of uncertainty: the larger the uncertainty, the larger the entropy. For continuous random variables the shape of the distribution with maximum entropy depends on the constraints, see Table 3.1.

Any continuous distribution must fulfil the normalization constraint $\int f_Z(z, \mathbf{s})\mathrm{d}z = 1$. If we add the constraint that z is bounded on both sides ($z \in [\alpha; \beta]$), then the maximum entropy distribution is the uniform distribution between α and β

Table 3.1: Constraints and resulting probability distribution function with maximum entropy. After D'Or (2003)

Constraints	Resulting pdf	Entropy expression
$\int f_Z(z, \mathbf{s})\mathrm{d}z = 1$ $z \in [\alpha; \beta]$	$Z \sim Uni(\alpha, \beta)$	$H_Z = -\log(\beta - \alpha)$
$\int f_Z(z, \mathbf{s})\mathrm{d}z = 1$ $z \in [0; +\infty]$ $E[Z] = a$	$Z \sim Exp(1/a)$	$H_Z = 1 + \log(a)$
$\int f_Z(z, \mathbf{s})\mathrm{d}z = 1$ $E[Z] = 0$ $Var[Z] = \sigma^2$	$Z \sim N(0, \sigma^2)$	$H_Z = (1/2)\log(2\pi e\sigma^2)$

$(f_Z(z, \mathbf{s}) = 1/(\beta - \alpha))$. In this case the maximum entropy equals $H_Z = -\ln(\beta - \alpha)$. If Z is non-negative ($z \in [0; +\infty]$) and the expectation of the distribution is finite and known, then the maximum entropy distribution will be an exponential distribution given by $f_Z(z, \mathbf{s}) = (1/a)e^{(-z/a)}$, where a is the mean. The entropy is then given by $H_Z = 1 + \ln(a)$. Finally, if the mean is zero and the variance is equal to σ^2, then the maximum entropy distribution takes a Gaussian shape, i.e. $Z \sim N(0, \sigma^2)$. In this case, the entropy is given by $H_Z = (1/2)\ln(2\pi e\sigma^2)$.

Choosing the pdf with maximum entropy makes sense because we do not want to go beyond the data. For discrete pdfs, we prefer pdfs with (nearly) equal probabilities for all categories over pdfs with strongly different probabilities, unless there is strong evidence from the data that some categories are more likely. In other words, we do not claim to be more certain than we are. As will be shown in section (3.1.1), the entropy is maximized under several constraints, which makes that in practice the estimated discrete pdf deviates from the uniform distribution.

In BME two important steps can be distinguished:

1. Computing the unconditional multi-point pdf (prior step)
2. Conditioning the multi-point pdf (posterior step)

Christakos (2000) distinguishes an intermediate step, the meta-prior step, in which the site-specific information is analyzed, evaluated and translated in mathematical relations so that it can be used in the posterior step to condition the unconditional multi-point pdf. In this report we will focus on the prior and posterior steps. We assume that the conditioning data are there, and that the quality of the data is quantified.

In the remaining part of this chapter we will elaborate on the prior and posterior step for continuous variables. In chapter (6) we will do this for categorical variables.

With respect to the availability of data and information used in conditioning the unconditional multi-point pdf, we distinguish between the following situations:

1. observations at points only

2. observations at points and map of target variable

3.1 Observations at points only

We first consider the situation where one has observations of the target variable at a finite number of locations only. A map of the target variable or of a variable related to the target variable (covariable) is lacking.

3.1.1 Computing the unconditional multi-point pdf

Eq. (3.2) defines the entropy of a one-point (univariate) pdf, i.e. the pdf of a random variable at a single location. This definition can be extended to the entropy for the multi-point (joint) pdf:

$$\begin{gathered} H(f_{Z\ldots Z}(z_0,\ldots,z_n,\mathbf{s}_0,\ldots,\mathbf{s}_n)) = \\ -\int\limits_{D_Z} f_{Z\ldots Z}(z_0,\ldots,z_n,\mathbf{s}_0,\ldots,\mathbf{s}_n)\log f_{ZZ}(z_0,\ldots,z_n,\mathbf{s}_0,\ldots,\mathbf{s}_n)\mathrm{d}\mathbf{z}, \end{gathered} \tag{3.3}$$

$\int$ denotes a multiple integration with the number of dimensions equal to the number of sample points + 1 (1 for the prediction location $\mathbf{s}_0$).

For convenience, in the following of this subsection we will write $f_{Z\ldots Z}(z_0,\ldots,z_n,\mathbf{s}_0,\ldots,\mathbf{s}_n)$ as $f(\mathbf{z})$, where $\mathbf{z}$ denotes the unknown value of the target variable at the prediction location plus the observations at the sample locations: $\mathbf{z} = (z(\mathbf{s}_0), z(\mathbf{s}_1), \cdots, z(\mathbf{s}_n))$, and $z(\mathbf{s}_\alpha)$ $(Z(\mathbf{s}_\alpha))$ is written as z_α (Z_α).

The entropy of this multi-point pdf is maximized under several constraints. Although there are no problems with constraints on statistical moments of higher order, in practice the following constraints are enforced:

1. Normalization constraint
2. Constraints on the mean
3. Constraints on the variance
4. Constraints on the (auto-)covariance

The normalization constraint $\int\limits_{D_Z} f(\mathbf{z})\mathrm{d}z = 1$ enforces the multi-point pdf to "sum" to 1. If we have n sample points, then we have $n+1$ constraints on the mean, one for each sample location and one for the prediction location. For the variance we have again $n+1$ constraints. For the (auto-)covariance we have $(n+1)n/2$ constraints, as many constraints as pairs that can be formed with the $n+1$ locations. The mean, variance and (auto-)covariances are unknown and must be estimated from the observations.

For further development, it is convenient to write the constraints as

$$E[g_i] = \int g_i(\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z} \qquad i = 0,\ldots,N_c, \tag{3.4}$$

with $N_c + 1$ the total number of constraints, and the $g_i(\mathbf{z})$ known functions of $\mathbf{z}$. $g_0(\mathbf{z}) = 1$ leading to the normalization constraint. For the constraints on the mean $g_i(z_\alpha) = Z_\alpha$, for the constraints on the variances $g_i(z_\alpha) = [Z_\alpha - \mu_\alpha]^2$, and for the (auto-)covariances $g_i(z_\alpha, z_\beta) = [Z_\alpha - \mu_\alpha][Z_\beta - \mu_\beta]$.

Maximizing the entropy under these constraints can be done with the Lagrange multiplier method:

$$\begin{aligned} L[f(z)] = &- \int f(\mathbf{z}) \log f(\mathbf{z}) \mathrm{d}\mathbf{z} \\ &- \sum_{i=0}^{N_c} \mu_i [\int g_i(\mathbf{z}) f(\mathbf{z}) \mathrm{d}\mathbf{z} - E[g_i(\mathbf{z})]]. \end{aligned} \tag{3.5}$$

Setting the partial derivatives to zero and solving the system of equations with respect to the μ_i yields the maximum entropy solution for the unconditional multi-point pdf:

$$f(\mathbf{z}) = \frac{1}{A} \exp(\sum_{i=1}^{N_c} \mu_i g_i(\mathbf{z})), \tag{3.6}$$

where A enforces the normalization constraint:

$$A = \int \exp(\sum_{i=1}^{N_c} \mu_i g_i(\mathbf{z})) d\mathbf{z}. \tag{3.7}$$

With constraints on the mean, variance and (auto-)covariance only, the unconditional multi-point pdf is multivariate Gaussian (Rao, 1973, p. 162-163 and p. 532-533).

3.1.2 Conditioning the multi-point pdf

In the posterior step the unconditional multi-point pdf is conditioned on the observations at the neighbouring sample locations. This can be done by using the definition of a conditional distribution:

$$f(z_0 \mid z_1, \cdots, z_n) = \frac{f(z_0, z_1, \cdots, z_n)}{f(z_1, \cdots, z_n)}. \tag{3.8}$$

The multi-point pdf in the denominator is the marginal of the unconditional multi-point pdf in the numerator and therefore can simply be obtained by integration.

If the observations of the target variable are uncertain at part of the locations, it is important to account for these differences in observation accuracy. In this case the sample data must be split in a subsample of hard (certain) data $z_1, \cdots, z_n$ denoted as $\mathbf{z}_\mathrm{h}$, and a subsample of soft data $z_{n+1}, \cdots, z_{n+m}$ denoted as $\mathbf{z}_\mathrm{s}$. If one has hard conditioning data only, BME is equivalent to simple kriging (Christakos, 2000; Christakos et al., 2002).

The uncertainty of soft observations is often expressed by specifying a lower and upper bound for each soft observation (interval data). In some cases a complete pdf is specified, denoted hereafter as $f_\mathrm{O}(\mathbf{z}_\mathrm{s})$ (probabilistic data). The subscript O of this pdf refers to the source of uncertainty, the observation method.

For probabilistic data the conditional pdf can be obtained by multiplying the unconditional $(n+m+1)$-point probability densities and the $(n+m)$-point probability densities by the m-point probability densities associated with the soft data, followed by multiple integration (D'Or, 2003, p. 34). In formula:

$$f(z_0 \mid \mathbf{z}_{\mathrm{h}}, \mathbf{z}_{\mathrm{s}})) = \frac{\int f(z_0,\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) f_{\mathrm{O}}(\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}}{\int f(\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) f_{\mathrm{O}}(\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}} = \frac{\int f(z_{\mathrm{s}}|z_0,\mathbf{z}_{\mathrm{h}}) f_{\mathrm{O}}(\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}} f(z_0,\mathbf{z}_{\mathrm{h}})}{\int f(\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) f_{\mathrm{O}}(\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}}, \tag{3.9}$$

with

$$f_{\mathrm{O}}(\mathbf{z}_{\mathrm{s}}) = \Pi_{i=1}^{m} f_{\mathrm{O}}(z_{\mathrm{s},i}). \tag{3.10}$$

For interval data the conditional pdf Eq. (3.9) can be simplified to

$$f(z_0) = \frac{\int_{\mathbf{l}}^{\mathbf{u}} f(z_0,\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}}{\int_{\mathbf{l}}^{\mathbf{u}} f(\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}} = \frac{\int_{\mathbf{l}}^{\mathbf{u}} f(z_{\mathrm{s}}|z_0,\mathbf{z}_{\mathrm{h}}) \mathrm{d}\mathbf{z}_{\mathrm{s}} f(z_0,\mathbf{z}_{\mathrm{h}})}{\int_{\mathbf{l}}^{\mathbf{u}} f(\mathbf{z}_{\mathrm{h}},\mathbf{z}_{\mathrm{s}}) \mathrm{d}\mathbf{z}_{\mathrm{s}}}. \tag{3.11}$$

In words, the conditional pdf is obtained by multiple integration of the unconditional $(n+m+1)$-point pdf (numerator) and of the $(n+m)$-point pdf (nominator) with as many dimensions as there are soft data in the neighbourhood.

Unlike ordinary kriging, BME requires that the unknown local mean $\mu(\mathbf{s}_0)$ be estimated explicitly from the data. This is done by Generalized Least Squares, which works fine if the soft observations follow a normal distribution. Orton and Lark (2007) show that for interval data, especially censored observations, GLS-estimation of the local mean can be biased, and may result in suboptimal BME predictions. They propose to estimate the local mean by a maximum likelihood method.

3.2 Observations at points and map of target variable

In this case one has, besides hard observations of the target variable at locations, a map of the target variable. This map can be used to derive a soft observation at any location. In practice, a soft observation is derived at the estimation (simulation) location only. This soft observation at the estimation (simulation) location can be used to condition the prior distribution in the same way as soft observations in the neighbourhood. The procedure is analogous to that described in section (3.1.2).

3.3 Simulation

Using the Bayesian Maximum Entropy approach for unconditional simulation and for conditional simulation with hard conditioning observations only, is equivalent to sequential Gaussian simulation, see section (2.3). This is because the conditional distribution obtained with BME in these cases are Gaussian.

Chapter 4

Markov Random Fields

The goal of the Markov Random Field (MRF) approach is similar to that of Kriging and Bayesian Maximum Entropy. It aims to define and calibrate statistical models of quantitative and qualitative spatially distributed variables, and use these models for spatial prediction and simulation. The difference between the three approaches is that the models used are different. In the case of MRF, the models are defined in terms of *conditional* probability distributions. The modeller defines or calibrates the probability of occurrence of the variable of interest at some location, conditional to the value of the variable at neighbouring locations. When a grid representation of the geographic domain is used, as assumed throughout this work, the neigbourhood is typically taken as the four or eight cells surrounding the cell of interest. However, larger neighbourhoods may also be used. Critical assumption of the MRF approach is that the local neighbourhood contains all information necessary to characterize the probability distribution of the variable in the centre cell (i.e. information from outside the local neighbourhood has no longer added value). This will be explained in more detail below, but it is important here to underline that while this key assumption may greatly facilitate estimation, prediction and simulation, one should always question the practical validity of the assumption in real-world cases.

The Markov Random Field approach has not yet been widely used in soil science and the environmental sciences in general. Some recent examples are Norberg et al. (2002); Wu et al. (2004); Kasetkasem et al. (2005); Hartman (2006). Compared to Kriging and to a lesser extent Bayesian Maximum Entropy, application of the Markov Random Field approach to the problems addressed in this work is still in the development phase with many important problems yet to be resolved. However, the approach has many attractive properties and is potentially very valuable. Recently, it enjoys much interest from various disciplines, ranging from land use change modelling to image analysis. The method has a sound statistical basis, starting with the fundamental work of Besag (1974). The method seems particularly attractive for modelling qualitative soil properties, because unlike with Kriging, the extension from the quantitative to the qualitative domain does not greatly complicate matters or introduce major heuristic choices. However, for completeness sake and because the method may be interesting for quantitative appliations as well, in this chapter we first describe the use of MRF to model, predict and simulate quantitative soil properties.

4.1 The Markov property

Let us consider the SRF Z again, which represents an uncertain quantitative soil property and which was introduced in chapter (2). The probability distribution of Z may be thought of as a prior distribution of the continuous soil property that has not yet exploited the information contained in the point observations and any other relevant information (such as spatially exhaustive covariates). Conditioning of Z to the point observations and ancillary information yields the SRF $\hat{Z}$, which has a narrower conditional distribution that conveys that uncertainty may be reduced by incorporating the ancillary information. Before discussing how conditioning is done in the MRF case, let us first characterize Z.

Because Z is continuously distributed, it has a multi-point probability density:

$$f_Z(z_1, \ldots, z_n, \mathbf{s}_1, \ldots, \mathbf{s}_n). \tag{4.1}$$

If we discretize the geographic domain and assume that all locations are on a rectangular grid, then n refers in this context to the total number of grid cells in the domain. Using the definition of conditional probability and dropping the explicit naming of the locations, we can now define the conditional probability of Z at some location $\mathbf{s}_k$ given the value of Z at all other locations as

$$f_Z(z_k | Z_i = z_i, i = 1, \ldots, n, i \neq k) = \frac{f_Z(z_i, i = 1, \ldots, n)}{f_Z(z_i, i = 1, \ldots, n, i \neq k)}. \tag{4.2}$$

This shows how the conditional distribution can be derived from the full distribution. Brooks lemma (Banerjee et al., 2004, section 3.2) can be used to do the reverse and derive the full distribution from the conditional. This is more difficult though, and not all hypothesized conditional distributions render a unique and valid full probability distribution. The *Markov property* now states that the conditional distribution of Z_k satisfies:

$$f_Z(z_k | Z_i = z_i, i = 1, \ldots, n, i \neq k) = f_Z(z_k | Z_i = z_i, \mathbf{s}_i \in \delta\mathbf{s}_k), \tag{4.3}$$

where $\delta\mathbf{s}_k$ is a set containing the locations that are in the neighbourhood of $\mathbf{s}_k$ (Figure 4.1). The size of the neighbourhood may be chosen as preferred. A larger neighbourhood produces a more flexible and realistic model but is more complex. In practice, neigbourhoods are rarely larger than the four immediate neighbours in the grid (i.e. the *von Neumann* neighbourhood) or extends these with the diagonal neighbours to eight neighbouring cells (i.e. the *Moore* neighbourhood). Equation (4.3) states that given the value of Z in the neighbourhood, Z_k does not depend on the value of Z at locations outside the neighbourhood. The equation also shows the great similarity between Markov Random Fields and stochastic cellular automata (Baltzer et al., 1998; Baltzer, 2000).

We can now summarize the rationale of the MRF approach. The idea is to define the full multi-point probability of Z through the conditional distribution Eq. (4.2), which is greatly simplified such that it can be estimated in practice from available information, by invoking the Markov property. Conditioning to observations is very easy as we will see in sections (4.3) and (4.4).

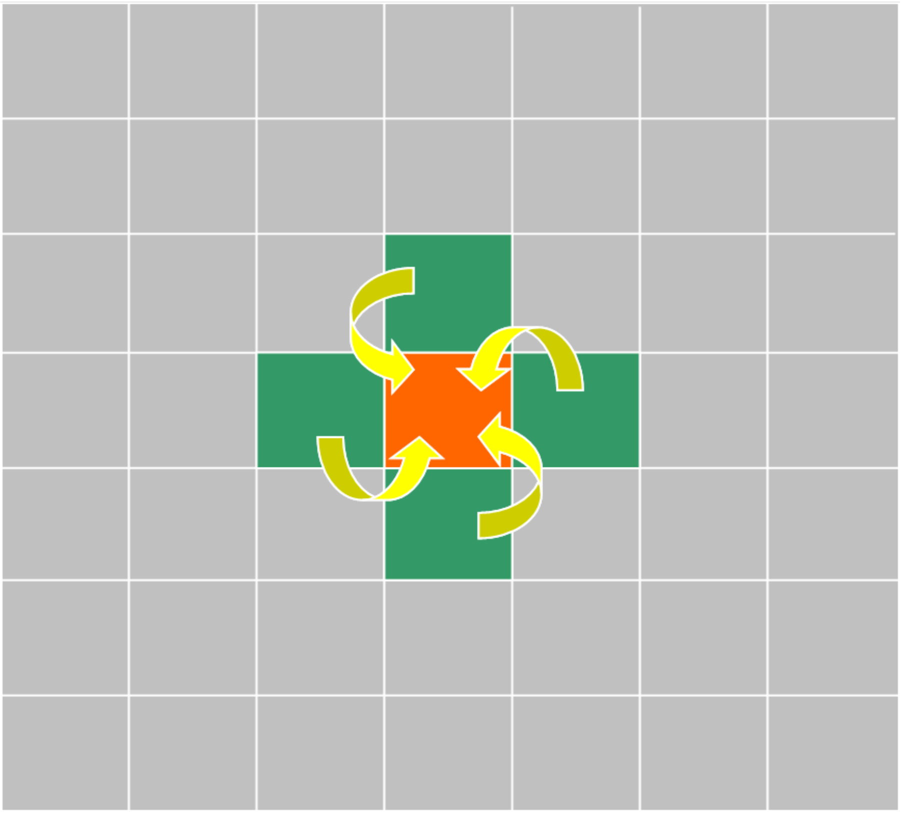

Figure 4.1: The Markov property states that the value at a location only depends directly on the values in its neighbourhood. In this example the neighbourhood consist of the four immediate neighbours of the centre cell.

4.2 Estimation of the conditional density

To estimate the shape and parameters of the conditional density, we might take a pragmatic approach whereby we assume that the variable is normally distributed with mean and variance equal to that of the sample of neighbourhood values:

$$f_Z(z_k|Z_i = z_i, \mathbf{s}_i \in \delta\mathbf{s}_k) \sim N(\bar{z}, s^2), \tag{4.4}$$

where $\bar{z}$ and s^2 are the sample mean and sample variance of the neighbouring values z_i, respectively. A better alternative, however, is to parametrize the conditional density and estimate its parameters with additional information such as contained in observations of the target variable across the entire study area. For example, one might assume that the conditional distribution is Gaussian with a mean that is a weighted linear combination of the neighbouring observations and a constant variance that is not affected by the neighbouring values (see also Banerjee et al. (2004), Eq. (3.12)):

$$f_Z(z_k|Z_i = z_i, \mathbf{s}_i \in \delta\mathbf{s}_k) \sim N(\sum_i b_i y_i, \tau^2). \tag{4.5}$$

In practice, one might achieve the calibration (i.e. the estimation of the parameters of the distribution) by generating possible realities of Z using the method discussed in the next section, computing an experimental variogram from these realities, comparing it with that of the observations, and adjusting the parameters of the conditional density such that a closer match is obtained. Clever calibration algorithms such as simulated annealing (Van Groenigen et al., 1999; Brus and Heuvelink, 2007) may be needed to speed up the computationally demanding calibration process.

4.3 Simulation

Unlike for Kriging and Bayesian Maximum Entropy, for the MRF approach we first discuss simulation and next prediction. This is because with MRF prediction is done by using the results of the simulation.

Stochastic simulation of Z under the MRF approach is done using *Markov Chain Monte Carlo (MCMC) methods* (Banerjee et al., 2004). These methods have recently been introduced to statistical computing, mainly in the context of Bayesian statistics, and have in fact revolutionary changed statistical computing. The MCMC methods are very flexible and have made it possible to numerically solve problems that cannot be solved using analytical methods and were therefore practically unresolvable until the introduction of these methods. It is beyond the scope of this work to explain and give an overview of MCMC methods. We will only briefly explain one example from the multitude of techniques, namely the MCMC technique that we will use to simulate Z. This is the so-called *Gibbs sampler*, which is remarkably simple and generic. It is explained in more detail by Casella and George (1992).

Using the Gibbs sampler we generate a realization of Z at the n grid locations $\mathbf{s}_i$ as follows:

1. Start with an arbitrary initial soil map z;
2. Visit an arbitrary grid location $\mathbf{s}_i$, compute the conditional distribution Eq. (4.3) at $\mathbf{s}_i$ and draw a random number from the conditional distribution using a pseudo-random number generator;
3. Replace the current value of $z(\mathbf{s}_i)$ by the number drawn and go to a next grid location not yet visited;
4. Repeat the simulation procedure for the new grid cell and continue with selecting grid cells and replacing values until all grid cells have been visited;
5. Repeat the grid simulation and replacement procedure many times.

Provided the number of repetitions is large enough, the final map will be a realization from the joint probability distribution Eq. (4.1). The procedure can be repeated to generate multiple realizations of Z, but alternatively one can also use the previously simulated map as a starting point for the next simulation, effectively running the simulation procedure only once for a longer time, periodically using intermediate maps as independent realizations from Z. To ensure that the realizations are indeed independent, one must allow a sufficiently large number of runs in between subsequent samples (i.e. the so-called *thinning ratio* must be sufficiently large). Moreover, the initial number of runs that must be executed before the first realization is taken (the so-called *burn-in period*) must be sufficiently large to guarantee that the first realization drawn is not affected by the initial soil map and is a realization from the distribution of Z.

Although the procedure is remarkably simple, the downside is that it is computationally demanding, also for the many modifications to the basic algorithm that have been proposed to speed up the computations. Another difficulty is finding the right values for the burn-in period and thinning ratio. This particular problem has been thoroughly investigated in the literature, but results are case-specific to some extent.

A more fundamental problem is that the Gibbs sampler will produce realizations even when the conditional distributions are ill-specified and do not have a proper corresponding full distribution. Thus it is imperative that a sound check is made that the conditional distributions used yield a valid full distribution. The statistical literature provides valuable entries to this problem (Griffith and Layne, 1999; Cressie, 1993; Banerjee et al., 2004).

4.3.1 Conditioning to point observations

We can also use the method described in the previous section to sample from the conditional distribution $\hat{Z}(\mathbf{s}_0 = Z(\mathbf{s}_0|Z(\mathbf{s}_i) = z(\mathbf{s}_i), i = i, \ldots, n)$, where n represents the number of observations and $\mathbf{s}_0$ is an unobserved location. The objective is to condition the simulations of Z to the observations. With the Gibbs sampler, this is extremely simple. One runs the sampling procedure described above with fixed values at the observation locations. Thus, each time a grid location is visited that is a sampling location, the value of Z at the location is not sampled from the conditional probability distribution but the observed value is used instead (note that this assumes that observation locations coincide with grid locations, which may be a problem when using coarse grids).

4.3.2 Conditioning to maps of covariables

Conditioning to maps of covariables is more difficult with MRF because the covariables do not represent true values of Z at some locations but provide a spatially exhaustive 'hint' about Z. In this case we need to adjust the probability model of Z to the ancillary information. The logical approach is then to let the conditional density Eq. (4.5) depend on the ancillary information as well. For instance, the conditional probability density of clay content of the topsoil would not only depend on the clay content at neigbouring cells but also on the soil type at the centre cell. The difficulty here is that this means an extension of the probability model with additional parameters that somehow need to be derived from the data. Also, proving that the extended conditional model yields a proper full model may be more difficult.

4.4 Prediction

Recall from section (1.4) that prediction usually means that we compute the expected value of $\hat{Z}$ at all locations in the geographical domain. In (linear) Kriging, computation of the conditional probability distribution of Z is solved by setting $\hat{Z}$ equal to a linear combination of the $Z(\mathbf{s}_i)$, so that its expected value can directly be computed once the values of the $Z(\mathbf{s}_i)$ are known. This is not possible with MRF because we have no simple explicit equation expressing $\hat{Z}$ as a linear combination of the observations. Rather, we have a numerical algorithm that allows us to simulate from $\hat{Z}$ as discussed in the previous section. Prediction with MRF is therefore done simply by averaging a large number of simulations.

Prediction with MRF makes use of the fact that the frequency distribution of a sufficiently large number of realizations from a probability distribution approximates the probability distribution. The approximation error vanishes when the number of realizations gets large. Properties of the frequency distribution, such as the mean and variance, also approximate the corresponding properties of the probability distribution. Thus we can compute the mean (or other properties) of $\hat{Z}$ simply by generating a sufficiently large sample and computing the corresponding sample statistic (i.e. the sample mean).

This is easy but we must be aware that the sample statistic only approximates the true mean. We can approximate it arbitrary well, but only at the expense of increasing computation costs. Also, we do not have an analytical expression for the mean and other properties of $\hat{Z}$, which means that we cannot easily generalize the results obtained to other situations (e.g. minor changes in the model parameters). Generalization is only possible by repeating the entire procedure with the new parameter settings.

Part II

QUALITATIVE SOIL MAPS

Chapter 5

Kriging methods

When adopting a kriging approach for simulation of realizations of a categorical Spatial Random Function (SRF), the single categorical random field is replaced by n_c indicator random fields. For this transformation the observations on the categorical variable are coded as indicators:

$$\delta_i(\mathbf{s}_\alpha) = \delta_{i_\alpha} = \begin{cases} 1 \text{ if } C(\mathbf{s}_\alpha) = c_i \\ 0 \text{ if } C(\mathbf{s}_\alpha) \neq c_i \end{cases} \tag{5.1}$$

For ordinal categorical variables this coding (Eq. (5.1)) can be replaced by a cumulative coding:

$$\delta_{i_\alpha} = \begin{cases} 1 \text{ if } C(\mathbf{s}_\alpha) \leq c_i \\ 0 \text{ if } C(\mathbf{s}_\alpha) > c_i \end{cases} \tag{5.2}$$

If we have n_c categories each observation is transformed into an n_c-vector with a 1 at the entry for the observed category, and 0's at the remaining entries. This gives n_c random fields of indicators.

The expectation of the indicator random variable equals the unconditional probability of occurrence of category c_i at location $\mathbf{s}_\alpha$ (Eq. 1.4):

$$E[\delta_{i_\alpha}] = \pi_{i_\alpha}. \tag{5.3}$$

The expectation of the product of the indicator random variables δ_{i_α} and δ_{j_β} equals the multi-point probability of occurrence of categories c_i and c_j at locations $\mathbf{s}_\alpha$ and $\mathbf{s}_\beta$, respectively:

$$E[\delta_{i_\alpha}\delta_{j_\beta}] = \pi_{ij}(\mathbf{h}_{\alpha\beta}) = P(C(\mathbf{s}_\alpha) = c_i, C(\mathbf{s}_\beta) = c_j). \tag{5.4}$$

Similar to the SRF of a continuous random variable, the spatial structure of the SRF of an indicator random variable is characterized by the (auto)covariance function or the variogram. Note that we have n_c indicator random fields, one for each category. The spatial interaction between two indicator random fields can be characterized by the indicator cross-covariance function or the indicator cross-variogram. The indicator (cross-)covariance function is defined as:

$$\begin{array}{rcl} C_{\delta_{i_\alpha},\delta_{j_\beta}}(\mathbf{h}_{\alpha\beta}) & \equiv & E[\delta_{i_\alpha}\delta_{j_\beta}] - E[\delta_{i_\alpha}]E[\delta_{j_\beta}] \\ & = & \pi_{ij}(\mathbf{h}_{\alpha\beta}) - \pi_{i_\alpha}\pi_{j_\beta} \\ & = & \pi_{ij}(\mathbf{h}_{\alpha\beta}) - \pi_i\pi_j \end{array} . \tag{5.5}$$

The estimated covariance function (experimental covariance function) can be obtained by substituting the estimated bivariate and univariate probabilities in Eq. (5.5):

$$C_{\delta_i,\delta_j}(\mathbf{h}) = p_{ij}(\mathbf{h}) - p_i \cdot p_j, \tag{5.6}$$

with $p_{ij}(\mathbf{h})$, the estimated bivariate probability, given by:

$$p_{ij}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{\alpha=1}^{N_{\mathbf{h}}} \delta_{i_\alpha} \cdot \delta_{j_{\alpha+\mathbf{h}}}, \tag{5.7}$$

with $N_{\mathbf{h}}$ denoting the number of pairs of points in the distance class corresponding to lag $\mathbf{h}$, and p_i and p_j the estimated univariate probabilities, computed with:

$$p_i = \frac{1}{N} \sum_{\alpha=1}^{N} \delta_{i_\alpha}. \tag{5.8}$$

After one has estimated the experimental covariance function, a model must be selected and calibrated. To ensure positive-definiteness a permissible model must be selected. In case of co-kriging this can be ensured by fitting a linear model of coregionalization, see section (5.1.3).

With indicator kriging it is neither guaranteed that the probabilities are in the interval [0,1], nor that the sum of the probabilities over all categories equals 1. Perhaps this problem can be tackled using compositional kriging (Walvoort and de Gruijter, 2001)

With respect to the availability of data and information, we distinguish between the following situations

1. observations at points only;
2. observations at points and map of covariable.

5.1 Observations at points only

We first consider the situation where one has no local information about the (unconditional) probability of occurrence of the categories. The only assumption made is that these probabilities are constant over the area of interest (global kriging), or within a neighbourhood of the prediction location (local kriging).

5.1.1 Simple indicator kriging

Simple indicator kriging assumes that the (unconditional) probabilities of occurrences of the categories are constant over the area of interest. Moreover, it is assumed that these probabilities are known without error. The conditional probability, conditioned on the indicators at locations in the neighbourhood, $\{\delta_{i_\alpha}\}$, for a given category c_i is predicted as a linear combination of the deviations from this unconditional probability at the data locations in the neighbourhood:

$$p_i(\mathbf{s}_0 \mid \{\delta_{i_\alpha}\}) = p_i + \sum_{\alpha=1}^{n} \lambda_{\alpha,i}(\delta_{i_\alpha} - p_i). \tag{5.9}$$

where $\lambda_{\alpha,i}$ is the weight attributed to the data location $\mathbf{s}_\alpha$ for the prediction of category c_i, and δ_{i_α} is the Kronecker delta operator defined in Eq. (5.1). For ordinal categorical variables, cumulative indicators may be used, Eq. (**??**), and the predicted conditional probability obtained with Eq. (5.9) must be interpreted as the cumulative conditional probability of having categories up to c_i. The weights $\lambda_{\alpha,i}$ are chosen such that the prediction-error variance is minimized. The prediction-error variance equals

$$\sigma^2_{\mathrm{SIK}}(\mathbf{s}_0) = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_{\alpha,i}\,\lambda_{\beta,i}\,C_{ii}(\mathbf{h}_{\alpha\beta}) - 2\sum_{\alpha=1}^{n} \lambda_{\alpha,i}\,C_{ii}(\mathbf{h}_{0\alpha}) + C_{ii}(0). \tag{5.10}$$

To minimize Eq. (5.10), its partial derivatives with respect to the $\lambda_{\alpha,i}$ are equated to zero:

$$\frac{\partial\sigma^2_{\mathrm{SIK}}(\mathbf{s}_0)}{\partial\lambda_{\alpha,i}} = 2\sum_{\beta=1}^{n} \lambda_{\beta,i}\,C_{ii}(\mathbf{h}_{\alpha\beta}) - 2C_{ii}(\mathbf{h}_{0\alpha}) = 0 \qquad \alpha = 1,\ldots,n\,. \tag{5.11}$$

This results in the following system of n equations, referred to as the 'simple indicator kriging system':

$$\begin{bmatrix} C_{ii}(\mathbf{h}_{11}) & \cdots & C_{ii}(\mathbf{h}_{1n}) \\ \vdots & \ddots & \vdots \\ C_{ii}(\mathbf{h}_{n1}) & \cdots & C_{ii}(\mathbf{h}_{nn}) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{ii}(\mathbf{h}_{10}) \\ \vdots \\ C_{ii}(\mathbf{h}_{n0}) \end{bmatrix}, \tag{5.12}$$

where $C_{ii}(\mathbf{h})$ is the indicator (auto-)covariance model for category c_i.

Using Eq. (5.6) this system can also be written as:

$$\begin{bmatrix} p_{ii}(\mathbf{h}_{11}) - p_i^2 & \cdots & p_{ii}(\mathbf{h}_{1n}) - p_i^2 \\ \vdots & \ddots & \vdots \\ p_{ii}(\mathbf{h}_{n1}) - p_i^2 & \cdots & p_{ii}(\mathbf{h}_{nn}) - p_i^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} p_{ii}(\mathbf{h}_{10}) - p_i^2 \\ \vdots \\ p_{ii}(\mathbf{h}_{n0}) - p_i^2 \end{bmatrix}. \tag{5.13}$$

The minimum of the prediction-error variance, Eq. (5.10) equals

$$\sigma^2_{\mathrm{SIK}}(\mathbf{s}_0) = C_{ii}(0) - \sum_{\alpha=1}^{n} \lambda_{\alpha,i}\,C_{ii}(\mathbf{h}_{\alpha 0}). \tag{5.14}$$

5.1.2 Ordinary indicator kriging

The simple kriging variance Eq. (5.10) does not include uncertainty about the unconditional probability, which might be unrealistic. If we want to incorporate this uncertainty, then we recommend to predict the conditional probability of a given category c_i with the ordinary indicator kriging predictor in which the unconditional probability (within neighbourhoods) is estimated from the sample data:

$$p_i(\mathbf{s}_0 \mid \{\delta_{i_\alpha}\}) = \sum_{\alpha=1}^{n} \lambda_{\alpha,i}\delta_{i_\alpha}. \tag{5.15}$$

To obtain unbiased predictions, the weights must sum to 1. The prediction-error variance can be minimized under this constraint by the Lagrangian multiplier method. This yields the following ordinary indicator kriging system:

$$\begin{bmatrix} C_{ii}(\mathbf{h}_{11}) & \cdots & C_{ii}(\mathbf{h}_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{ii}(\mathbf{h}_{n1}) & \cdots & C_{ii}(\mathbf{h}_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{bmatrix} = \begin{bmatrix} C_{ii}(\mathbf{h}_{10}) \\ \vdots \\ C_{ii}(\mathbf{h}_{n0}) \\ 1 \end{bmatrix}, \tag{5.16}$$

where ν is a Lagrange parameter. In terms of bivariate probabilities this system can be written as

$$\begin{bmatrix} p_{ii}(\mathbf{h}_{11}) & \cdots & p_{ii}(\mathbf{h}_{1n}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ p_{ii}(\mathbf{h}_{n1}) & \cdots & p_{ii}(\mathbf{h}_{nn}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \nu \end{bmatrix} = \begin{bmatrix} p_{ii}(\mathbf{h}_{10}) \\ \vdots \\ p_{ii}(\mathbf{h}_{n0}) \\ 1 \end{bmatrix}. \tag{5.17}$$

The variance of the ordinary indicator kriging error equals

$$\sigma^2_{\mathrm{OIK}}(\mathbf{s}_0) = C_{ii}(0) - \sum_{\alpha=1}^{n} \lambda_{\alpha,i}\, C_{ii}(\mathbf{h}_{\alpha 0}) - \nu. \tag{5.18}$$

5.1.3 Indicator-cokriging

In simple and ordinary indicator kriging the indicators associated with the other categories (the categories not being processed) are not used. This can be inefficient, especially when the number of categories is small, and a clear dependence exists between the indicators. In that case co-kriging of the indicators can be advantageous. The simple indicator co-kriging predictor can be written as

$$p_i(\mathbf{s}_0 \mid \{\delta_{i_\alpha}\}) = \sum_{i=1}^{n_c}\sum_{\alpha=1}^{n} \lambda_{\alpha,i}(\delta_{i_\alpha} - p_i) + p_i, \tag{5.19}$$

while the ordinary indicator co-kriging predictor is given by

$$p_i(\mathbf{s}_0 \mid \{\delta_{i_\alpha}\}) = \sum_{i=1}^{n_c}\sum_{\alpha=1}^{n} \lambda_{\alpha,i}\delta_{i_\alpha}, \tag{5.20}$$

subject to the unbiasedness constraints

$$\left\{ \begin{array}{lcll} \sum_{\alpha=1}^{n} \lambda_{\alpha,i} & = & 1 & \\ \sum_{\alpha=1}^{n} \lambda_{\alpha,\beta} & = & 0 & \forall j \neq i \end{array} \right. \tag{5.21}$$

In case of indicator-cokriging, the (auto)covariance and cross-covariance models cannot be fitted independently. They must satisfy the positive-definiteness condition. This can be ensured by fitting a Linear Model of Coregionalization (LMC) (Goovaerts, 1997, p. 107-123). For categorical variables, Bogaert (2002) has shown that the only valid expression for the LMC is

$$\Sigma_{\mathbf{h}} = \Sigma_{\mathbf{0}} \sum_{m} w_m C^m(\mathbf{h}), \tag{5.22}$$

with $\sum_m w_m = 1$ and m the number of models considered in the nested structure. In this expression, the $C^m(\mathbf{h})$'s are permissible models of covariance functions with $C^m(\mathbf{0}) = 1$, the w_m are scalar constants, and

$$\Sigma_{\mathbf{h}} = \begin{bmatrix} C_{\delta_1\delta_1}(\mathbf{h}) & \dots & C_{\delta_1\delta_{n_c}}(\mathbf{h}) \\ \vdots & \ddots & \vdots \\ C_{\delta_{n_c}\delta_1}(\mathbf{h}) & \dots & C_{\delta_{n_c}\delta_{n_c}}(\mathbf{h}) \end{bmatrix}, \tag{5.23}$$

and

$$\Sigma_{\mathbf{0}} = \begin{bmatrix} C_{\delta_1\delta_1}(\mathbf{0}) & \dots & C_{\delta_1\delta_{n_c}}(\mathbf{0}) \\ \vdots & \ddots & \vdots \\ C_{\delta_{n_c}\delta_1}(\mathbf{0}) & \dots & C_{\delta_{n_c}\delta_{n_c}}(\mathbf{0}) \end{bmatrix}. \tag{5.24}$$

Any other form of the LMC leads to inconsistent models for bivariate probabilities. The fitting of the LMC for the indicators can become difficult, especially when the number of categories is large, and data are scarce.

5.1.4 Soft-indicator kriging

In some situations we are uncertain about the categories at a sample location. As an example, consider the water table class. The mean highest water table (MHW) often is difficult to determine on the basis of hydromorphic soil properties, and consequently we may conclude that there are two or even more water table classes with non-zero probabilities at a measurement location. These soft observations of the categories can be used if we are able to quantify the probabilities for the categories. These probabilities are then used in simple or ordinary indicator kriging.

5.2 Observations at points and map of categories

We now discuss the case where besides observations of categories at points, we also have a map of the categories. The mapping units of this map are impure, i.e. a mapping unit of a given category also contains locations with other categories. Nonetheless, the map provides useful information. The map is used by letting the unconditional probabilities of occurrence of the categories depend on the mapping unit.

5.2.1 Simple indicator kriging with local prior means

In simple indicator kriging with local prior means a table with (one-point) bivariate probabilities is constructed, and the entries of this table are used as unconditional probabilities in simple indicator kriging. The procedure is analogous to that described in section (5.1.1).

5.3 Indicator simulation of soil categories

Maps with soil categories can be generated by sequential simulation. The procedure is analogous to that desccribed in section (2.3). The simulation nodes are visited in a random sequence. In the absence of prior conditioning data (unconditional simulation), at the first node a random category is drawn from the unconditional probability distribution of the soil categories, which is assumed known. The drawn category is used as a conditioning, hard observation in simulating soil categories at subsequent nodes. The conditional probabilities are estimated by simple indicator kriging, Eq. (5.9). Software for indicator simulation of soil categories is available in GSLIB (Deutsch and Journel, 1992), and in GSTAT (Pebesma and G.Wesseling, 1998). Finke et al. (1999) used the latter software for stratified simulation of 28 soil/vegetation categories in Europe.

If prior data are available, and one wants to use these data as conditioning data in conditional simulation, then the conditional probabilities can be estimated by ordinary indicator kriging, Eq. (5.15).

Cross-correlations between the indicators can be taken into account by estimating the conditional probabilities by simple indicator co-kriging, Eq. (5.19) or ordinary indicator co-kriging, Eq. (5.20). This requires calibration or the postulation of a complex model with many parameters.

Chapter 6

Bayesian Maximum Entropy

In chapter (3) the Bayesian Maximum Entropy approach was introduced for spatial prediction of continuous soil properties. Here we will elaborate on BME for categorical variables. Similar to BME estimation for continuous variables, the pdf for categorical variables is obtained in two steps:

- computing the unconditional multi-point pdf
- conditioning the unconditional multi-point pdf

With respect to the availability of data and information, in the sections hereafter the following situations with respect to the available data and information are considered:

1. observations at points only
2. observations at points and map of categories

6.1 Observations at points only

This section deals with the situation where one has observations of the categories at points only. A map depicting the categories is absent, so at the prediction location itself we have no information about the category. For this situation we define the multi-point discrete probability distribution function (pdf) for the random variables $C(\mathbf{s}_0), \ldots, C(\mathbf{s}_n)$ as:

$$\begin{aligned} &\pi_{C\ldots C}(c_i, \ldots, c_j; \mathbf{s}_0, \ldots, \mathbf{s}_n) \equiv \\ &P\left[C(\mathbf{s}_0) = c_i, \ldots, C(\mathbf{s}_n) = c_j\right]. \end{aligned} \tag{6.1}$$

This pdf can be represented as an $n_c \times ... \times n_c$ hypersquare probability table with $n + 1$ dimensions. For instance, for a prediction location with two neighbouring data points, this probability table is a cube with $n_c \times n_c \times n_c$ cells with three-point probabilities (see Figure 6.1).

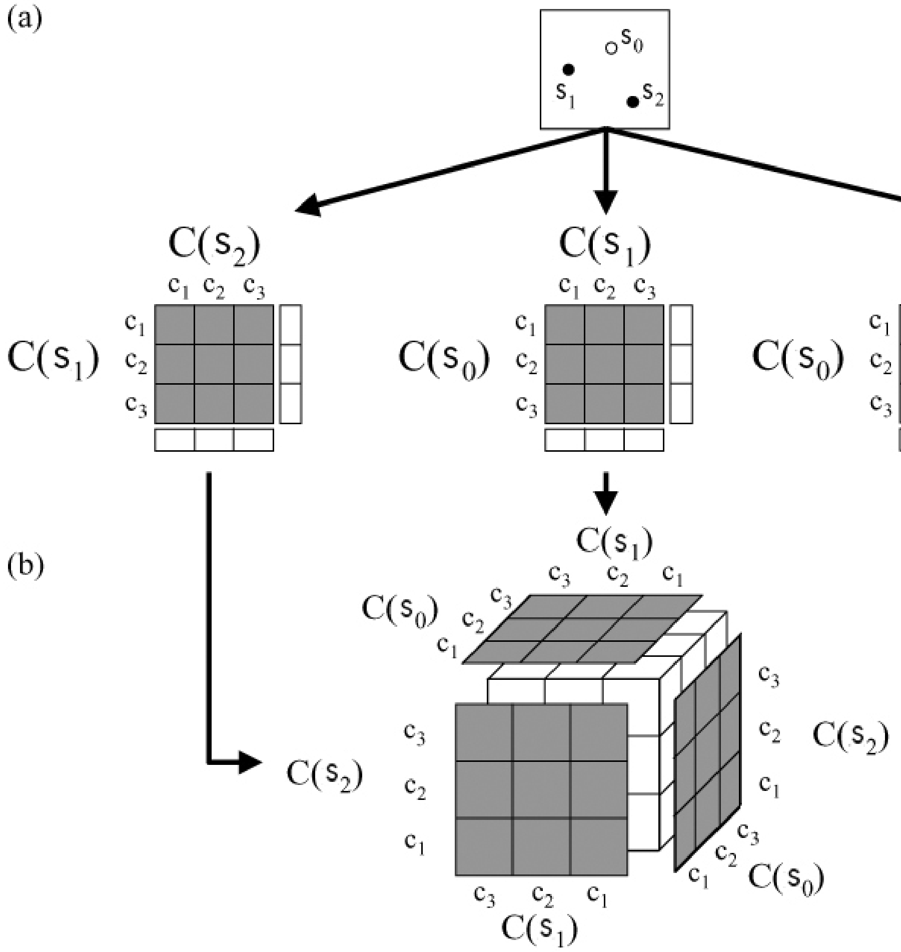

Figure 6.1: Maximization of the entropy for three categories and three locations ($\mathbf{s}_0$: prediction location; $\mathbf{s}_1$ and $\mathbf{s}_2$: data locations) (a) Construction of the two-point probability tables. (b) Searching for the three-point probabilities with maximum entropy that satisfies the constraints on the three tables with two-point probabilities. Copied with permission from D'Or (2003)

.

6.1.1 Computing the unconditional multi-point pdf

In the first step the unconditional multi-point pdf is computed by maximizing the entropy under constraints specified hereafter. The entropy of the multi-point discrete pdf is defined as (compare with Eq. (3.1))

$$\begin{aligned} H &= -\textstyle\sum_{i,\ldots,j,k} \pi_{C\ldots C}(c_i,\ldots,c_j;\mathbf{s}_0,\ldots,\mathbf{s}_n) \\ &\quad \log \pi_{C\ldots C}(c_i,\ldots,c_j;\mathbf{s}_0,\ldots,\mathbf{s}_n), \end{aligned} \tag{6.2}$$

where $\sum_{i,\ldots,j}$ is a sum over al possible combinations of categories at the prediction location and observation locations.

Constraints are imposed on the two-point probabilities of occurrence of the categorical variable C. The two-point probabilities for a given pair of locations can be obtained by summing the multi-point probabilities over the cells along the axes represented by the remaining locations. For instance, in Figure 6.1, the three marginal tables of the cube contain two-point probabilities. In general, for any dimension > 2, the two-point probabilities are obtained by summing the multi-point probabilities over the cells along the axes represented by the locations not considered.

In computing the multi-point probabilities the marginal two-point probabilities must equal the two-point probabilities as estimated from the data:

$$\pi_{CC}(c_i, c_j; \mathbf{s}_\alpha, \mathbf{s}_\beta) = p_{CC}(c_i, c_j; \mathbf{s}_\alpha, \mathbf{s}_\beta) \qquad \forall \alpha \neq \beta; \forall i, j = 1, \ldots, n_c, \tag{6.3}$$

where $\pi_{CC}(c_i, c_j; \mathbf{s}_\alpha, \mathbf{s}_\beta)$ is the marginal two-point probability, and $p_{CC}(c_i, c_j; \mathbf{s}_\alpha, \mathbf{s}_\beta)$ the two-point probability estimated from the data.

Assuming invariance under translation, Eq. (1.8), and rotation, the two-point probabilities $p_{CC}(c_i, c_j; \mathbf{s}_\alpha, \mathbf{s}_\beta)$ equal $p_{CC}(c_i, c_j; h_{\alpha\beta})$. For a chosen series of lags h, these two-point probabilities can be estimated from the data by the method-of-moments estimator

$$p_{CC}(c_i, c_j; h) = \frac{1}{N_h} \sum_{\alpha=1}^{N_h} \delta(c_i; \mathbf{s}_\alpha) \cdot \delta(c_j; \mathbf{s}_\alpha + h), \tag{6.4}$$

with N_h the number of pairs of points separated by lag h, and $\delta(c_i; \mathbf{s}_\alpha)$ the Kronecker delta operator:

$$\delta(c_i; \mathbf{s}_\alpha) = \begin{cases} 1 \text{ if } C(\mathbf{s}_\alpha) = c_i \\ 0 \text{ if } C(\mathbf{s}_\alpha) \neq c_i \end{cases} \tag{6.5}$$

In practice a tolerance interval for lag h is used to obtain enough pairs of points for each lag, as classically done in variogram estimation. To be used as constraints, we also need two-point probabilities for any intermediate value of h. This can be achieved by a Gaussian kernel smoothing procedure (Bogaert, 2002).

Maximizing the multi-point entropy, Eq. (6.2), under the constraints of Eq. (6.3) can be done with iterative scaling, also referred to as iterative proportional fitting (Bishop et al., 1976, p. 83). The iterative search starts with a uniform pdf. In our case this means that the probabilities in all cells of the hypersquare probability table equal $\frac{1}{n_c^{n+1}}$ with n the number of neighbouring observation locations. In subsequent iterations the probabilities are multiplied by the ratio of the two-point probabilities used as constraints and the marginal two-point probabilities obtained by summing the multi-point probabilities of the previous iteration. For instance, if we have two neighbouring observation location, then the hypersquare probability table is a cube with n_c^3 cells, see Figure 6.1. In this case three marginal tables are used in scaling the multi-point probabilities. In formula, the first three iterations proceed as follows:

$$\begin{aligned} \pi^{(1)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) &= \pi^{(0)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) \frac{p_{CC}(c_i, c_j; \mathbf{s}_0, \mathbf{s}_1)}{\pi^{(0)}_{CC+}(c_i, c_j; \mathbf{s}_0, \mathbf{s}_1)} \\ \pi^{(2)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) &= \pi^{(1)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) \frac{p_{CC}(c_i, c_k; \mathbf{s}_0, \mathbf{s}_2)}{\pi^{(1)}_{C+C}(c_i, c_k; \mathbf{s}_0, \mathbf{s}_2)} \\ \pi^{(3)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) &= \pi^{(2)}_{CCC}(c_i, c_j, c_k; \mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) \frac{p_{CC}(c_j, c_k; \mathbf{s}_1, \mathbf{s}_2)}{\pi^{(2)}_{+CC}(c_i, c_k; \mathbf{s}_1, \mathbf{s}_2)} \end{aligned} \tag{6.6}$$

with $\pi^{(0)}_{CC+}(c_i, c_j; \mathbf{s}_0, \mathbf{s}_1)$ the marginal two-point probability at iteration 0 (start of iteration), obtained by summing the multi-point probabilities along the axis represented by the categories at the second observation location $\mathbf{s}_2$, $\pi^{(1)}_{C+C}(c_i, c_k; \mathbf{s}_0, \mathbf{s}_2)$ the marginal two-point probability at iteration 1, obtained by summing the multi-point probabilities along the axis represented by the categories at the first observation location $\mathbf{s}_1$, and $\pi^{(2)}_{+CC}(c_j, c_k; \mathbf{s}_0, \mathbf{s}_2)$ the marginal two-point probability at iteration 2, obtained by summing the multi-point probabilities along the axis represented by the categories at the prediction location $\mathbf{s}_0$. The iteration proceeds until the differences between the marginal probabilities and the probabilities used as constraints become smaller than a user-specified value.

6.1.2 Conditioning the multi-point pdf

In this step the unconditional distribution of the previous step is conditioned on the data (observations of categories) in the neighbourhood. These data and information are referred to as the specific knowledge K_S. The conditional probability distribution, also referred to as the posterior distribution, at a prediction location is defined as:

$$\begin{aligned}\pi_{C|K_S}(c_i; \mathbf{s}_0) &= P[C(\mathbf{s}_0) = c_i \mid K_S] \\ &= \frac{P[C(\mathbf{s}_0) = c_i, K_S]}{P[K_S]}, \quad \forall i_0 = 1, \ldots, n_c,\end{aligned} \tag{6.7}$$

with $P[K_S] = \sum_i P[C(\mathbf{s}_0) = c_i, K_S]$.

The observations can be either hard (errorless) or soft. In the latter case the observation errors cannot be neglected.

Hard observations

We first consider the case where the specific knowledge consists of errorless observations of the categories. The specific knowledge can then be represented as:

$$K_S \equiv \{C(\mathbf{s}_1) = c_i, \ldots, C(\mathbf{s}_n) = c_j\}. \tag{6.8}$$

The conditional probability (posterior probability) can then be obtained by substituting for the numerator in Eq. (6.7):

$$P[C(\mathbf{s}_0) = c_i, K_S] = \pi_{C\ldots C}(c_i, \ldots, c_j; \mathbf{s}_0, \ldots, \mathbf{s}_n), \tag{6.9}$$

and for the denominator:

$$P[K_S] = \sum_i \pi_{C\ldots C}(c_i \ldots c_j; \mathbf{s}_1, \ldots, \mathbf{s}_n). \tag{6.10}$$

Soft observations

The observed categories at the neighbouring locations need not be errorless to be used in the conditioning of the unconditional distribution. Also uncertain observations of the categories can be used. If the soft observations have the form of the probability of a given combination of categories at the neighbouring locations

$$K_S \equiv \{P_S[C(\mathbf{s}_1) = c_i, \ldots, C(\mathbf{s}_n) = c_j)],\ i, \ldots, j = 1, \ldots, n_c\}, \tag{6.11}$$

then the numerator in Eq. (6.7) is given by

$$\begin{gathered}P[C(\mathbf{s}_0) = c_i, K_S] = \\ \textstyle\sum_{j,\ldots,k} \pi_{C\ldots C}(c_i, c_j, \ldots, c_k; \mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_n) P_S[C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k].\end{gathered} \tag{6.12}$$

6.2 Observations at points and map of categories

This section considers the situation where one has a map depicting the categories, besides observations of the categories at points. An example of BME prediction and simulation making use of a traditional soil map as soft information is Brus et al. (2008). The map can be used to obtain a first estimate of the category at any prediction location. Of course the map is not errorless, i.e. the map units contain impurities, and consequently at any prediction location, the category as depicted on the map must be used as a soft observation of the category. These soft observations are treated as observations on a second categorical variable D, which is related to the primary variable C. The sets of possible outcomes for the two categorical variables C and D are equal: $\Omega_C \equiv \{c_i, i = 1, \ldots, n_c\} = \Omega_D \equiv \{d_i, i = 1, \ldots, n_d\}$, and consequently $n_c = n_d$.

For this situation, it is convenient to define the one-point bivariate probability. The probability that at a location $\mathbf{s}_\alpha$ the two categorical random variables C and D take the values c_i and d_j, respectively, is defined as

$$\pi_{CD}(c_i, d_j; \mathbf{s}_\alpha) \equiv P[C(\mathbf{s}_\alpha) = c_i, D(\mathbf{s}_\alpha) = d_j], \tag{6.13}$$

with $i = 1, \ldots, n_c$ and $j = 1, \ldots, n_d$.

Similar to the univariate case (one categorical variable), we may define two-point bivariate probabilities and multi-point bivariate probabilities. The multi-point bivariate probability is defined as:

$$\begin{gathered}\pi_{C\ldots CD\ldots D}(c_i, \ldots, c_j, d_k, \ldots, d_l; \mathbf{s}_\alpha, \ldots, \mathbf{s}_\beta) \equiv \\ P\left[C(\mathbf{s}_\alpha) = c_i, \ldots, C(\mathbf{s}_\beta) = c_j, D(\mathbf{s}_\alpha) = d_k, \ldots, D(\mathbf{s}_\beta) = d_l\right].\end{gathered} \tag{6.14}$$

In this section the categorical variable D is considered at the prediction location $\mathbf{s}_0$ only, the soft categories at the neighbouring locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are not used in prediction, leading to the multi-point bivariate probability:

$$\begin{gathered}\pi_{C\ldots CD}(c_i, \ldots, c_j, d_k; \mathbf{s}_0, \ldots, \mathbf{s}_n, \mathbf{s}_0) \equiv \\ P\left[C(\mathbf{s}_0) = c_i, \ldots, C(\mathbf{s}_n) = c_j, D(\mathbf{s}_0) = d_k\right].\end{gathered} \tag{6.15}$$

This pdf can be represented as an $n_c \times \ldots \times n_c$ hypersquare probability table with $n+2$ dimensions. Compared to the previous section, which dealt with hard observations only, there is one dimension extra, for the soft categories at the prediction location.

Again, first the unconditional multi-point pdf is computed, after which this pdf is conditioned on the data and information (specific knowledge). As we will see, the soft observation at the prediction location is used in the first step, i.e. to compute the unconditional multi-point pdf.

6.2.1 Computing the unconditional multi-point pdf

The unconditional multi-point pdf is computed much in the same way as in the situation with observations at neighbouring points only. The only difference is that in maximizing the entropy an extra constraint is used. The entropy of the multi-point discrete pdf, Eq. (6.15), equals

$$\begin{aligned} H &= -\textstyle\sum_{i,\ldots,j,k} \pi_{C\ldots CD}(c_i,\ldots,c_j,d_k;\mathbf{s}_0,\ldots,\mathbf{s}_n,\mathbf{s}_0) \\ &\quad \log \pi_{C\ldots CD}(c_i,\ldots,c_j,d_k;\mathbf{s}_0,\ldots,\mathbf{s}_n,\mathbf{s}_0), \end{aligned} \tag{6.16}$$

where $\sum_{i,\ldots,j,k}$ is a sum over al possible combinations of hard categories at all locations (prediction location plus observation locations) and soft categories at the prediction location.

The entropy is maximized under constraints on the two-point probabilities, see section (6.1.1), Eq. (6.3). Additional constraints are enforced on the one-point bivariate probabilities of hard and soft categories at the prediction location. The marginal bivariate probabilities obtained by summing the multi-point probabilities over the cells along the axes represented by the neighbouring locations must equal the bivariate probabilities estimated from the data:

$$\pi_{CD}(c_i,d_j;\mathbf{s}_0) = p_{CD}(c_i,d_j;\mathbf{s}_0) \qquad \forall i=1,\ldots,n_c, \forall j=1,\ldots,n_d. \tag{6.17}$$

To estimate the bivariate probabilities, the soft category (soil map unit) is determined for all locations with hard observations of the category. The probability of occurrence of each pair of hard and soft category can then be estimated by

$$p_{CD}(c_i,d_j;\mathbf{s}_0) = \frac{1}{N}\sum_{\alpha=1}^{N} \delta(c_i;\mathbf{s}_\alpha)\cdot\delta(d_j;\mathbf{s}_\alpha), \tag{6.18}$$

with N the number of locations with hard observations.

Again, the optimal solution can be found by iterative scaling (compare section (6.1.1)). For instance, if we have only one neighbouring observation location, then the hypersquare probability table is a cube with n_c^3 cells. In this case only two marginal tables are used in scaling the multi-point probabilities. The square with two-point probabilities of hard and soft categories is not used. In mathematical terms, the first two iterations proceed as follows:

$$\begin{aligned} \pi^{(1)}_{CCD}(c_i,c_j,d_k;\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_0) &= \pi^{(0)}_{CCD}(c_i,c_j,d_k;\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_0)\tfrac{p_{CC}(c_i,c_j;\mathbf{s}_0,\mathbf{s}_1)}{\pi^{(0)}_{CC+}(c_i,c_j;\mathbf{s}_0,\mathbf{s}_1)} \\ \pi^{(2)}_{CCD}(c_i,c_j,d_k;\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_0) &= \pi^{(1)}_{CCD}(c_i,c_j,d_k;\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_0)\tfrac{p_{CD}(c_i,d_k;\mathbf{s}_0)}{\pi^{(1)}_{C+D}(c_i,d_k;\mathbf{s}_0)}, \end{aligned} \tag{6.19}$$

with $\pi^{(0)}_{CC+}(c_i,c_j;\mathbf{s}_0,\mathbf{s}_1)$ the marginal two-point probability at iteration 0 (start of iteration), obtained by summing the multi-point probabilities along the axis represented by the soft categories at the prediction location $\mathbf{s}_0$, and $\pi^{(1)}_{C+D}(c_i,d_k;\mathbf{s}_0)$ the marginal bivariate probability of hard and soft categories at iteration 1, obtained by summing the multi-point probabilities along the axis represented by the hard categories at the observation location $\mathbf{s}_1$. The iteration proceeds until the differences between the marginal probabilities and the probabilities used as constraints become smaller than a user specified value.

6.2.2 Conditioning the multi-point pdf

In the second step the unconditional pdf of the first step is conditioned on the observations at the neighbouring sample locations and on the observed soft category at the prediction location. Using the definition of the conditional probability, we obtain:

$$\begin{aligned} P\left[(C(\mathbf{s}_0) = c_i) \mid C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right] &= \\ \frac{P\left[C(\mathbf{s}_0) = c_i, C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right]}{P\left[C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right]} \end{aligned} \tag{6.20}$$

Applying the law of total probability for the denominator, we finally obtain for the conditional pdf at the prediction location

$$\begin{aligned} P\left[(C(\mathbf{s}_0) = c_i) \mid C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right] &= \\ \frac{P\left[C(\mathbf{s}_0) = c_i, C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right]}{\sum_i P\left[C(\mathbf{s}_0) = c_i, C(\mathbf{s}_1) = c_j, \ldots, C(\mathbf{s}_n) = c_k, D(\mathbf{s}_0) = d_l\right]} \end{aligned} \tag{6.21}$$

6.3 Simulation of soil categories with Bayesian Maximum Entropy

Categorical (qualitative) soil maps can be simulated by sequential simulation, much in the same way as sequential simulation of quantitative soil maps. A random visit path through the nodes is followed. If there are no conditioning, prior data (unconditional simulation), then at the first node the prior probabilities are taken as the posterior probabilities from which a category is randomly drawn. This category is used in computing the conditional distribution (posterior distribution) at the next node, etc. If there are prior data (conditional simulation), then these data are used from the beginning of the sequence to compute the conditional distributions.

If a soil map exists, then this soil map can be used in simulating possible categorical soil maps. At the first simulation node a category is drawn from the prior distribution for the soil map unit this node is falling in. In subsequent steps, besides the soil map unit (soft category) at these nodes, the categories drawn in previous steps are used as hard conditioning data.

The BMELIB package contains several MATLAB scripts for simulation of categorical random fields (simucat.m, simucatcond.m, simucatcondPdf.m, simucatHardk.m).

Chapter 7

Markov Random Fields

Markov Random Fields (MRF) were already introduced in chapter (4). We mentioned that the MRF approach is particularly useful for modelling, prediction and simulation of qualitative spatial variables, which is the subject of this chapter. However, the methodology is largely the same as that used for MRF of quantitative variables. In this chapter, we will therefore frequently refer to chapter (4).

Our interest in this chapter is in the probability distribution of C as defined in Eq. (1.6). As before, we take an approach in which we characterize the probability distribution of C by means of conditional probability distributions, with the possibility to use Brooks Lemma to derive the full distribution, Eq. (1.6), from the conditional distribution if desired, although in practice we will rarely do so because we will use simulation methods that work with the conditional distribution.

The natural analog of the Markov property, Eq. (4.3), for qualitative variables is

$$P[C = c_k | C_i = c_i, i = 1, \ldots, n, i \neq k] = P[C = c_k | C_i = c_i, i \in \delta \mathbf{s}_k]. \tag{7.1}$$

where the neighbourhood $\delta \mathbf{s}_k$ may for instance be chosen as the Von Neumann or Moore neighbourhood as discussed in chapter (4).

The central object of the MRF approach is the conditional distribution Eq. (7.1). It is all that is needed for simulation and prediction of C. Once it is known, algorithms to generate realizations of C run smoothly although they will likely use up a lot of computing time. However, estimation of the conditional distribution from point observations and ancillary information is a tricky problem, particularly because the distributions must be chosen such that they yield a unique and proper full probability model (see also the discussion in chapter (4)).

7.1 Estimation of the conditional probability

Even though the conditional distribution of C has been confined to the values of C in its local neighbourhood, it is still enormously complex. For instance, if C can take on eight values and four neighbours are considered, Eq. (7.1) requires that $5^8 = 390625$ probabilities are specified. In order to reduce the complexity, one must make assumptions about the structure of the probabilities. One option that has been successfully applied in land use change modelling is to use a logistic regression approach:

$$Log(\frac{P}{1-P}) = \sum_j \beta_j F_j, \tag{7.2}$$

where P is the conditional probability in Eq. (7.1) and where the F_j are explanatory variables that depend on the values of C in the neighbourhood. Verburg et al. (2004) use so-called enrichment factors as explanatory variables. These are computed as the frequency of cells in the neighbourhood that have a particular outcome relative to the overall occurrence of the outcome. Other explanatory variables can also be used, such as maps of covariates (see below) or the presence/absence of a particular outcome in the neighbourhood. The difficulty is in choosing relevant explanatory variables without increasing the dimensionality too much. If the number of explanatory variables is large, then it will be more difficult to estimate the logistic regression coefficients β_j and there will be an increased risk of introducing artefacts through over-fitting.

Estimation of the β_j is a difficult task in itself. A large body of literature and tools are available from the logistic regression literature, but these typically assume that simultaneous observations of the dependent and explanatory variables are available. In our case they are not, because rarely will the soil be sampled using a grid sample with resolution equal to the resolution of the target map (in fact, if this were the case then there would be no need for modelling because the spatial distribution of the soil property would be known). Thus, estimation of the logistic regression coefficients will involve a computationally demanding calibration procedure in which a combination of regression coefficients is tried, realizations of C are generated using this combination, statistics are computed and compared with those of the observations, after which the calibration will modify one or more of the coefficients to obtain a closer match between the statistics from the generated realities and that from the observations. Realistically speaking, such a method would only work if the number of coefficients is smaller than 10, preferably even smaller.

To circumvent the computer-intensive calibration routine one might also design sampling experiments specifically dedicated to estimation of the logistic regression coefficients. This would mean that multiple locations in the field are visited, each time sampling the soil at the centre cell as well as in the neighbourhood cells.

One other option that may be tried is to let the conditional probabilities or alternatively the explanatory variables and associated regression coefficients be determined by an experienced soil surveyor. For instance, the soil surveyor should be able to tell what the chances of soil type A are at some location, given that three of its four neighbours also have soil type A and the fourth neighbour has soil type B.

7.2 Simulation and Prediction

Simulation and prediction of qualitative soil variables is practically identical to the methods described in chapter (4). The Gibbs sampler works equally well with categorical data as with continuous data and conditioning to point observations is done in the same way as before. Conditioning to maps of covariables is done by letting the conditional probabilities depend on the covariable, either directly or through using the covariables as explanatory variables in the logistic regression Eq. (7.2).

Calibration is relatively easy in this case because the covariable is available as a map and exhaustively known.

Prediction for qualitative variables is again done by generating many simulations and summarising the resulting frequency distributions. For instance, the soil type with maximum frequency (i.e. probability) may be used as a prediction, or alternatively one may use the observed frequencies for all outcomes as an estimate of the probability vector of the true categorical soil variable.

Part III

EXAMPLES: MAPPING SOIL PROPERTIES USED BY THE SMART - SUMO - MOVE MODEL CHAIN

Chapter 8

Introduction

SMART (Simulation Model for Acidification's Regional Trends) is a dynamic soil acidification model aimed at the evaluation of the effectiveness of emission control strategies for SO_2, NO_x and NH_3 at the European scale. SMART includes only geochemical buffer processes such as weathering and cation exchange. SMART2, developed for evaluations at the national level, furthermore includes nutrient cycling and solute input through upward seepage (Kros et al., 1999; Kros, 2002).

SMART2 requires quite a few of soil variables as input, such as (Kros, 2002, Table 8, p. 157).

- bulk density in the root zone (ρ)
- volumetric moisture content in the soil (θ)
- base saturation fraction (Bsat)
- cation exchange capacity (CEC)
- selectivity constant for Al - (Ca,Mg) exchange (KAlBC2)
- selectivity constant for H - (Ca,Mg) exchange (KHBC2)
- dissolution constant for Al-hydroxide (KAl_{ox})
- Al-content in secondary Al compounds in the soil (Al)
- C/N ratio of the soil (C/N)

In this report we present the simulation results for three soil properties, viz. KHBC2, KAlBC2, and KAl_{ox}. KAlBC2 appears to be related to soil type to some extent. To simulate maps of this soil property, a universal kriging model was used, with the means of soil types as fixed effects (deterministic trend). Simulation of maps of KAl_{ox} illustrates how to deal with a clearly non-normal distribution.

In the next chapter we explain how we simulated the soil types assumed to have a different frequency distribution for one or more of the soil properties used as input by the model SMART. These soil types are 1. poor sand (SP); 2. rich sand (SR); 3. calcareous sand (SC); 4. non-calcareous clay (CN); 5. calcareous clay

(CC); 6. non-calcareous loam (LN); 7. peat (PN). These soil categories are groups of units of the soil map of the Netherlands at scale 1 : 50 000. In chapter (10) the simulation of the three selected continuous soil properties is explained. The multi-point (multivariate) distribution after subtraction of the mean per soil type (KAlBC2) or after transformation (KAl$_{\mathrm{ox}}$) is assumed to be normal (Gaussian).

Chapter 9

BME simulation of soil type

The data and information used to estimate the probabilities of occurrence of the seven soil categories consist of 8369 observations of the soil categories at points taken from the Dutch Soil Information System (http://www.bodemdata.nl), and a map of the categories at scale 1 : 50 000 (Fig. 9.1). The observations at points are considered as error free (hard observations), whereas the map is treated as spatially exhaustive soft information on the soil categories.

The 8369 hard observations of the soil categories at points were used to estimate the two-point probabilities, Eq. (6.4). Figure 9.2 shows the results. Note that categories 3 to 7 show almost no spatial auto-correlation, as seen from the probabilities in the diagonal subplots that fall sharply down and reach more or less its minimum value at the second lag (ca. 1000 m) in the subplots. For categories 1 and 2 spatial auto-correlation is stronger, whereas categories 4, 5 and 6 take an intermediate position.

Table 9.1 shows the one-point bivariate probabilities for the hard and soft categories. Note that the row-totals correspond with the intercepts in the diagonal subplots of Figure 9.2, and that the table margins correspond with the category frequencies as estimated either from the hard observations or the soil map.

All computations have been made using BMELIB 2.0b, a free Matlab toolbox (see Christakos et al. (2002)) and the website http://www.unc.edu/depts/case/BMELIB). To simulate we extended the BMELIB 2.0b toolbox with an additional script named simucatHardkcond.m (available upon request). Figure 9.3 depicts the soil category with the largest estimated probability. Figure 9.4 depicts four example simulations. Differences between these four maps are noticeable but small, indicating that the information contained in the 1 : 50 000 soil map and the hard observations leaves little uncertainty about the soil categories at unobserved locations.

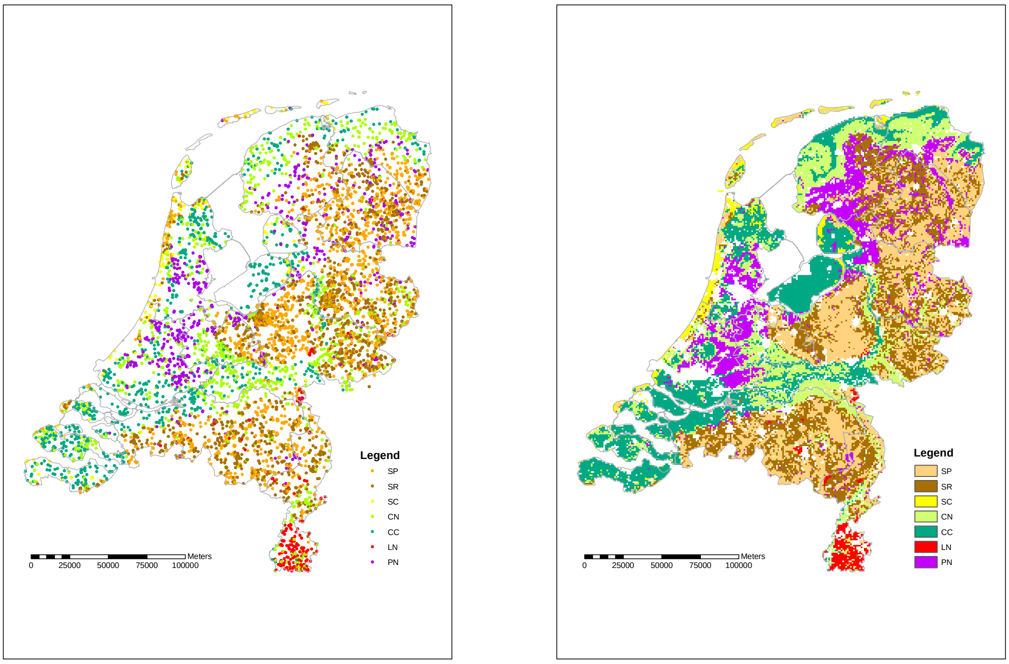

Figure 9.1: Hard observations of soil categories at points (left) and soft soil categories as derived from the soil map of the Netherlands 1 : 50 000 (right).

Table 9.1: One-point bivariate probabilities for hard and soft categories. Rows: hard categories, columns: soft categories.

	SP	SR	SC	CN	CC	LN	PN	Σ
SP	0.236	0.064	0.001	0.006	0.000	0.000	0.011	0.319
SR	0.050	0.198	0.001	0.022	0.000	0.001	0.010	0.282
SC	0.001	0.000	0.015	0.000	0.003	0.000	0.001	0.021
CN	0.006	0.020	0.000	0.113	0.012	0.003	0.007	0.160
CC	0.000	0.000	0.004	0.013	0.091	0.000	0.001	0.109
LN	0.001	0.001	0.000	0.002	0.000	0.019	0.000	0.023
PN	0.005	0.008	0.000	0.004	0.000	0.000	0.067	0.085
Σ	0.299	0.292	0.021	0.160	0.107	0.023	0.097	1.000

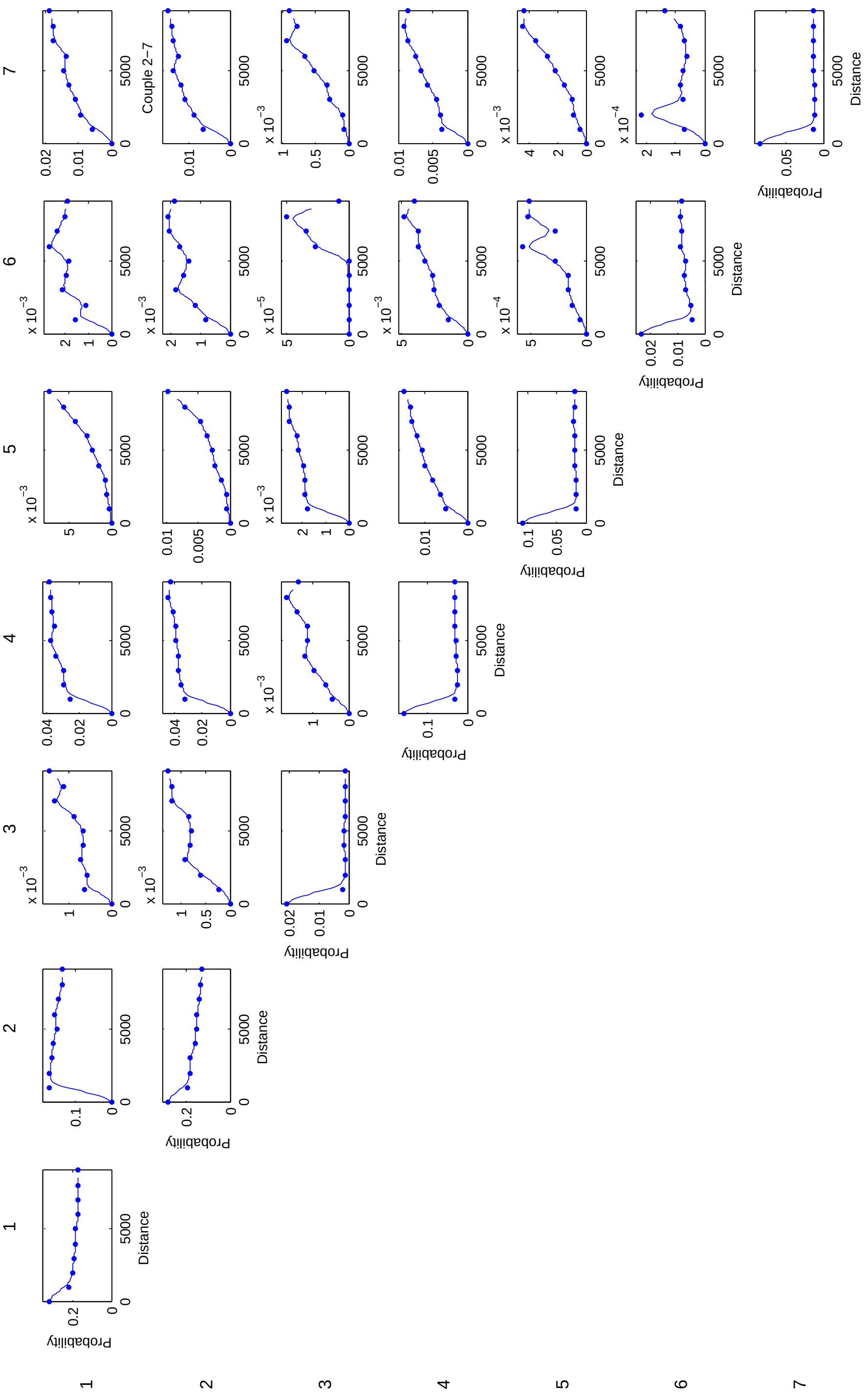

Figure 9.2: Model for two-point univariate (diagonal plots) and bivariate (off-diagonal plots) probabilities of SMART-categories. Distance in m.

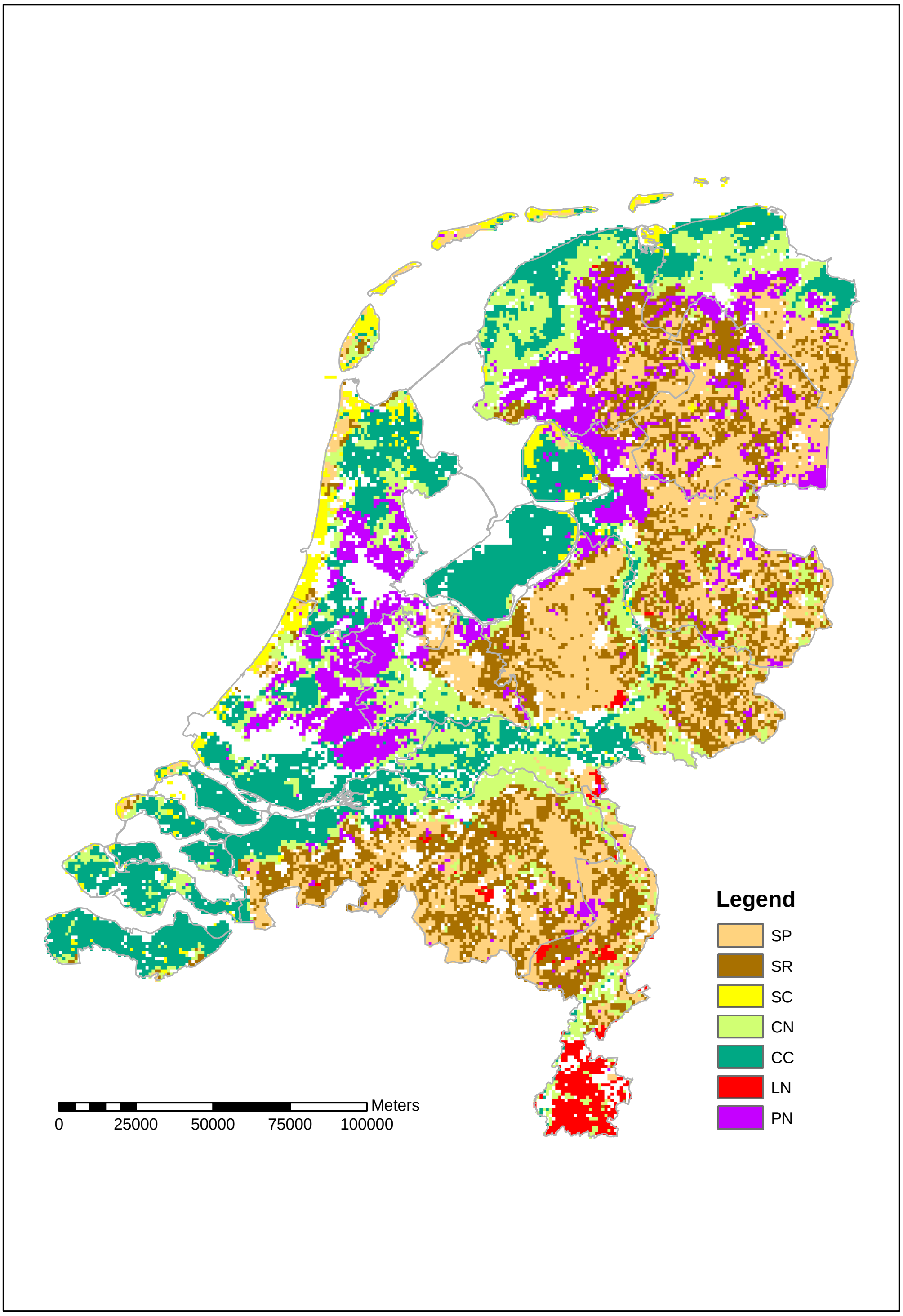

Figure 9.3: Soil categories predicted with BME from hard observations plus existing soil map.

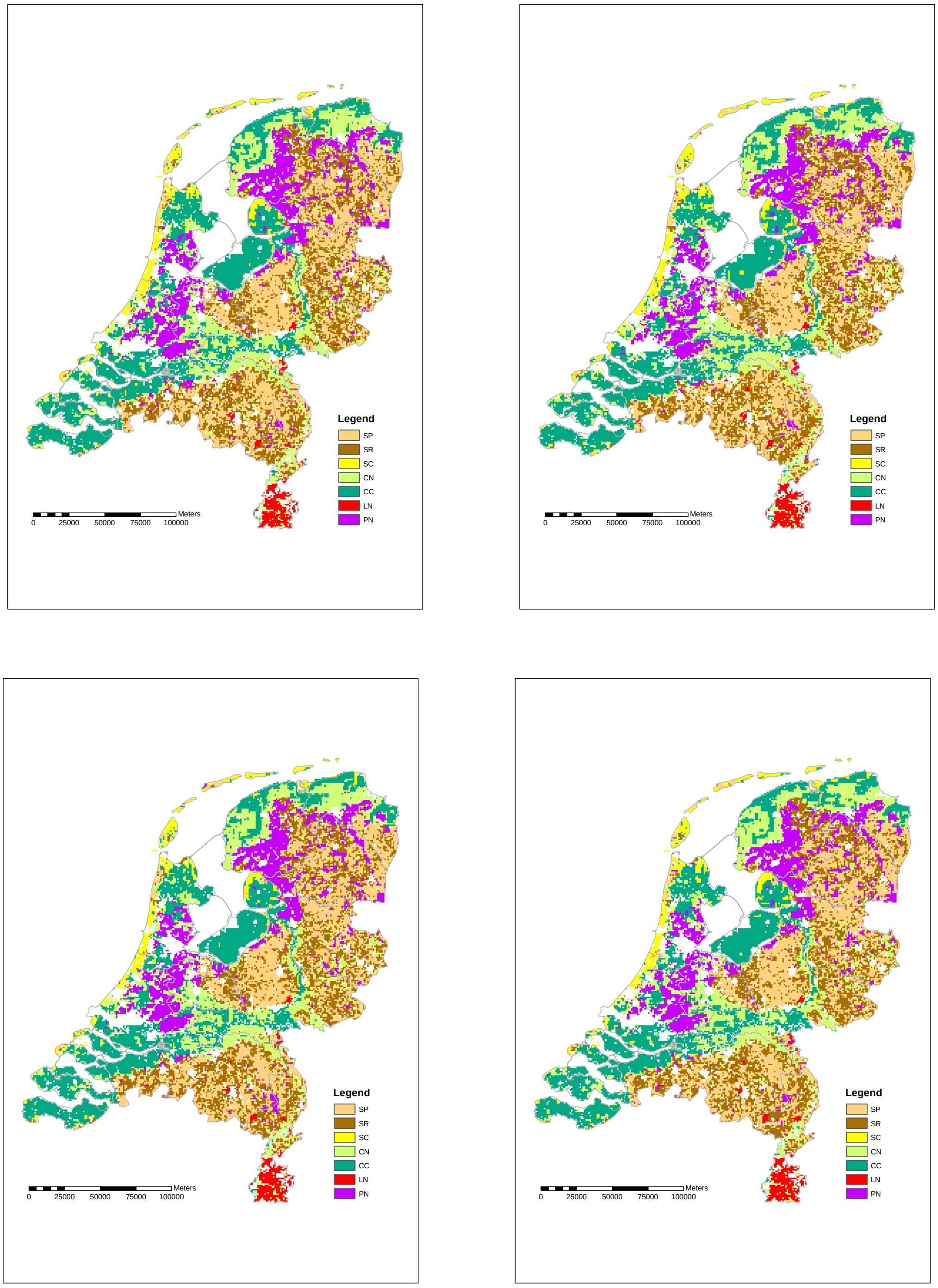

Figure 9.4: Four possible realities of the soil type map generated with BME simulation, using both hard and soft data.

Chapter 10

Sequential Gaussian simulation of continuous soil properties

Three continuous soil properties were selected for which statistical models are built and simulations are generated. These are the selectivity constant for H - (Ca,Mg) exchange (KHBC2), the selectivity constant for Al - (Ca,Mg) exchange (KAlBC2), and the dissolution constant for Al-hydroxide (KAl$_{\mathrm{ox}}$). For all three properties we use the methodology described in chapter (2) and take the model (2.1) as a starting point. The trend component μ may be a constant or a function of soil type, depending on whether soil type has predictive power for the particular property. When the trend of the continuous soil property depends on soil type, then simulated maps of the soil property use simulated maps of soil type, as created in the previous chapter, as the trend. In this way, uncertainty about soil type will affect the uncertainty about the continuous soil property. Before building the probability models and using these for spatial prediction and simulation of the three soil properties, we first explore the available data and the strength of the relationships between soil properties and soil type.

10.1 Exploratory data analysis

For exploratory data analysis and simulation of the three selected soil properties we used 317 observations from forests in the Netherlands (Klap et al., 1999). Locations were chosen such that all important soils of the Netherlands were represented. Besides this, spatial coverage and coverage of soil parent materials was taken into account in the selection of the sites.

Figure 10.1 presents histograms of the three soil properties. These are useful to detect outliers and verify whether the frequency distributions are sufficiently normal. None of the three distributions show clear outliers, but the distributions of KAl$_{\mathrm{ox}}$ and to a lesser extent KAlBC2 are clearly skewed. The histogram of the log-transfomed KAl$_{\mathrm{ox}}$ (using a shift parameter of 100) is therefore also given. It is apparent that log-transformation removes much of the skewness in KAl$_{\mathrm{ox}}$.

The distribution of the three soil properties over the seven soil types is given in Figure 10.2. Note that soil type CC (calcareous clay) has no observations. The box

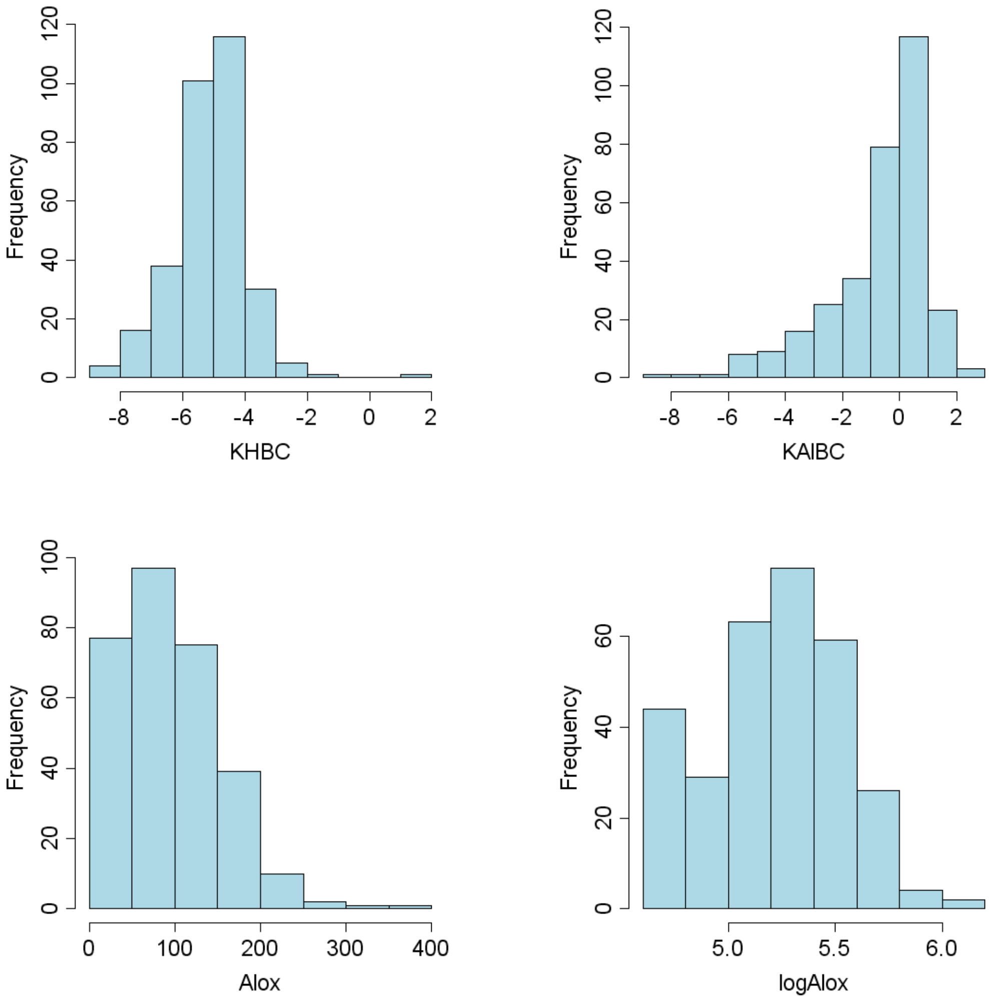

Figure 10.1: Histograms of the three selected soil properties and the log-transformed $\mathrm{KAl_{ox}}$.

plots show that KAlBC2 strongly depends on soil type, while KHBC2 and $log\mathrm{KAl_{ox}}$ are less influenced by soil type. The $\mathrm{KAl_{ox}}$ (and $log\mathrm{KAl_{ox}}$) values for soil type SC (calcareous sand) are markedly different from those for other soil types, but the frequency distribution is based on 7 observations only. KHBC2 has smaller values for soil type PN (peat), but overall KHBC2 does not seem to vary much with soil type.

Based on the results of the exploratory data analysis it was decided to log-transform $\mathrm{KAl_{ox}}$ prior to model building and not to tranform KHBC2 and KAlBC2. Also, it was assumed that $log\mathrm{KAl_{ox}}$ and KHBC2 had a constant model mean, whereas the model mean of KAlBC2 was assumed to vary with soil type. In all three cases it was assumed that the variance of the soil properties is the same for all soil types. It must be noted that these decisions are based on a preliminary assessment. A more thorough analysis might reveal more appropriate assumptions about the characteristics of the uncertain soil properties. However, the assumptions made are judged

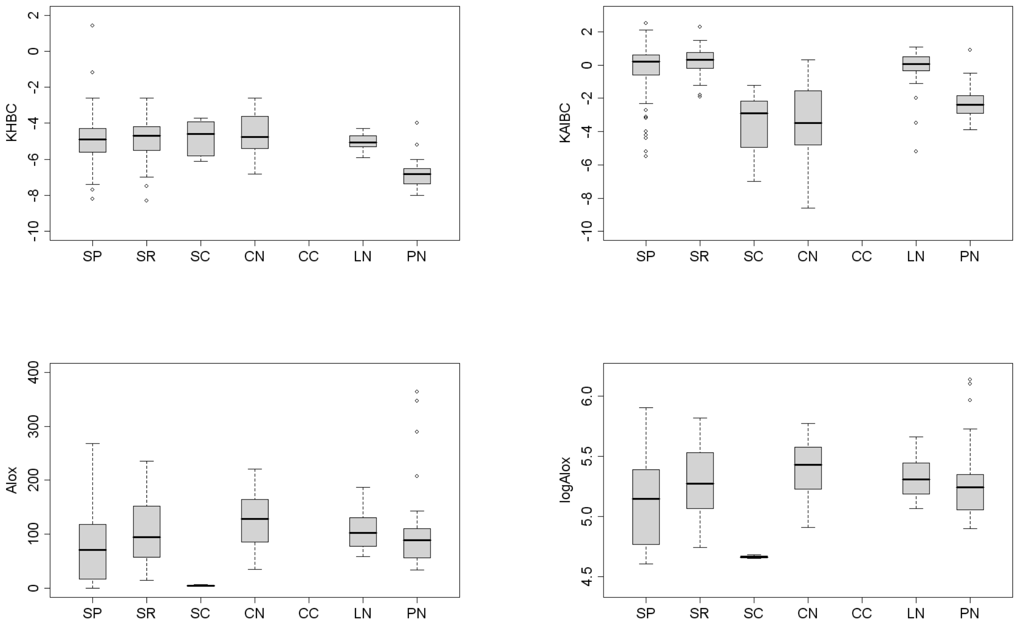

Figure 10.2: Box plots of the soil properties for each of the seven soil types.

satisfactory, because the purpose of this chapter is mainly to illustrate the theory of previous chapters, not to produce the definitive probability models of the three soil properties. For illustration purposes, we prefer relatively simple models.

10.2 Variography

Experimental semivariograms of the three soil properties are given in Figure 10.3. Note that for KAlBC2 the semivariogram was computed on its residual, i.e. after removal of the soil type-dependent mean. Note also that anisotropy was not considered, since there is no plausible reason why the spatial correlation of the soil properties should depend on direction. All three semivariograms show clear spatial dependence, with spatial correlation lengths ranging from 60 (KHBC2 and KAlBC2) to 100 km ($log\text{KAl}_{\text{ox}}$). The nugget-to-sill ratio is approximately 0.5 for KHBC2 and KAlBC2, indicating substantial short-distance spatial variation. For $log\text{KAl}_{\text{ox}}$ it is only about 0.2. This, together with the fact that its spatial correlation length is large, implies that $log\text{KAl}_{\text{ox}}$ is much more strongly spatially correlated than KHBC2 and (residual) KAlBC2. This should be confirmed by less noise in the simulated maps to be presented in the next section. Consultation with soil scientists specialized in the spatial distribution of these soil properties may help explain the similarities and differences of the three semivariograms.

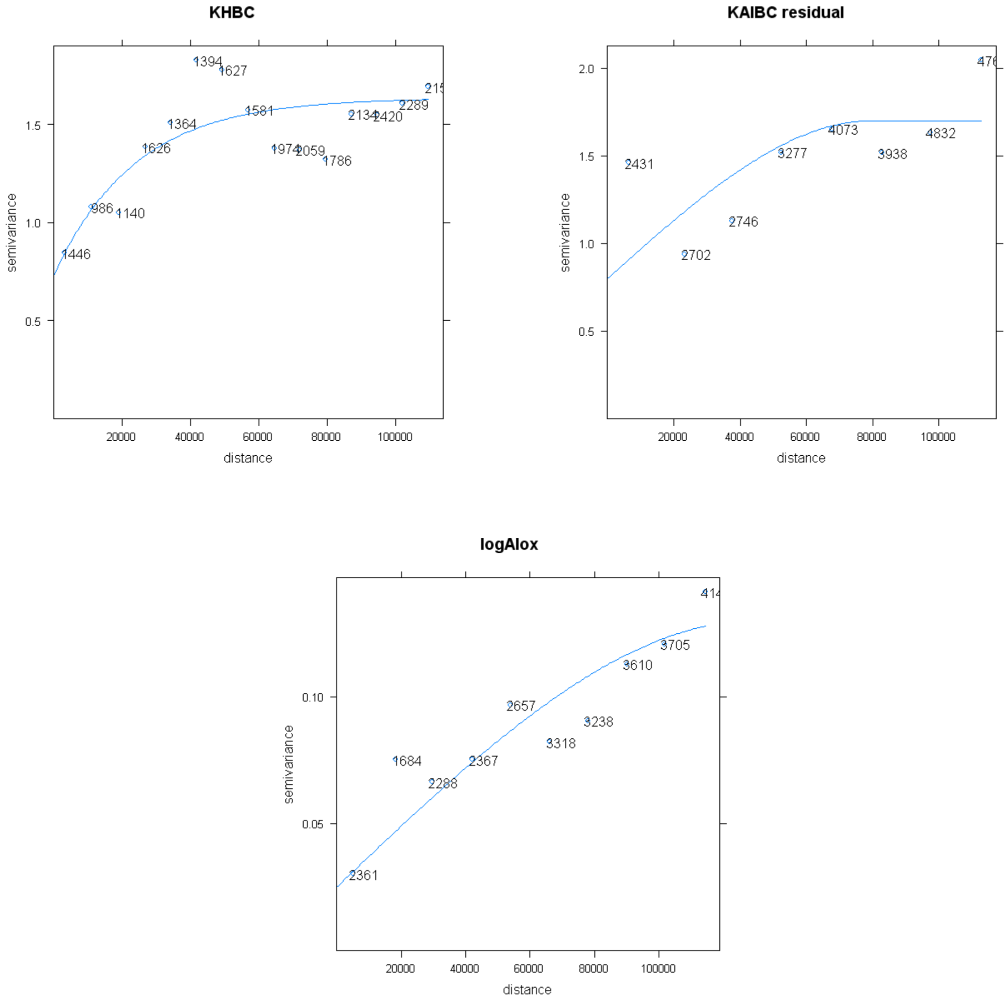

Figure 10.3: Experimental and fitted semivariograms of the three soil properties.

10.3 Spatial prediction and simulation

10.3.1 Selectivity constant for H - (Ca,Mg) exchange (KHBC2)

Since the model mean and variance of KHBC2 was assumed constant and spatial correlation was assumed to depend only on the geographical distance between locations, a map of KHBC2 may be created from the 317 point observations using ordinary kriging (section (2.1.2)). The maps of the ordinary kriging prediction and kriging variance are given in Figure 10.4. The spatial pattern in the prediction map cannot easily be explained. Small values are generally predicted in the lower parts of the Netherlands with marine and fluviatile sediments, with an exception for the North-West, where large predictions are made. Perhaps these are caused by a spatial outlier in the data. Greater values are generally predicted for the sandy, higher elevation parts in the East and South-East, but again a few spatial outliers mask this effect. The pattern of the kriging variance map reflects the spatial configuration

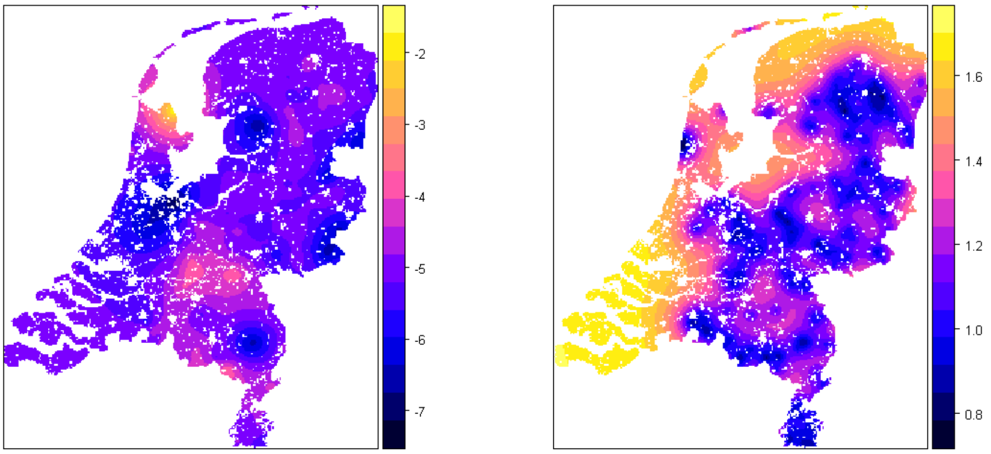

Figure 10.4: Maps of kriging predictions (left) and kriging variances (right) for KHBC2.

of the 317 point locations: the variance is small near observation locations and large further away from them. Indeed, few observations were taken in the South-West and North of the country. The average kriging variance is about 1.0, indicating a fairly large uncertainty about the predicted KHBC2 (i.e. standard deviations are on average 1.0 as well, with predicted values ranging from -7 to -2).

Figure 10.5 presents four possible realities of KHBC2 from an infinite number that could have been generated. These were created using conditional sequential Gaussian simulation, using the 317 observations as conditioning data (see section (1.4.2)). The global pattern of the four simulated maps is the same as that of the kriging prediction map, but the simulated realities are more erratic. The amount of 'noise' that is superimposed over the prediction map signifies our uncertainty about KHBC2. The uncertainty is also reflected in the differences between the four simulated realities: the greater the differences, the greater the uncertainty. Note that the differences are indeed greater in the South-West, there where the kriging variance is large and where there are no conditioning data.

10.3.2 Selectivity constant for Al - (Ca,Mg) exchange (KAlBC2)

The kriging prediction and prediction variance maps of KAlBC2 are given in Figure 10.6. In this case universal kriging was used, because the model of KAlBC2 assumes a non-constant mean. Since the mean of KAlBC2 depends on soil type, the underlying pattern of the Dutch soil map (right panel of Figure 9.1) is clearly visible in the kriging prediction map. Superimposed on the per-unit mean is the interpolated residual of KAlBC2, which causes gradual changes in the interpolated values within the mapping units. However, these variations are small compared to the differences between the map unit means. Consequently, fairly discrete jumps in the predicted KAlBC2 are observed, such as at the border between the clayey river soils and sandy aeolian deposits. Note also that the kriging variance map is hardly influenced by the soil map. This indicates that the uncertainty in the kriging predictions is mainly caused by the interpolation error, not by the estimation errors of

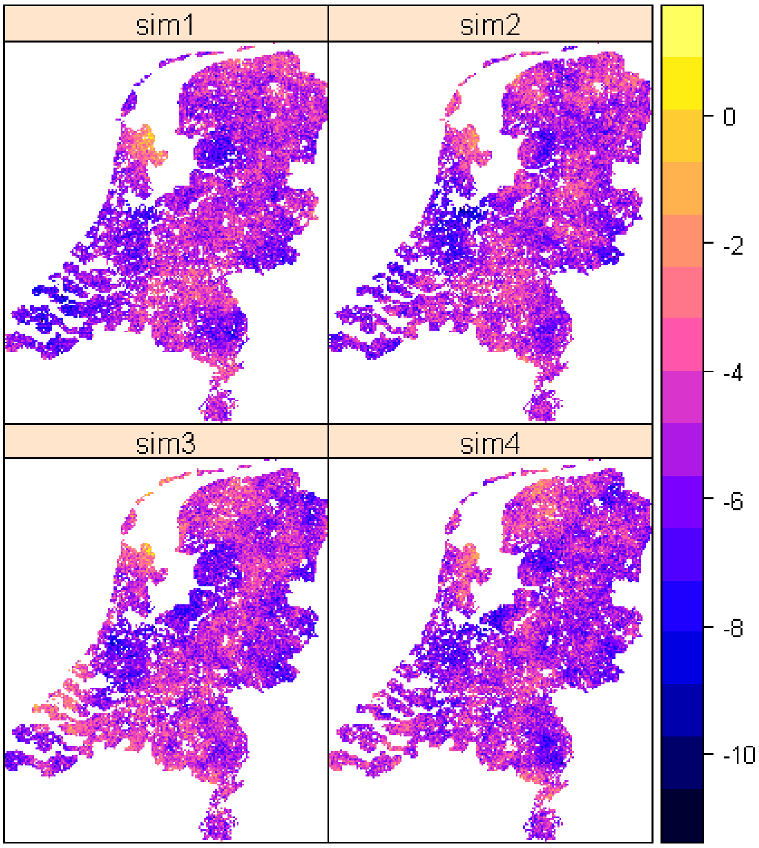

Figure 10.5: Four possible realities of KHBC2 generated with conditional sequential simulation.

the per-unit means. In fact, this should not come as a surprise, because the number of observations used to estimate the means are fairly large.

The four example realizations of KAlBC2 given in Figure 10.7 all used the Dutch soil map (i.e. right panel of Figure **??**) to stratify the Netherlands in different soil mapping units. However, the soil map itself is uncertain as well. In the previous chapter we simulated possible realities of the soil map using the Bayesian Maximum Entropy method. To acknowledge the uncertainty about the soil map, we must therefore also use these simulated soil maps (i.e. such as those presented in Figure 9.4) as starting points for simulation of KAlBC2. The simulated maps also show the influence of the soil map, but much less pronounced than for the kriging prediction map. Here, the residual variation is not averaged out by the smoothing effect of kriging, but adds a substantial spatially correlated 'noise' to the per-unit means. Nonetheless, the relatively small differences between the simulated realities shows that the uncertainty about the spatial distribution of the true KAlBC2 is small compared to the predicted pattern. Note that the differences between the maps are small.

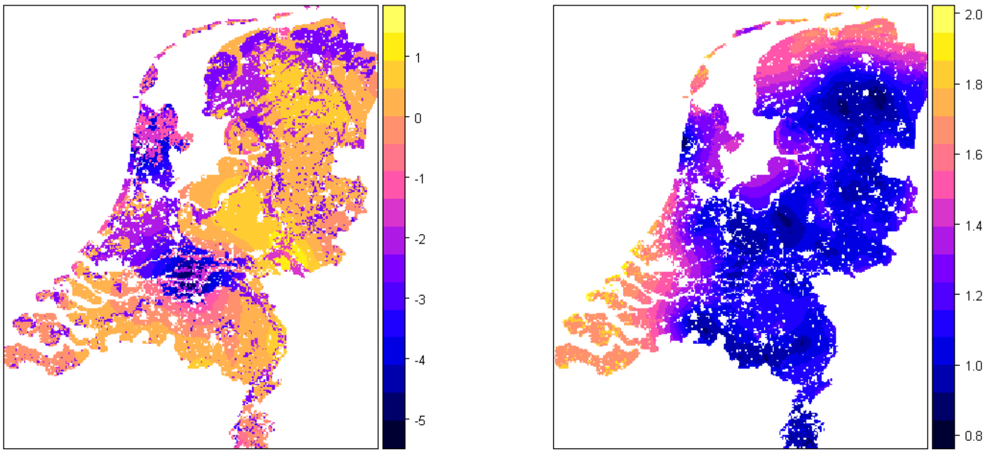

Figure 10.6: Maps of kriging predictions (left) and kriging variances (right) for KAlBC.

10.3.3 Dissolution constant for Al-hydroxide ($\mathrm{KAl_{ox}}$)

The kriging maps for $log\mathrm{KAl_{ox}}$ are given in Figure 10.8. Note that the legends of these maps refer to the log-transformed $\mathrm{KAl_{ox}}$. Back-transformation to the original scale is not trivial, but feasible (Journel and Huijbregts, 1978). Direct transformation of the kriged map by taking its antilog produces the median $\mathrm{KAl_{ox}}$ value, not the mean. The spatial pattern of the predicted $log\mathrm{KAl_{ox}}$ show small values in the North-West and in the marine clay soils of the polders near the IJsselmeer and in the South-Western part of the Netherlands. However, these patterns cannot be directly related to soil type because the West has similar soils but has large predicted values (see also Figure 10.2).

Four simulations of $\mathrm{KAl_{ox}}$ are presented in Figure 10.9. Back-transformation is easy and involves a simple antilog operation. Note the skew distribution of the simulated $\mathrm{KAl_{ox}}$ (i.e. the main part of the map has small values shown in blue, whereas there are much fewer locations with large values shown in orange and yellow). The patterns are similar to what we observed before and show a superposition of the kriged map and a spatially correlated 'noise'. As before, the differences between the four maps are a measure of the uncertainty about the spatial distribution of $\mathrm{KAl_{ox}}$.

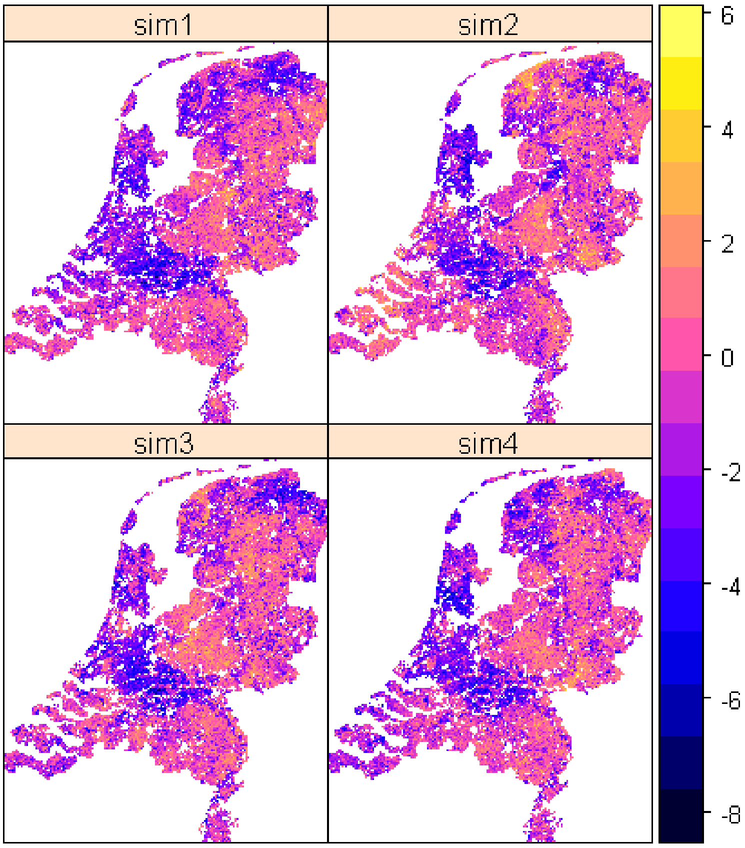

Figure 10.7: Four possible realities of KAlBC2 generated with conditional sequential simulation, taking uncertainty in the soil categories into account.

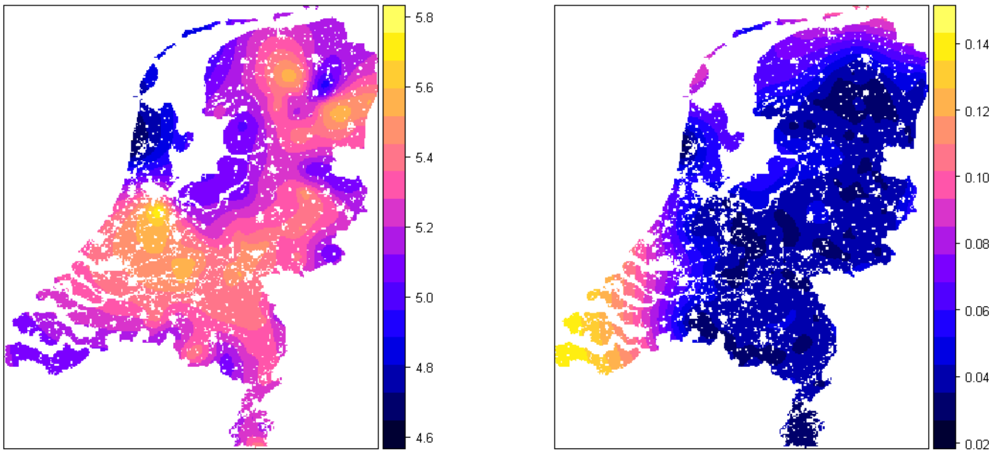

Figure 10.8: Maps of kriging predictions (left) and kriging variances (right) for log-transformed $\mathrm{KAl}_{\mathrm{ox}}$.

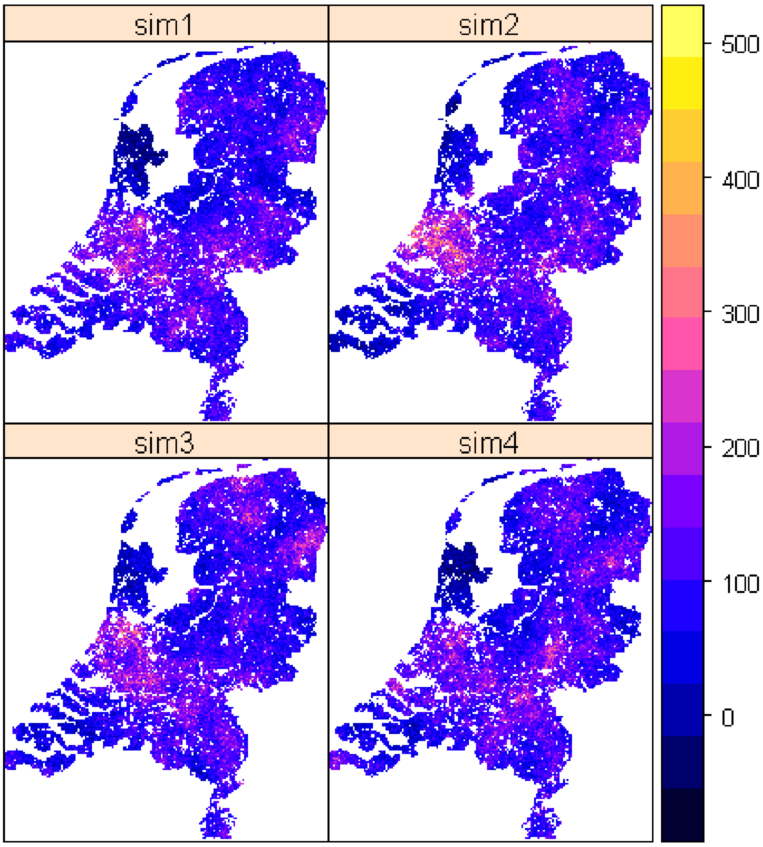

Figure 10.9: Four possible realities of KAl_{ox} generated with conditional sequential simulation.

Chapter 11

Discussion and conclusions

This chapter discusses the suitability of the three methods (Kriging, Bayesian Maximum Entropy and Markov Random Fields) for prediction and stochastic simulation of soil maps. For quantitative soil properties, sequential simulation making use of one of the Kriging models has proven its value. There are many examples in the literature, and well-tested public domain software is widely available.

Spatial prediction and simulation with Bayesian Maximum Entropy is far less well-known. There are not many case studies yet that have been published, and its potential still has to be explored further. Nevertheless, Bayesian Maximum Entropy has proven its value already for prediction and simulation of categorical variables such as soil type. Compared to the kriging alternative, indicator kriging, this method is much better founded on theory and, perhaps more importantly, results have shown to be superior to those obtained with indicator kriging. A serious problem encountered in this project is the computational burden of the Bayesian Maximum Entropy method for categorical methods. Only for a very limited number of categories (say < 10), and a very limited number of neighbouring observations (say < 5), computing time is acceptable for projects at a national scale. For larger numbers of categories, one may possibly make use of the hierarchical nature of the soil classification system. We recommend this as an interesting topic for future research.

Bayesian Maximum Entropy also seems to have interesting potentials for prediction and simulation of continuous soil variables in situations where (part of) the observations are uncertain (soft). The Soil Information System contains numerous soft data, such as soil profile descriptions with interval estimates of soil organic matter, clay content and Mean Highest Water table (MHW), and censored observations such as clay content $< 1\%$ and MHW $> 1.20\ m$. Kriging also offers possibilities for dealing with soft data, but these possibilities seem to be more restricted, because it assumes that the soft observations are Gaussian distributed, which is clearly not realistic for interval estimates and censored observations of soil properties. Public domain software for Bayesian Maximum Entropy prediction and simulation is available (BMELIB), but is relatively new and, at least some programmes in this library, are not yet widely tested.

Even more unexplored than Bayesian Maximum Entropy is the approach based on Markov Random Fields. Compared to Bayesian Maximum Entropy the computational burden is small, even for a large number of soil categories. However, little work has been done on calibrating the Markov Random Field model. One problem, yet to be explored in more detail, is that the calibrated model must satisfy some basic requirements to arrive at a valid distribution for the variable of interest. It is only fair to say that the ins and outs of conditional density estimation under the Markov Random Field approach is a largely undiscovered field that needs to be thoroughly investigated before the method can be applied in practice. Both practical as well as theoretical problems need to be tackled.

Bibliography

Baltzer, H. (2000). Markov chain models for vegetation dynamics. *Ecological Modelling*, 126:139–154.

Baltzer, H., Braun, P. W., and Kohler, W. (1998). Cellular automata models for vegetation dynamics. *Ecological Modelling*, 107:113–125.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC Press, Boca raton.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1976). *Discrete Multivariate Analysis. Theory and Practice.* The Massachusetts Institute of Technology Press, Cambridge.

Bogaert, P. (2002). Spatial prediction of categorical variables: the Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment*, 16:425–448.

Bogaert, P. and D'Or, D. (2002). Estimating soil properties from thematic soil maps: the Bayesian maximum entropy approach. *Soil Science Society of America Journal*, 66:1492–1500.

Brus, D. J., Bogaert, P., and Heuvelink, G. B. M. (2008). Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. *European Journal of Soil Science*, (in press).

Brus, D. J., de Gruijter, J. J., Marsman, B. A., Visschers, R., Bregt, A. K., and A. Breeuwsma, J. B. (1996). The performance of spatial interpolation methods and choropleth maps to estimate properties at points: a soil survey case study. *Environmetrics*, 7:1–16.

Brus, D. J. and Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138:86–95.

Brus, D. J. and Jansen, M. J. W. (2004). Uncertainty and sensitivity analysis of spatial predictions of heavy metals in wheat. *Journal of Environmental Quality*, 33:882–890.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46:167–174.

Christakos, G. (1990). A bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22:763–777.

Christakos, G. (2000). *Modern spatiotemporal geostatistics.* Oxford University Press, New York.

Christakos, G., Bogaert, P., and Serre, M. L. (2002). *Temporal GIS. Advanced Functions for Field-Based Applications.* Springer-Verlag, New York.

Christensen, R. (1991). *Linear Models for Multivariate, Time Series and Spatial Data.* Springer, New York.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, 2nd ed.* Wiley, New York.

de Marsily, G. (1986). *Quantitative Hydrogeology; Groundwater Hydrology for Engineers.* Academic Press, Orlando.

Delhomme, J. P. (1978). Kriging in the hydrosciences. *Advances in Water Resources*, 1:251–266.

Deutsch, C. V. and Journel, A. G. (1992). *GSLIB. Geostatistical Software Library and User's guide, 2nd ed.* Oxford University Press, Oxford.

D'Or, D. (2003). *Spatial prediction of soil properties, the Bayesian Maximum Entropy approach.* PhD thesis, Universite catholique de Louvain.

Finke, P. A., Brus, D. J., Bierkens, M. F. P., Hoogland, T., Knotters, M., and de Vries, F. (2004). Mapping groundwater dynamics using multiple sources of exhaustive high resolution data. *Geoderma*, 123:23–39.

Finke, P. A., D.Wladis, J.Kros, J.Pebesma, E., and J.Reinds, G. (1999). Quantification and simulation of errors in categorical data for uncertainty analysis of soil acidification modelling. *Geoderma*, 93:177–194.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford University Press, New York.

Griffith, D. A. and Layne, L. J. (1999). *A Casebook for Spatial Statistical Data Analysis.* Oxford University Press, New York.

Hartman, L. W. (2006). Bayesian modelling of spatial data using markov random fields, with application to elemental composition of forest soil. *Mathematical Geology*, 38:113–133.

Heuvelink, G. B. M., Brown, J., and van Loon, E. E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographic Information Science*, 21:497–513.

Jansen (ed), J. (2004). Kwaliteitsborging databestanden en modellen: Balanceren tussen chaotische dynamiek en geordende stilstand. Alterra Rapport 956, Alterra, Wageningen, the Netherlands.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

Kasetkasem, T., Arora, M. K., and Varshney, P. K. (2005). Super-resolution land cover mapping using a markov random field based approach. *Remote Sensing of Environment*, 96:302–314.

Klap, J. M., de Vries, W., and Leeters, E. E. J. M. (1999). Effects of acid atmospheric deposition on the chemical composition of loess, clay and peat soils under forest in the netherlands. Winand Staring Centre Report 97, Winand Staring Centre, Wageningen, the Netherlands.

Kros, J. (2002). *Evaluation of biogeochemical models at local and regional scale.* Alterra Scientific Contributions 7, Alterra, P.O. Box 47, 6700 AA, Wageningen, The Netherlands.

Kros, J., Pebesma, E. J., Reinds, G. J., and Finke, P. (1999). Uncertainty assessment in modelling soil acidification at the european scale: a case study. *Journal of Environmental Quality*, 28:366–377.

Lark, R. M. and Webster, R. (2006). Geostatistical mapping of geomorphic variables in the presence of trend. *Eart Surface Processes and Landforms*, 31:862–874.

M. Voltz, R. W. (1990). A comparison of kriging, cubic splines and classification for predicting soil propoerties from sample information. *Journal of soil Science*, 41:473–490.

Marsman, B. A. and de Gruijter, J. J. (1986). Quality of soil maps: a comparison of survey methods in a sandy area. Soil Survey Papers 15, Soil Survey Institute, Wageningen, the Netherlands.

Norberg, T., Rosen, L., Baran, A., and Baran, S. (2002). On modelling discrete geological structures as markov random fields. *Mathematical Geology*, 34:63–77.

Orton, T. G. and Lark, R. M. (2007). Estimating the local mean for Bayesian maximum entropy by generalized least squares and maximum likelihood, and an application to the spatial analysis of a censored variable. *European Journal of Soil Science*, 58:60–73.

Pebesma, E. J. and G.Wesseling, C. (1998). Gstat, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, 24:17–31.

Rao, C. R. (1973). *Linear statistical inference and its applications.* Wiley, New York.

Schouwenberg, E. P. A. G., Houweling, H., Jansen, M. J. W., Kros, J., and Mol-Dijkstra, J. P. (2000). Uncertainty propagation in model chains: a case study in nature conservation. Alterra Rapport 001, Alterra, Wageningen, the Netherlands.

Stein, A., Hoogerwerf, M., and Bouma, J. (1988). Use of soil-map delineations to improve (co-)kriging of point data on moisture deficits. *Geoderma*, 43:163–177.

Stichting voor Bodemkartering (1961-1995). *Bodemkaart van Nederland, schaal 1 : 50.000; toelichtingen bij de kaartbladen.* Pudoc, Wageningen, the Netherlands.

van Dobben, H. F., van Elswijk, M., Groben, M. S., Groenendijk, P., Houweling, H., Jansen, M. J. W., Mol-Dijkstra, J. P., Otjens, A. J., te Roller, J. A., Schouwenberg, E. P. A. G., and Wamelink, G. W. W. (2002a). Technische documentatie modellen raamwerk ecologie. Alterra Rapport 549, Alterra, Wageningen, the Netherlands.

van Dobben, H. F., Wamelink, G. W. W., Schouwenberg, E. P. A. G., and Mol, J. P. (2002b). Use of coupled models to predict biodiversity in managed ecosystems. In *Sustainable Forestry in Temperate Regions*, Proceedings of the SUFO International Workshop. April 7-9, 2002, Lund, Sweden, pages 76–85.

Van Groenigen, J. W., Siderius, W., and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, 87:239–259.

Verburg, P. H., de Nijs, T. C. M., van Eck, J. R., Visser, H., and de Jong, K. (2004). A method to analyse neighbourhood characteristics of land use patterns. *Computers, Environment and Urban Systems*, 28:667–690.

Visschers, R., A.Finke, P., and de Gruijter, J. (2007). A soil sampling program for the netherlands. *Geoderma*, 139:60–72.

Walvoort, D. J. and de Gruijter, J. J. (2001). Compositional kriging: a spatial ainterpolation method for compositional data. *Mathematical Geology*, 33:951–966.

Wamelink, G. W. W., ter Braak, C. J. F., and van Dobben, H. F. (2003). Changes in large-scale patterns of plant biodiversity predicted from environmental economic scenarios. *Landscape Ecology*, 18:513–527.

Wu, K., Nunan, N., Crawford, J. W., Young, I. M., and Ritz, K. (2004). An efficient markov chain model for the simulation of heterogeneous soil structure. *Soil Science Society of America Journal*, 68:346–351.

WOt-onderzoek

Verschenen documenten in de reeks Rapporten van de Wettelijke Onderzoekstaken Natuur & Milieu

WOt-rapporten zijn verkrijgbaar bij het secretariaat van Unit Wettelijke Onderzoekstaken Natuur & Milieu te Wageningen. T 0317 – 47 78 44; F 0317 – 41 90 00; E info.wnm@wur.nl
WOt-rapporten zijn ook te downloaden via de WOt-website www.wotnatuurenmilieu.wur.nl

1 *Wamelink, G.W.W., J.G.M. van der Greft-van Rossum & R. Jochem* (2005). Gevoeligheid van LARCH op vegetatieverandering gesimuleerd door SUMO

2 *Broek, J.A. van den* (2005). Sturing van stikstof- en fosforverliezen in de Nederlandse landbouw: een nieuw mestbeleid voor 2030

3 *Schrijver, R.A.M., R.A. Groeneveld, T.J. de Koeijer & P.B.M. Berentsen (2005).* Potenties bij melkveebedrijven voor deelname aan de Subsidieregeling Agrarisch Natuurbeheer

4 *Henkens, R.J.H.G., S. de Vries, R. Jochem, R. Pouwels & M.J.S.M. Reijnen, (2005).* Effect van recreatie op broedvogels op landelijk niveau; Ontwikkeling van het recreatiemodel FORVISITS 2.0 en koppeling met LARCH 4.1

5 *Ehlert, P.A.I. (2005).*Toepassing van de basisvrachtbenadering op fosfaat van compost; Advies

6 *Veeneklaas, F.R., J.L.M. Donders & I.E. Salverda (2006).*Verrommeling in Nederland

7 *Kistenkas, F.H. & W. Kuindersma (2005).* Soorten en gebieden; Het groene milieurecht in 2005

8 *Wamelink, G.W.W. & J.J. de Jong (2005).* Kansen voor natuur in het veenweidegebied; Een modeltoepassing van SMART2-SUMO2, MOVE3 en BIODIV

9 *Runhaar, J., J. Clement, P.C. Jansen, S.M. Hennekens, E.J. Weeda, W. Wamelink, E.P.A.G. Schouwenberg (2005).* Hotspots floristische biodiversiteit

10 *Cate, B. ten, H.Houweling, J. Tersteeg & I. Verstegen (Samenstelling) (2005).* Krijgt het landschap de ruimte? – Over ontwikkelen en identiteit

11 *Selnes. T.A., F.G. Boonstra & M.J. Bogaardt (2005).* Congruentie van natuurbeleid tussen bestuurslagen

12 *Leneman, H., J. Vader, E. J. Bos en M.A.H.J. van Bavel (2006).* Groene initiatieven in de aanbieding. Kansen en knelpunten van publieke en private financiering

13 *Kros, J, P. Groenendijk, J.P. Mol-Dijkstra, H.P. Oosterom, G.W.W. Wamelink (2005).* Vergelijking van SMART2SUMO en STONE in relatie tot de modellering van de effecten van landgebruikverandering op de nutriëntenbeschikbaarheid

14 *Brouwer, F.M, H. Leneman & R.G. Groeneveld (2007).* The international policy dimension of sustainability in Dutch agriculture

15 *Vreke, J., R.I. van Dam & F.H. Kistenkas (2005).* Provinciaal instrumentarium voor groenrealisatie

16 *Dobben, H.F. van, G.W.W. Wamelink & R.M.A. Wegman (2005).* Schatting van de beschikbaarheid van nutriënten uit de productie en soortensamenstelling van de vegetatie. Een verkennende studie

17 *Groeneveld, R.A. & D.A.E. Dirks (2006).* Bedrijfseconomische effecten van agrarisch natuurbeheer op melkveebedrijven; Perceptie van deelnemers aan de Subsidieregeling Agrarisch Natuurbeheer

18 *Hubeek, F.B., F.A. Geerling-Eiff, S.M.A. van der Kroon, J. Vader & A.E.J. Wals (2006).* Van adoptiekip tot duurzame stadswijk; Natuur- en milieueducatie in de praktijk

19 *Kuindersma, W., F.G. Boonstra, S. de Boer, A.L. Gerritsen, M. Pleijte & T.A. Selnes (2006).* Evalueren in interactie. De mogelijkheden van lerende evaluaties voor het Milieu- en Natuurplanbureau

20 *Koeijer, T.J. de, K.H.M. van Bommel, M.L.P. van Esbroek, R.A. Groeneveld, A. van Hinsberg, M.J.S.M. Reijnen & M.N. van Wijk (2006).* Methodiekontwikkeling kosteneffectiviteit van het natuurbeleid. De realisatie van het natuurdoel 'Natte Heide'

21 *Bommel, S. van, N.A. Aarts & E. Turnhout (2006).* Over betrokkenheid van burgers en hun perspectieven op natuur

22 *Vries, S. de & Boer, T.A. de, (2006) .* Toegankelijkheid agrarisch gebied voor recreatie: bepaling en belang. Veldinventarisatie en onderzoek onder in- en omwonenden in acht gebieden

23 *Pouwels, R., H. Sierdsema & W.K.R.E. van Wingerden (2006).* Aanpassing LARCH; maatwerk in soortmodellen

24 *Buijs, A.E., F. Langers & S. de Vries (2006).* Een andere kijk op groen; beleving van natuur en landschap in Nederland door allochtonen en jongeren

25 *Neven, M.G.G., E. Turnhout, M.J. Bogaardt, F.H. Kistenkas & M.W. van der Zouwen (2006).* Richtingen voor Richtlijnen; implementatie Europese Milieurichtlijnen, en interacties tussen Nederland en de Europese Commissie.

26 *Hoogland, T. & J. Runhaar (2006).* Neerschaling van de freatische grondwaterstand uit modelresultaten en de Gt-kaart

27 *Voskuilen, M.J. & T.J. de Koeijer (2006).* Profiel deelnemers agrarisch natuurbeheer

28 *Langeveld, J.W.A. & P. Henstra (2006).* Waar een wil is, is een weg; succesvolle initiatieven in de transitie naar duurzame landbouw .

29 *Kolk, J.W.H. van der, H. Korevaar, W.J.H. Meulenkamp, M. Boekhoff, A.A. van der Maas, R.J.W. Oude Loohuis & P.J. Rijk (2007).* Verkenningen duurzame landbouw. Doorwerking van wereldbeelden in vier Nederlandse regio's

30 *Vreke, J., M. Pleijte, R.C. van Apeldoorn, A. Corporaal, R.I. van Dam & M. van Wijk (2006).* Meerwaarde door gebiedsgerichte samenwerking in natuurbeheer?

31 *Groeneveld, R.A., R.A.M. Schrijver & D.P. Rudrum (2006).* Natuurbeheer op veebedrijven: uitbreiding van het bedrijfsmodel FIONA voor de Subsidieregeling Natuurbeheer

32 *Nieuwenhuizen, W., M. Pleijte, R.P. Kranendonk & W.J. de Regt (2007).* Ruimte voor bouwen in het buitengebied; de uitvoering van de Wet op de Ruimtelijke Ordening in de praktijk

33 *Boonstra, F.G., W.W. Buunk & M. Pleijte (2006).* Governance of nature. De invloed van institutionele veranderingen in natuurbeleid op de betekenisverlening aan natuur in het Drents-Friese Wold en de Cotswolds

34 *Koomen, A.J.M., G.J. Maas & T.J. Weijschede (2007).* Veranderingen in lijnvormige cultuurhistorische landschapselementen; Resultaten van een steekproef over de periode 1900-2003

35 *Vader, J. & H. Leneman (redactie) (2006).* Dragers landelijk gebied; Achtergronddocument bij Natuurbalans 2006

36 *Bont, C.J.A.M. de, C. van Bruchem, J.F.M. Helming, H. Leneman & R.A.M. Schrijver (2007).* Schaalvergroting en verbreding in de Nederlandse landbouw in relatie tot natuur en landschap.

37 *Gerritsen, A.L., A.J.M. Koomen & J. Kruit (2007)* .Landschap ontwikkelen met kwaliteit; een methode voor het evalueren van de rijksbijdrage aan een beleidsstrategie

38 *Luijt, J. (2007).* Strategisch gedrag grondeigenaren; Van belang voor de realisatie van natuurdoelen.

39 *Smits, M.J.W. & F.A.N. van Alebeek, (2007).* Biodiversiteit en kleine landschapselementen in de biologische landbouw; Een literatuurstudie.

40 *Goossen, C.M. & J. Vreke. (2007).* De recreatieve en economische betekenis van het Zuiderpark in Den Haag en het Nationaal Park De Hoge Veluwe

41 *Cotteleer, G., Luijt, J., Kuhlman, J.W. & C. Gardebroek, (2007*. Oorzaken van verschillen in grondprijzen. Een hedonische prijsanalyse van de agrarische grondmarkt.

42 *Ens B.J., N.M.J.A. Dankers, M.F. Leopold, H.J. Lindeboom, C.J. Smit, S. van Breukelen & J.W. van der Schans (2007).* International comparison of fisheries management with respect to nature conservation.

43 *Janssen, J.A.M. & A.H.P. Stumpel (red.) (2007).* Internationaal belang van de nationale natuur; Ecosystemen, Vaatplanten, Mossen, Zoogdieren, Reptielen, Amfibieën en Vissen

44 *Borgstein, M.H., H. Leneman, L. Bos-Gorter, E.A. Brasser, A.M.E. Groot & M.F. van de Kerkhof (2007).* Dialogen over verduurzaming van de Nederlandse landbouw. Ambities en aanbevelingen vanuit de sector

45 *Groot, A.M.E, M.H. Borgstein, H. Leneman, M.F. van de Kerkhof, L. Bos-Gorter & E.A Brasser (2007).* Dialogen over verduurzaming van de Nederlandse landbouw. Gestructureerde sectordialogen als onderdeel van een monitoringsmethodiek

46 *Rijn, J.F.A.T. van & W.A. Rienks (2007).* Blijven boeren in de achtertuin van de stedeling; Essays over de duurzaamheid van het platteland onder stedelijke druk: Zuidoost-Engeland versus de provincie Parma

47 *Bakker, H.C.M. de, C.S.A. van Koppen & J. Vader (2007).* Het groene hart van burgers; Het maatschappelijk draagvlak voor natuur en natuurbeleid

48 *Reinhard, A.J., N.B.P. Polman, R. Michels & H. Smit (2007).* Baten van de Kaderrichtlijn Water in het Friese Merengebied; Een interactieve MKBA vingeroefening

49 *Ozinga, W.A., M. Bakkenes & J.H.J. Schaminée (2007).* Sensitivity of Dutch vascular plants to climate change and habitat fragmentation; A preliminary assessment based on plant traits in relation to past trends and future projections

50 *Woltjer, G.B. (met bijdragen van R.A. Jongeneel & H.L.F. de Groot) (2007).* Betekenis van macro-economische ontwikkelingen voor natuur en landschap. Een eerste oriëntatie van het veld

51 *A.Corporaal, A.H.F.Stortelder, J.H.J.Schaminée en H.P.J. Huiskes (2007).* Klimaatverandering, een nieuwe crisis voor onze landschappen ?

52 *Oerlemans, N., J.A. Guldemond & A. Visser (2007).* Meerwaarde agrarische natuurverenigingen voor de ecologische effectiviteit van Programma Beheer. Ecologische effectiviteit regelingen natuurbeheer: Achtergrondrapport 3

53 *Leneman, H., J.J. van Dijk, W.P. Daamen & J. Geelen (2007).* Marktonderzoek onder grondeigenaren over natuuraanleg: methoden, resultaten en implicaties voor beleid. Achtergronddocument 'Evaluatie omslag natuurbeleid'

54 *G.L. Velthof & B. Fraters (2007).* Nitraatuitspoeling in duinzand en lössgronden.

55 *Broek, J.A. van den, G. van Hofwegen, W. Beekman & M. Woittiez (2007).* Options for increasing nutrient use efficiency in Dutch dairy and arable farming towards 2030; an exploration of cost-effective measures at farm and regional levels

56 *Melman, Th.C.P., C.Grashof-Bokdam, H.P.J.Huiskes, W. Bijkerk, J.E. Plantinga, Th.Jager, R. Haveman & A. Corporaal (2007).* Veldonderzoek effectiviteit natuurgericht beheer van graslanden. Ecologische effectiviteit regelingen natuurbeheer: Achtergrondrapport 2

57 *Massop, H.Th.L., J.G. Kroes, J. Hoogewoud, R. Pastoors, T. Kroon & P.J.T. van Bakel (2007).* Actualisatie Hydrologie voor STONE 2.3. Aanpassing randvoorwaarden en parameters, koppeling tussen NAGROM en SWAP, en plausibiliteitstoets

58 *Brus, D.J. & G.B.M. Heuvelink (2007).* Towards a Soil Information System with quantified accuracy. Three approaches for stochastic simulation of soil maps

wot

Wettelijke Onderzoekstaken Natuur & Milieu