Promoter **Prof. dr. ir. Marc Van Meirvenne**
Department of Soil Management
Faculty of Bioscience Engineering
Ghent University

Dean **Prof. dr. ir. Guido Van Huylenbroeck**

Rector **Prof. dr. Paul Van Cauwenberge**

Eef Meerschman

Multiple-point geostatistics for the reconstruction of complex spatial patterns in soil science

Thesis submitted in fulfilment of the requirements
for the degree of Doctor (PhD) in Applied Biological Sciences

Dutch translation of the title:
Meerpuntsgeostatistiek voor de reconstructie van complexe ruimtelijke patronen in de bodemkunde

Illustration on the cover:
Left: a polygonal network of ice-wedge pseudomorphs that was exposed during excavation at a study site in Sint-Niklaas, Belgium (© Marc Van Meirvenne, Universiteit Gent); right: a binary polygonal network training image.

Citation:
Meerschman, E. 2013. *Multiple-point geostatistics for the reconstruction of complex spatial patterns in soil science*, PhD thesis, Ghent University.

ISBN-number: 978-90-5989-631-4

Dankwoord

Deze eerste en vaak gelezen pagina van mijn doctoraat wil ik graag gebruiken om de mensen te bedanken die een grote rol hebben gespeeld tijdens mijn doctoraat. Dank aan/thanks to

mijn promotor Prof. Marc Van Meirvenne om me vijf jaar lang te hebben begeleid tijdens mijn thesis- en doctoraatsonderzoek, voor de twee boeiende onderwerpen die je hiervoor aanbracht, voor het delen van je enthousiasme voor geostatistiek, voor je motivatie en uitstekende hulp bij het uitschrijven van mijn onderzoeksresultaten en voor de ruimte die je me gaf om eigen keuzes te maken.

FWO-Vlaanderen voor het steunen van mijn doctoraatsonderzoek.

mijn ORBit-collega's, Valentijn, Islam, Timothy, Philippe, Ellen en Sam, voor de samenwerking, de vriendschap en de vele plezante momenten in ons bureau; oud-ORBit'ers, Udaya, Meklit en Liesbet, voor jullie hulp en luisterend oor tijdens mijn beginjaren en vakgroepcollega's, voor de collegiale samenwerking, de babbels tussendoor en de gezellige lunchmomenten.

the people I collaborated with, in particular Marie Cooreman, Gunther Ghysels, Murray Lark, Grégoire Mariethoz, Guillaume Pirot, Philippe Renard and Julien Straubhaar, for the fruitful discussions that taught me so much.

the members of the Examination Board for their thoughtful review of this thesis.

Annelies, Elena, Ellen en Sara, voor jullie vriendschap en de deugddoende lunchmomenten, mijn ouders, voor jullie trouwe rol achter de schermen die evolueerde van het verhuizen van mijn kot tot het opvangen van jullie kleinzoon en natuurlijk mijn twee mannen thuis, Kenneth en kleine Senn, binnenkort is er weer tijd voor jullie.

juni 2013
Eef Meerschman

Table of contents

List of abbreviations and acronyms

AUC	area under curve
ccdf	conditional cumulative distribution function
cdf	cumulative distribution function
CLT	continuous local trend
cpdf	conditional probability density function
DOE	depth of exploration
DS	Direct Sampling algorithm
ECa	apparent electrical conductivity
EMI	electromagnetic induction
IK	indicator kriging
MAEE	mean absolute estimation error
MEE	mean estimation error
MPG	multiple-point geostatistics
OK	ordinary kriging
PCLT	Poisson continuous local trend
pdf	probability density function
PVT	Poisson Voronoi Tessellation
RF	random function or random field
RMSEE	root mean squared estimation error
ROC	receiver operating characteristic
SISIM	sequential indicator simulation
TG	trans-Gaussian
TI	training image
TPG	two-point geostatistics

Samenvatting

In dit proefschrift wordt onderzocht welke meerwaarde meerpuntsgeostatistiek (MPG) kan bieden voor bodemkundige toepassingen. MPG is een recent ontwikkelde, geostatistische techniek die het variogram vervangt door een trainingsbeeld. De techniek werd ontwikkeld door aardolie- en hydrogeologen omdat zij vaak te maken krijgen met ruimtelijke patronen, zoals repetitieve, curvilineaire of verbonden patronen, die te complex zijn om met een variogramfunctie te modelleren. Bodemkundigen krijgen echter ook te maken met analoge complexe ruimtelijke patronen. Voorbeelden zijn paleogeulen, catena's, duinpatronen, lapiaz, krimpscheuren, permafrostpatronen, landgebruikspatronen, sedimentaire gesteentelagen en bodemporiën. MPG zou dus ook bruikbaar kunnen zijn in de bodemkunde, maar de techniek wordt momenteel nog niet toegepast.

Het kernidee van MPG is dat meerpuntsstatistieken nodig zijn om complexe ruimtelijke patronen te modelleren. Meerpuntsstatistieken relateren dezelfde variable op meer dan twee plaatsen tegelijk. Traditionele geostatistiek is gebaseerd op het variogram of de covariantiefunctie. Dit zijn tweepuntsstatistieken omdat ze dezelfde variable slechts op twee verschillende locaties relateren. Daarom gebruiken we in dit proefschrift de term *tweepuntsgeostatistiek* (TPG). MPG ontleent de vereiste meerpuntsstatistieken van een trainingsbeeld, omdat het aantal observaties vaak te beperkt is om er rechtstreeks meerpuntsstatistieken van af te leiden. Een trainingsbeeld is een conceptueel beeld van de verwachte ruimtelijke structuur dat vaak is opgebouwd uit voorkennis. Er werden verschillende MPG simulatie-algoritmes ontwikkeld, waarvan de meeste gebaseerd zijn op het principe van sequentiële simulatie. De belangrijkste verschillen van MPG t.o.v. TPG simulatie-algoritmes zijn dat de conditionele cumulatieve verdelingsfunctie voor elke locatie $\mathbf{x}$ wordt voorspeld door alle observaties (typisch tussen 20 en 100 observaties) in de omgeving van $\mathbf{x}$ gezamenlijk te beschouwen, in plaats van paarsgewijs en door het trainingsbeeld te scannen, in plaats van een kriging systeem op te lossen.

We verzamelden eerst een dataset van complexe bodempatronen. Polygonale patronen van fossiele ijswiggen zijn een duidelijk voorbeeld van complexe bodempatronen. De structuren zijn overblijfselen van thermische krimpscheuren die tijdens een ijstijd in permafrostbodems werden gevormd. We selecteerden een landbouwperceel in België met fossiele ijswiggen in de ondergrond op basis van een luchtfoto met polygonale gewassporen. Een klein deel (6 x 6-m) van het testgebied (0.63 ha) werd opgegraven waardoor een fossiele ijswig zichtbaar werd. De wiggen waren gevormd in kleirijke Tertiaire mariene sedimenten tijdens de laatste ijstijd, en werden later afgedekt door een 0.6 m dikke laag van eolische zandige sedimenten. We namen 94 bodemstalen (0.6 - 0.8 m) verdeeld over het testgebied en analyseerden de bodemtextuur. Er was een duidelijk

verschil tussen het Tertiair moedermateriaal (gemiddeld 21 % klei) en het Quaternaire wigmateriaal (gemiddeld 6 % klei). Het testgebied werd ook gescand met een proximale bodemsensor die de schijnbare elektrische geleidbaarheid (ECa) meet, wat resulteerde in een accuraat beeld van het polygonaal netwerk. Het was de eerste keer dat ondergrondse fossiele ijswiggen zo nauwkeurig in kaart werden gebracht met een proximale bodemsensor.

De lage ECa-waarden weerspiegelden het grover wigmateriaal en waren dus ruimtelijk verbonden, terwijl de hoge ECa-waarden eerder ruimtelijk geïsoleerd waren. De algemeen toegepaste benadering die steunt op een multi-Gaussiaanse random functie is niet geschikt om de connectiviteit van extreme waarden te reconstrueren. Daarom hebben we als alternatief een niet-Gaussiaans random functiemodel geselecteerd en het vermogen van dit model om de ECa-variable te reconstrueren geëvalueerd. We vertrokken van algemene kennis over de bodem (pedogenetische processen en hun relatie tot ECa) en enkele beschrijvende statistieken van de ECa-dataset om het model te selecteren en de parameters ervan te schatten. Het gefitte continu lokaal trend (CLT) model werd dan vergeleken met een trans-Gaussiaans (TG) model voor dezelfde data met behulp van een teststatistiek die de ruimtelijke connectiviteit van lage ECa-waarden weerspiegelde. Het CLT-model scoorde beter dan het TG-model en was dus beter geschikt om het onderzochte bodemproces te modelleren. Omdat het CLT-model de connectiviteit van de bodemvariable kon reconstrueren, is het bijvoorbeeld geschikt om geparameterizeerde trainingsbeelden aan te leveren, wat een nood is in het domein van MPG.

Vooraleer we MPG toepasten op de ijswigdataset, hebben we eerst een geschikt MPG-algoritme geselecteerd en zijn workflow en het belang en de gevoeligheid van de input parameters grondig onderzocht. We selecteerden het Direct Sampling (DS) algoritme als het meest geschikte MPG-algoritme om het polygonaal netwerk reconstrueren, omdat dit algoritme in staat is om categorische, continue en multivariate simulaties te genereren. De sterkte van het DS-algoritme is dat het meteen een waarde uit het trainingsbeeld toekent aan $\mathbf{x}$ in plaats van de volledige conditionele cumulatieve verdelingsfunctie te voorspellen. We voerden een gevoeligheidsanalyse uit op de belangrijkste invoerparameters door simulaties te genereren met behulp van zeven verschillende trainingsbeelden, waaronder een 3D-trainingsbeeld. Dit resulteerde in een leidraad voor het uitvoeren van MPG simulaties met het DS-algoritme en enkele aanbevelingen omtrent het selecteren van de invoerparameters.

Na het verzamelen van een geschikte dataset en het bestuderen van het DS-algoritme, voerden we een eerste MPG-reconstructie van het polygonaal netwerk uit. We beschouwden het proximale bodemsensorbeeld als referentiebeeld en bemonsterden hieruit een continue (655 sensormetingen) en een categorische (100 puntobservaties) dataset. We gebruikten twee verschillende continue trainingsbeelden: het eerste werd opgebouwd op

basis van de sensormetingen van een ander deel van het veld, terwijl het tweede werd opgebouwd op basis van het CLT-model. Als categorisch trainingsbeeld gebruikten we een geclassificeerde foto van een ijswignetwerk in Alaska. De resulterende MPG-voorspellingen reconstrueerden de polygonale patronen goed en kwamen overeen met het referentiebeeld. We besloten hieruit dat MPG een veelbelovende techniek is om complexe bodempatronen te voorspellen.

Het voorgaande experiment leerde ons ook dat een raster geïnterpoleerd van dicht bemonsterde metingen in een nabijgelegen gebied (het eerste continue trainingsbeeld) een geschikt en makkelijk op te bouwen trainingsbeeld is. We bouwden verder op dit principe en gebruikten het om proximale bodemsensormetingen te voorspellen in ontoegankelijke gebieden, die resulteren in hiaten in sensorbeelden. Voorbeelden van gebieden die moeilijk toegankelijk zijn voor een bodemsensor zijn gebieden met een dichte vegetatie, steenachtige gebieden, bebouwde zones of perceelsgrenzen. We gebruiken de aangrenzende, dicht bemonsterde gebieden tegelijk als conditionele data en als trainingsbeeld en spraken daarom van ‘trainingsdata’. De aanpak werd geëvalueerd op twee verschillende datasets: het ECa-beeld van de ijswigpolygonen en een tweede ECa-beeld van een begraven paleogeul. We maskeerden systematisch gebieden uit de ECa-beelden om vervolgens de ECa-waarden in deze gebieden te simuleren. We besloten dat de gesimuleerde ECa-waarden hetzelfde ruimtelijk patroon hadden als de aangrenzende gebieden. Wanneer de gemaskeerde zone klein was in verhouding tot de grootte van het te reconstrueren bodemfenomeen, kon een nauwkeurige voorspellingskaart gemaakt worden. De conditionele variatiecoëfficiënt werd gebruikt om extra puntobservaties te lokaliseren wat de voorspellingskwaliteit verder verbeterde.

Een recente ontwikkeling is multivariate MPG waarbij meerdere variabelen gelijktijdig gesimuleerd worden met behulp van een multivariaat trainingsbeeld. De ijswigdataset liet ons toe te onderzoeken of multivariate MPG kan gebruikt worden voor de verwerking van sensordata. We pasten bivariate MPG toe om een selectie van de sensormetingen te interpoleren naar een regelmatige grid (ervan uitgaande dat er geen ontoegankelijke gebieden waren) en tegelijkertijd te voorspellen wat de kans is om wigmateriaal in de bodem te vinden. We construeerden een bivariaat trainingsbeeld met een categorisch beeld van een willekeurig polygonaal netwerk als eerste variabele en een continu beeld van de bijbehorende sensorwaarden als tweede variabele. Om een kader te schetsen voor de evaluatie van de nieuwe methode, vergeleken we ze met een veel gebruikte procedure waarbij de sensormetingen worden geïnterpoleerd met ordinary kriging en vervolgens worden geclassificeerd met fuzzy *k*-means. Een vergelijking tussen de resulterende kaarten en de luchtfoto met de polygonale gewassporen, toonde aan dat MPG het polygonaal patroon beter reconstrueerde. De accuraatheid van de MPG-kaarten werd bewezen door een kwantitatieve validatie op basis van negen meetlijnen die niet werden gebruikt tijdens

de simulatie en de 94 bodemstalen. Bijgevolg kan multivariate MPG met succes gebruikt worden bij de verwerking van sensordata.

Tenslotte vertaalden we onze onderzoeksresultaten naar een industriële toepassing in samenwerking met het baggerbedrijf DEME N.V. Zij presenteerden ons een probleem waarmee baggerbedrijven dagelijks geconfronteerd worden, meerbepaald het reconstrueren van de sedimentaire lagen in een kanaal dat moet uitgebaggerd worden. Het voorspellen van de dikte van de sedimentaire lagen heeft directe economische gevolgen. We stelden zowel een TPG- als een MPG-procedure voor om dit probleem aan te pakken. De MPG-procedure bestond uit het gebruik van het IMPALA-algoritme en een eenvoudig categorisch 3D-trainingsbeeld, terwijl de TPG-aanpak gebaseerd was op sequentiële indicator simulatie. Beide stochastische geostatistische technieken leverden eerder gelijkwaardige prestaties, doordat het ruimtelijk patroon van de sedimentlagen niet al te complex was (het kon worden gemodelleerd met een variogram) en omdat er voldoende boringen beschikbaar waren. Beide benaderingen gaven betere resultaten dan een *nearest neighbour* interpolatie, de deterministische methode die momenteel wordt gebruikt door het bedrijf als gevolg van een gebrek aan tijd, budget en opleiding.

De algemene conclusie van dit proefschrift is dat MPG een innovatieve techniek is die een waardevolle bijdrage kan leveren tot de bodemkunde en meerbepaald de geostatistiek. Dit werd bewezen door de verschillende, succesvolle MPG-toepassingen in dit proefschrift. TPG en MPG zijn complementaire technieken en het is aan de gebruiker om te beslissen welke techniek het meest geschikt is om een specifiek probleem op te lossen. We besluiten niet dat MPG een betere methode is dan TPG, maar geloven wel dat het een flexibelere methode is.

Summary

In this thesis we assessed the potential of multiple-point geostatistics (MPG) to be applied in soil science. MPG is a recently developed geostatistical toolbox that replaces the variogram by a training image (TI). It has been developed by petroleum geologists and hydrogeologists because they often face spatial patterns, such as repetitive, curvilinear or connected features, that are too complex to be modelled with a variogram function. However, soil scientists also face complex spatial patterns. Examples are paleochannels, catenas, dune patterns, limestone pavement, desiccation cracks, patterned ground, land-use patterns, sedimentary rock layers and soil pores. Consequently, MPG might be of use to soil scientists as well, but its application has not yet been investigated.

The main idea of MPG is that modelling complex spatial patterns requires multiple-point statistics. Multiple-point statistics relate the same property at more than two locations at a time. Traditional geostatistics is based on the variogram or the covariance function, which are two-point statistics because they relate the same property only at two different locations. Therefore, we use the term *two-point geostatistics* (TPG) in this thesis. MPG generally derives the required multiple-point statistics from a TI, because the number of observations is often too limited derive the multiple-point statistics directly. A TI is a conceptual image of the expected spatial structure that is often built from prior knowledge. The MPG toolbox consists of different simulation algorithms. Most of these are based on the sequential simulation principle. The main differences between TPG and MPG simulation algorithms are that the latter estimate the conditional cumulative distribution function at each $\mathbf{x}$ by considering the neighbouring data jointly (typically between 20 and 100 neighbours), instead of pairwise, and by scanning the TI, instead of solving a kriging system.

We first collected a comprehensive data set to evaluate the applicability of MPG to reconstruct complex soil patterns. Polygonal patterns of ice-wedge casts are a clear example of such complex soil patterns. They are remnants of thermal contraction cracks that were formed in permafrost-affected soils during an ice age. We selected an agricultural field in Belgium with ice-wedge casts in the subsoil based on an aerial photograph showing polygonal crop marks. A small part (6 x 6-m) of the test area (0.63 ha) was excavated revealing an ice-wedge cast. The wedges penetrated clay-rich Tertiary marine sediments, covered by a 0.6 m layer of aeolian sandy sediments, and were associated with the permafrost during the last glacial period. We took 94 subsoil (0.6 – 0.8 m) samples distributed over the test area and analyzed their texture. The results showed a clear difference between the Eocene host material (on average 21 % clay) and the Quaternary wedge filling (on average 6 % clay). The test area was also surveyed with a proximal soil

sensor measuring the apparent electrical conductivity (ECa) which resulted in an accurate image of the polygonal network. It was the first time that buried ice-wedge polygons were imaged so accurately with a proximal soil sensor.

Low ECa values reflected the coarser wedge material and were thus strongly spatially connected, whereas the large ECa values were rather spatially isolated. The generally applied approach of assuming a multi-Gaussian random function is not suited to reconstruct the connectivity of extreme values. Hence, we selected as an alternative a non-Gaussian random function model and evaluated its capacity to reconstruct the ECa values. We used general knowledge about the soil (pedogenetic processes and the relationship to ECa) and some summary statistics of the ECa data set, to select the model and to estimate its parameters. We then compared the fitted continuous local trend (CLT) model with a trans-Gaussian (TG) model of the same data using a multiple-point test parameter reflecting the connectivity of small ECa values. The CLT model scored higher than the TG model and is therefore more appropriate for process modelling in this environment. Since the CLT model succeeds in capturing the multiple-point behaviour of the soil variable, it could be used to provide parameterized TIs, which is a need in the field of MPG.

Before applying MPG to the test data set, we first selected an appropriate MPG algorithm and thoroughly investigated its workflow and the importance and sensitivity of its input parameters. We selected the Direct Sampling (DS) algorithm to be the most appropriate MPG algorithm to reconstruct the polygonal network, because it can generate categorical, continuous and multivariate simulations. The strength of the DS algorithm is that it directly assigns a value from the TI to each $\mathbf{x}$ instead of predicting the entire conditional cumulative distribution function. We performed a sensitivity analysis on the most important input parameters by generating unconditional simulations using seven different TIs, including a 3D TI. This resulted in a comprehensive guide to performing multiple-point statistical simulations with the DS algorithm providing recommendations on how to set the input parameters appropriately.

After collecting an appropriate test data set and studying the DS algorithm, we applied a first MPG reconstruction of the polygonal network test data set. We considered the high-resolution proximal soil sensor image as our reference image, and extracted a continuous (655 sensor data) and a categorical (100 point observations) data set from it. We used two different continuous TIs: the first TI was built from the proximal soil sensor data of another part of the field, whereas the second TI was built from the CLT model. As categorical TI we used a classified photograph of an ice-wedge network in Alaska. The resulting MPG maps reconstructed the polygonal patterns well and corresponded closely to the reference image. Consequently, we identified MPG as a promising technique to map complex soil patterns.

The previous experiments showed that a grid interpolated from densely sampled measurements in a nearby field can serve as an appropriate and easy-to-build TI. We expanded on this principle and used it to predict proximal soil sensor measurements in inaccessible areas showing up as gaps in a proximal soil sensor image. Examples of areas that might remain unsampled are areas with a dense vegetation, stony areas, building areas or field boundaries. The neighbouring densely sampled areas are then used as both conditioning data and as TI, and are called 'training data'. This technique was evaluated on two different test cases: the ECa image of the ice-wedge polygons and a second ECa image of a buried paleochannel. We systematically blanked zones from the ECa images and simulated the ECa values in the blanked zones. We found that the simulated ECa values had similar spatial characteristics to the neighbouring areas. When the gaps were small relative to the size of the features being reconstructed, an accurate prediction map could be made. The conditional coefficient of variation was used as a guide to determine the location of extra point observations to improve the prediction quality further.

A recent development is multivariate MPG in which an ensemble of variables can be simulated simultaneously using a multivariate TI. We investigated whether multivariate MPG can be used for the processing of proximal soil sensor data using the ice-wedge data set. We applied bivariate MPG to interpolate a selection of the sensor measurements to a regular grid (assuming no inaccessible areas) and to derive simultaneously a map predicting the location of the ice-wedge casts. We built a bivariate TI with a categorical image of a random polygonal network as primary variable and a continuous image of the corresponding sensor values as secondary variable. To set a comprehensive framework for the evaluation of the new method's prediction performance, we compared it with the often-applied procedure of interpolating the sensor data with ordinary kriging and then performing a fuzzy *k*-means classification to derive the possibility of finding wedge material in the subsoil. A comparison between the resulting maps and the aerial photograph showing the ice-wedges through polygonal crop marks, showed that MPG reconstructed the polygonal patterns much better. The local accuracy of the MPG maps was proven by a quantitative validation based on nine measurement lines, that had not been used during simulation, and the 94 bore hole samples. Consequently, multivariate MPG can be used for the processing of proximal soil sensor data.

Finally, our research findings were translated into an industrial application in collaboration with the dredging firm DEME N.V. They presented a problem that is daily faced by dredging firms, i.e. the reconstruction of the depositional pattern of sedimentary layers in a channel to be dredged. Predicting the thickness of sedimentary layers has direct economic consequences. We presented both a TPG and MPG solution to solve this problem. The MPG approach consisted of using the IMPALA algorithm and a simple categorical 3D TI, whereas the TPG approach was based on sequential indicator

simulation. Both stochastic geostatistical approaches showed a rather equivalent performance. This is because the sedimentary layers did not have a very complex spatial pattern (it could be represented with a variogram model) and because there were sufficient bore hole samples. Both approaches gave better reconstruction results than a nearest neighbour interpolation, a deterministic interpolation method that is currently used by the company due to budget and time constraints and a lack of training.

The general conclusion of this thesis is that MPG is an innovative technique that can be a valuable part of the pedometrician's toolbox. This was proven by the different successful MPG applications throughout this thesis. We believe that TPG and MPG are complementary techniques and the user should select the technique that is best suited to solve the particular problem. We do not state that MPG is a better method than TPG, but we believe that it is more flexible.

Chapter 1

Introduction

1.1 Motivation

'The soil varies from place to place. This is what makes the soil so fascinating. We place the variety we observe on record, and we seek explanation for it. Were the soil uniform, we should simply acknowledge the fact and switch our attention to something more interesting.' Heuvelink and Webster (2001).

The first attempts to map soil variation were made by classifying the soil into discrete classes. Soil surveyors based their maps on a few, often expensive, observations and laboratory measurements, but especially on their knowledge of the soil and how the soil is related to geology, geomorphology, vegetation and landuse (Heuvelink and Webster, 2001).

Today, we mainly use quantitative methods to map soil variation. *Pedometrics* is an emerging field in soil science that collectively categorises all mathematical, statistical and numerical soil prediction methods. One of the most common methods for spatial prediction of soil properties currently in use is geostatistics (Lark, 2012a).

Geostatistics was developed in the mining industry in the 1960s and was first implemented in soil science by Burgess and Webster (1980). It is based on the random function theory: soil properties are modelled as if they were realizations of random functions (Matheron, 1965). The geostatistical toolbox consists of prediction and simulation algorithms, most of them based on the principle of kriging (Goovaerts, 1997). The cornerstone of traditional geostatistics is the variogram function which is used as a model of the spatial structure (Webster and Oliver, 2007).

Although the past decades have shown numerous successful applications of variogram-based geostatistics, the variogram has two main shortcomings. First, variogram modelling is generally data driven: a first estimate of the variogram function is calculated from the

differences between paired data values at increasing distances. Translating prior relevant soil knowledge into a variogram function can be difficult. Second, the variogram is a two-point statistic and therefore incapable to model all sorts of random functions. Especially the modelling of *complex* spatial patterns, as curvilinear, repetitive or connected features, is problematic. In this thesis, 'complex spatial patterns' refers to all spatial patterns that are too complex to be successfully modelled by a two-point statistic.

Complex spatial patterns frequently appear in soils. They are induced by an interaction of soil-forming factors and can be observed at scales ranging from landscape to microscopic. Examples are paleochannels, catenas, dune patterns, limestone pavement, desiccation cracks, patterned ground, land-use patterns, sedimentary rock layers and soil pores. Figure 1.1 shows some examples of complex soil patterns.

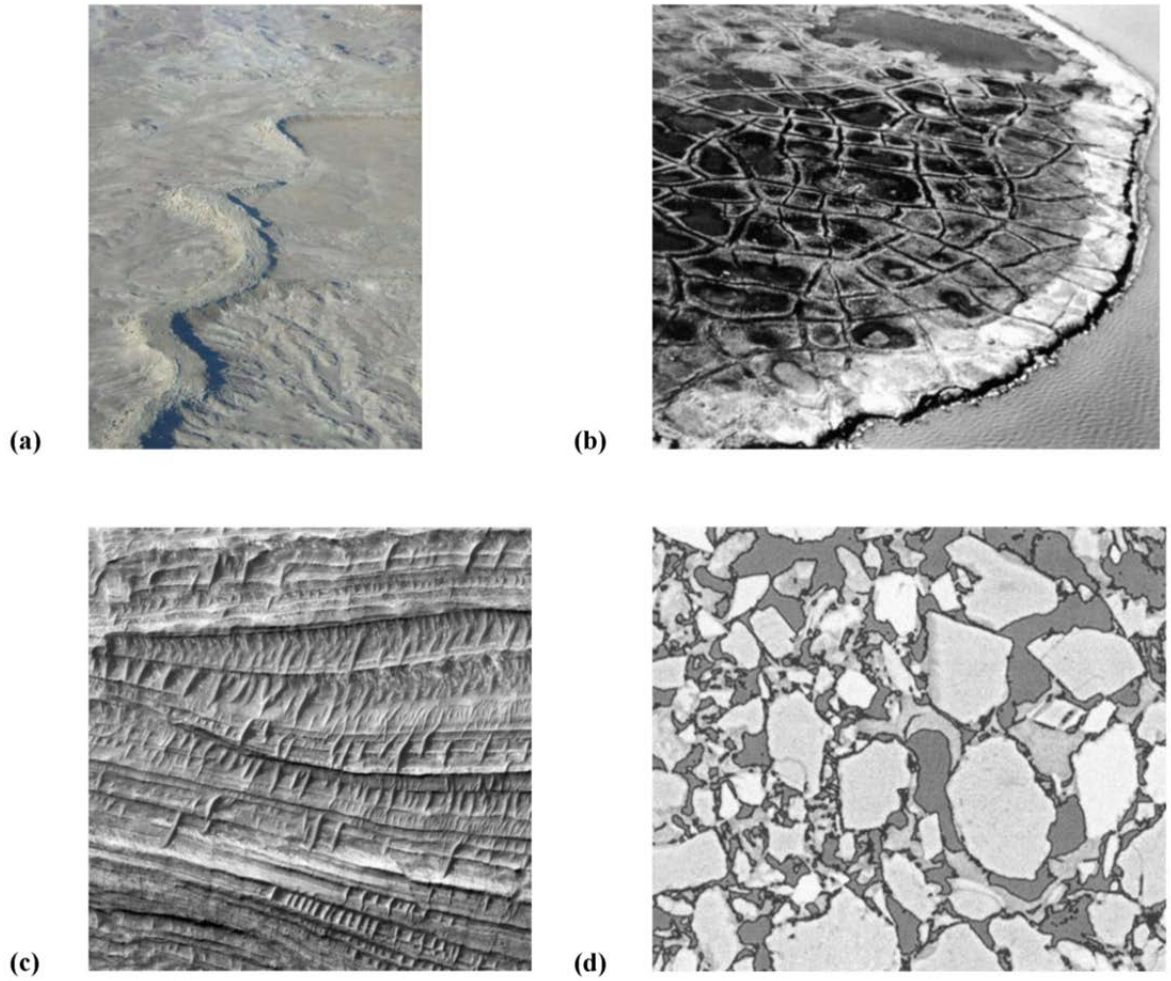


Figure 1.1 Examples of complex soil patterns: (a) aerial photograph of a paleochannel segment in Utah, USA (photograph: www.psi.edu); (b) aerial photograph of patterned ground due to frost action in Alaska (Jones et al., 2010); (c) HiRISE image of sedimentary layers in a valley on Mars (photograph: www.uahirise.org); (d) scanning electron microscope image of soil pores (Dathe et al., 2001).

Soil scientists might be interested in mapping these complex soil patterns or simulating them for probabilistic forecasting, for instance to predict infiltration rates. An often-used approach is complementing direct soil observations, as augerings or excavations, with high resolution indirect observations, such as air-borne, space-borne and proximal sensor or scanning data (McBratney et al., 2000). However, these data are not always available, may be of poor quality or are fragmented. Thus, getting a complete image of complex spatial structures requires an adequate geostatistical prediction or simulation method.

This method might be found in the field of oil and gas reservoir modelling, where scientists faced similar problems when simulating complex patterns. The lab of Strebelle (2002) at Stanford University (USA) developed multiple-point geostatistics (MPG). MPG covers a collection of simulation algorithms that use a training image (TI) as model of the spatial structure instead of a variogram function. A TI is a conceptual image of the expected spatial structure. A TI can be built from prior knowledge and can represent more complex spatial patterns than a variogram function, because it allows one to derive multiple-point statistics. To date, MPG has not been applied to soil science.

1.2 Research hypothesis and objectives

The research hypothesis of this thesis is that MPG can be used to reconstruct complex spatial patterns in soil science. To evaluate this research hypothesis, the following objectives were formulated.

1.2.1 Collect a test data set of complex soil patterns.

Evaluating a new methodology requires a good test case to set up experiments. Consequently, our first objective was to collect a comprehensive data set of complex soil patterns, consisting of both direct and indirect observations.

1.2.2 Fit a non-Gaussian model of spatial variation to the test data set.

Soil scientists generally rely on the multi-Gaussian random function and a set of predefined variogram models. However, it is known that this approach is not suited to reconstruct the connectivity of extreme values (Goovaerts, 1997), as is the case for the collected data set. Our second objective was to select an alternative, non-Gaussian model of spatial variation and to evaluate its capacity to reconstruct a multiple-point test parameter measuring the connectivity of extreme values.

1.2.3 Perform a sensitivity analysis on an appropriate MPG algorithm.

Although there is agreement about the fundamental principle of MPG, different algorithms have been developed to implement it. Because these are all very new, there is

little supporting material to help users getting started. Therefore, our third objective was to select the most appropriate algorithm, and to thoroughly investigate its workflow and the importance and sensitivity of its input parameters.

1.2.4 Evaluate the potential of MPG to reconstruct complex soil patterns using the test data set.

To date, most of the MPG applications can be found in the fields of petroleum geology and hydrogeology. Therefore, our fourth objective was to use the collected soil data set to set-up two straightforward experiments, i.e. a categorical and a continuous one, to evaluate the potential of MPG to reconstruct complex soil patterns. A focus here was finding an appropriate categorical and continuous soil TI. The latter is a challenging but crucial step in a MPG analysis.

1.2.5 Investigate whether MPG can be used for the processing of proximal soil sensor data.

Proximal soil sensing is an increasingly used data source for soil inventory. The typical sampling scheme of mobile soil sensing and the large density of the collected observations require adapted geostatistical procedures. Our fifth objective was to investigate whether MPG can be used for the processing of proximal soil sensor data. More specifically, we assessed if MPG could be used to interpolate the sensor data between measurements and in inaccessible areas, and if it could be used to predict the target variable of interest.

1.2.6 Evaluate the practical use of MPG in an industrial application.

Prediction or simulation methods that are successful at a field scale in 2D, cannot necessarily be extended to 3D studies at a landscape scale. Hence, the sixth objective was to apply the MPG approach in the context of an industrial application.

1.3 Structure of the thesis

The thesis is divided in ten chapters (Figure 1.2). Chapter 2 summarizes the state-of-the-art of two-point and multiple-point geostatistics and contains an overview of developed MPG algorithms. Chapters 3 explains how we collected a test data set of complex soil patterns by surveying a field with a polygonal network of ice-wedge casts in the subsoil using a proximal soil sensor. Chapter 4 presents a geometric model of spatial variation for polygonal soils that has been fitted to the test data set. Chapter 5 gives a detailed description of the Direct Sampling (DS) algorithm together with a sensitivity analysis on its input parameters. Chapter 6 shows a categorical and continuous MPG reconstruction of the polygonal network. Chapter 7 and 8 deal with the use of MPG to process proximal soil

sensor data. In chapter 9 MPG is used to reconstruct 3D sedimentary layers. The last chapter links the results to the originally defined research objectives. General conclusions are drawn concerning the applicability of MPG in soil science.

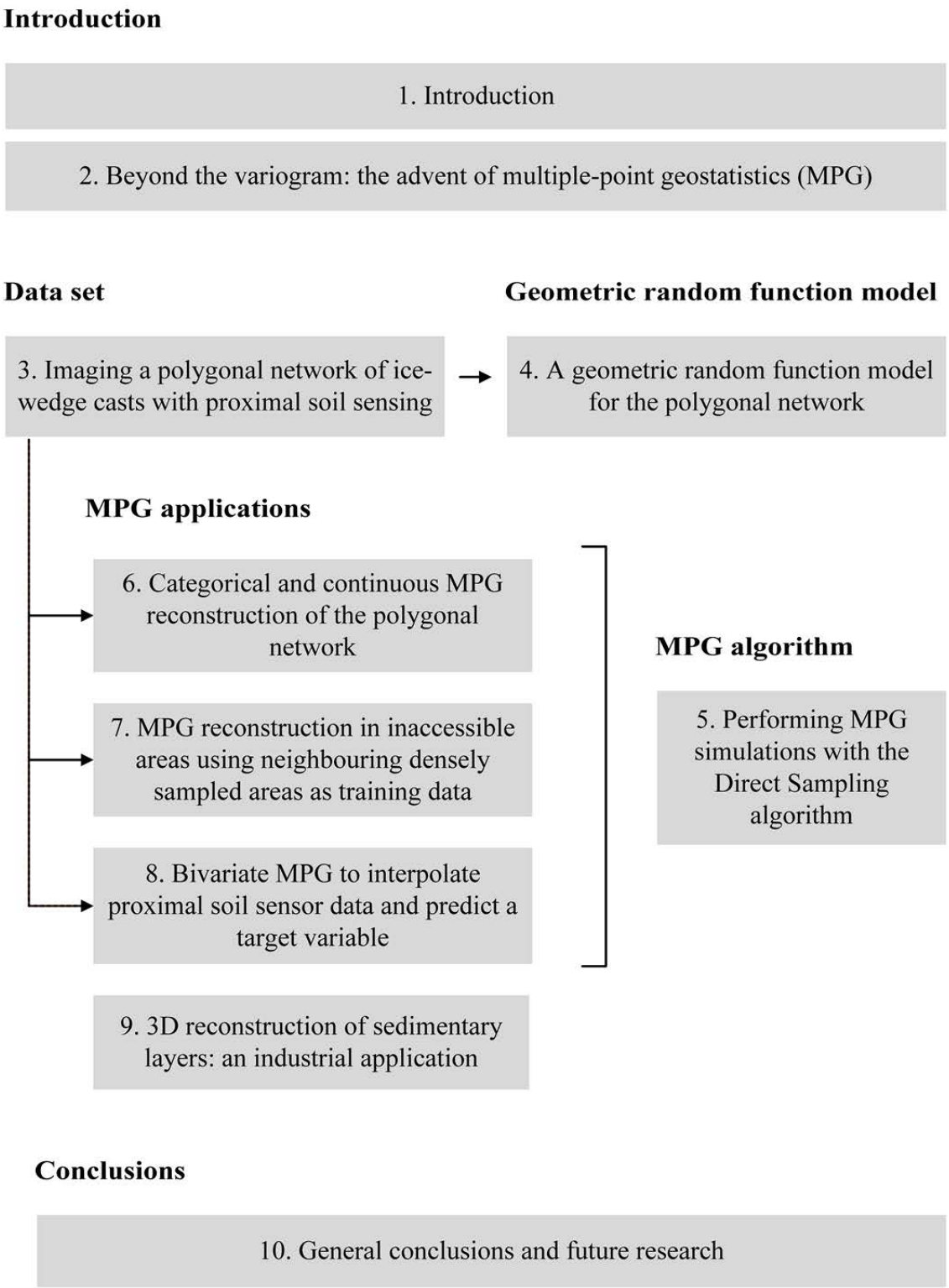


Figure 1.2 Structure of the thesis

Chapter 2

Beyond the variogram: the advent of multiple-point geostatistics

2.1 The random function model

The spatial prediction of soil properties requires a model of how the soil property behaves at unsampled locations. The probabilistic model that is at the root of geostatistics is the *random function model*. A probabilistic model recognises a fundamental uncertainty about how the soil property behaves at unsampled locations, in contrast with a deterministic model that assumes full knowledge about the soil process under study (Isaaks and Srivastava, 1989).

At each point in space $\mathbf{x}$ a property is treated as a *random variable* $\mathrm{Z}(\mathbf{x})$. This random variable is a model that represents the set of all possible values $\mathrm{z}(\mathbf{x})$ and their likelihood. Continuous random variables are characterized by their cumulative distribution function (cdf), which gives the probability that $Z(\mathbf{x})$ is no greater than any given threshold z:

$$F(\mathbf{x};z) = \mathrm{Prob}\{Z(\mathbf{x}) \leq z\} \in [0,1]. \qquad (2\text{-}1)$$

Its derivative is the probability density function (pdf) $f(\mathbf{x};z)$. Categorical random variables can only take a finite number K of values z_k with a probability of occurrence at $\mathbf{x}$:

$$p(\mathbf{x}, z_k) = \mathrm{Prob}\{Z(\mathbf{x}) = z_k\} \in [0,1]. \qquad (2\text{-}2)$$

A sampled observation $z(\mathbf{x}_\alpha)$ $(\alpha = 1,2,...)$ is assumed to be the outcome of a random variable. Predicting properties at unsampled locations $z^*(\mathbf{x})$ and assessing their uncertainty can thus be achieved by modelling the distribution function of $Z(\mathbf{x})$. This modelling is often restricted to assuming a parametric distribution and estimating its parameters, such as the mean and the variance for the frequently used Gaussian

distribution. However, it is important to note that a random variable is, in general, not fully described by a few parameters (Isaaks and Srivastava, 1989). Alternative approaches do not assume a theoretical distribution function of $Z(\mathbf{x})$ by applying a non-parametric indicator approach (Goovaerts, 1997) or by building algorithm-driven random variables through the generation of many alternative outcomes of the unknown (Remy et al., 2009).

The ensemble of all dependent random variables $\{Z(\mathbf{x}_1),...,Z(\mathbf{x}_N)\}$ at each $\mathbf{x}$ in the study area is called a *random function* or a *random field* (RF). Just as a random variable, a RF has several possible outcomes or realizations (maps), and is characterized by a distribution, known as the joint probability distribution function (Goovaerts, 1997):

$$F(\mathbf{x}_1,..., \mathbf{x}_N; z_1,..., z_N) = \text{Prob}\{Z(\mathbf{x}_1) \leq z_1,..., Z(\mathbf{x}_N) \leq z_N\}. \quad (2\text{-}3)$$

All geostatistical prediction or simulation methods require a decision of some degree of *stationarity*. This assumption allows one to infer statistics by pooling data over homogeneous areas. Strict stationarity implies that the joint probability distribution (Eq. 2-3) is independent of $\mathbf{x}$ (Goovaerts, 1997; Webster and Oliver, 2007).

2.2 Two-point geostatistics

Traditional geostatistics generally models the RF by specifying its variogram, or alternatively its covariance function, hereby weakening the stationarity decision to the first- and second-order moments (Goovaerts, 1997). Because the variogram and covariance function are two-point statistics, we will use the term *two-point geostatistics* (TPG) in the remaining of this thesis. The term 'two-point' refers to statistics that relate the same property at two different locations.

2.2.1 The covariance and variogram function

The covariance function describes the dependence between two random variables $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ separated by a vector $\mathbf{h}$:

$$C(\mathbf{h}) = \text{E}[(Z(\mathbf{x}) - \mu)(Z(\mathbf{x}+\mathbf{h}) - \mu)], \quad (2\text{-}4)$$

where μ is the mean $\text{E}(Z(\mathbf{x}))$ assumed to be stationary or constant for all $\mathbf{x}$.

The covariance cannot be defined when the mean changes over the study area. Therefore, Matheron (1965) weakened the condition of a stationary mean, and replaced the covariance by the variogram, that characterizes the spatial variation between two random variables $Z(\mathbf{x})$ and $Z(\mathbf{x}+\mathbf{h})$ separated by a vector $\mathbf{h}$:

$$\gamma(\mathbf{h}) = \frac{1}{2}\text{E}\left[(Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h}))^2\right]. \quad (2\text{-}5)$$

The use of a variogram does not assume that the mean is constant over the entire study area, but that it is at least constant for small **h**, so that $\mathrm{E}[Z(\mathbf{x})-Z(\mathbf{x}+\mathbf{h})]=0$. The less stringent conditions associated with the variogram make it much more useful than the covariance (Webster and Oliver, 2007).

The estimator of the variogram, known as Matheron's method of moments estimator (Matheron, 1962), is

$$\hat{\gamma}(\mathbf{h})=\frac{1}{2N(\mathbf{h})}\sum_{\alpha=1}^{N(\mathbf{h})}\{z(\mathbf{x}_\alpha)-z(\mathbf{x}_\alpha+\mathbf{h})\}^2, \qquad (2\text{-}6)$$

where $N(\mathbf{h})$ is the number of data pairs $\{z(\mathbf{x}_\alpha), z(\mathbf{x}_\alpha+\mathbf{h})\}$. A plot of the calculated $\hat{\gamma}(\mathbf{h})$ values versus **h** is called an experimental variogram. The theoretical variogram model is a continuous variogram function fit to the experimental variogram allowing to deduce variogram values for any possible **h** (Figure 2.1) (Webster and Oliver, 2007).

As can be seen in Figure 2.1, variogram values typically increase for larger **h**. This behaviour reflects Tobler's first law of geography: *'Everything is related to everything else, but near things are more related than distant things'* (Tobler, 1970). In most situations, the variogram stabilises around a maximum, which is generally the sum of the sill C_1 and the nugget effect C_0. The nugget effect is a discontinuity at the origin of the variogram that arises from measurement errors or spatial sources of variation at distances smaller than the shortest sampling interval. The distance at which the plateau is reached is the range *a* and can be interpreted as the distance of dependence or zone of influence of the property (Journel and Huijbregts, 1978).

Only functions that are conditionally negative definite can be considered as variogram models, so most geostatistical analyses are limited to a few permissible models, including the spherical, exponential or Gaussian model, or a combination of them (Goovaerts, 1997).

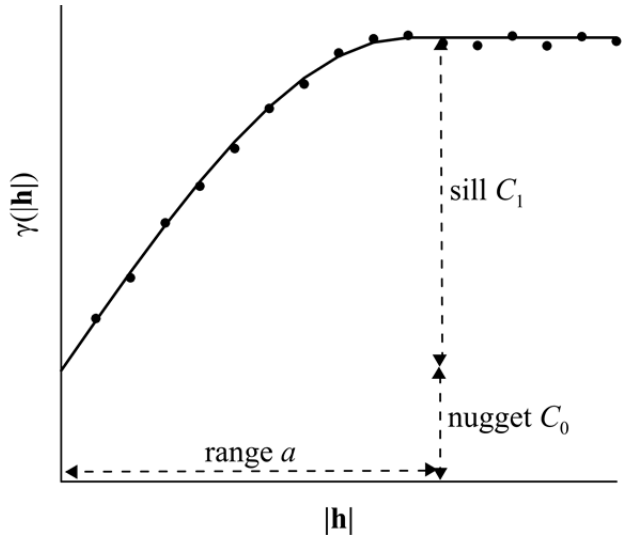


Figure 2.1 An experimental variogram (black dots) and a fitted theoretical exponential variogram model (solid line) with indication of the range *a*, the nugget C_0 and the sill C_1.

2.2.2 Two-point geostatistical prediction and simulation methods

The two-point geostatistical (TPG) toolbox consists of prediction and simulation techniques. Most of these techniques are based on the principle of kriging (Krige, 1951; Matheron, 1962). Kriging is a best linear unbiased predictor. It is *linear* because its predictions are weighted linear combinations of the sampled observations:

$$z^*(\mathbf{x}) = \sum_{\alpha=1}^{n(\mathbf{x})} \lambda_\alpha z(\mathbf{x}_\alpha), \qquad (2\text{-}7)$$

where λ_α are the kriging weights and $n(\mathbf{x})$ is the number of observations in the search neighbourhood. Kriging is *unbiased* because the expected value of the prediction error is zero, and it is *best* because the variance of the prediction error is minimized (Isaaks and Srivastava, 1989). The set of kriging weights λ_α that produces these best linear unbiased predictions can be found by solving the kriging equation system. The kriging system includes a matrix with data-to-unknown variogram values and data-to-data variogram values, causing the ability of kriging to induce declustering (Remy et al., 2009).

There exist different types of kriging, of which ordinary kriging (OK) is the most common algorithm to predict continuous variables and indicator kriging (IK) the most common for categorical variables. The latter is applied to binary indicators of occurrence of a category z_k.

$$i(\mathbf{x}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{x}_\alpha) = z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1,..., K \qquad (2\text{-}8)$$

IK can be used to predict continuous variables when the event k is that $Z(\mathbf{x})$ is valued below a given threshold z_k. Advantages of IK are that one can use separate indicator variogram models for each $i(\mathbf{x}_\alpha; z_k)$ accounting for class-specific patterns of spatial variation, and that the kriging predictions can be directly interpreted as predicted probabilities for z_k to occur at $\mathbf{x}$. In other words, they directly model the pdf of the discrete random variable (Goovaerts, 1997; Remy et al., 2009).

Prediction maps are created by calculating the best kriging prediction $z^*(\mathbf{x})$ for each $\mathbf{x}$ in a regular grid, without considering any neighbouring prediction. Kriging prediction maps are therefore smoothed in their spatial variation and do not represent the true spatial variation properly (Goovaerts, 1997). In other words, kriging prediction maps are locally accurate but the true spatial pattern may be poorly reproduced.

Simulation methods, on the other hand, aim at a more realistic reproduction of the spatial pattern by generating multiple realizations. Each realization represents the RF model, or the joint distribution in space of all random variables $Z(\mathbf{x})$. The ensemble of all simulated values for each $\mathbf{x}$ represents the distribution function of the random variable, from which, for instance, the local mean (E-type) can be derived. Depending on the

applied algorithm, realizations can honour the conditioning data (observations) and can reproduce the sample histogram and variogram. Most TPG simulation algorithms are sequential simulation algorithms.

Sequential simulation algorithms generate realizations by visiting the unsampled grid nodes **x** successively along a random path. At each **x**, a conditional cumulative distribution function (ccdf) is estimated:

$$F(\mathbf{x}; z|(n)) = \text{Prob}\{Z(\mathbf{x}) \leq z|(n)\}, \quad (2\text{-}9)$$

where $|(n)$ expresses the conditioning to local information, i.e. the conditioning data and the previously simulated grid nodes. From this ccdf a simulated value is drawn and the algorithm proceeds to the next **x**. The realization is finished when each grid node has been visited. To generate a next realization, the algorithm defines a new random path (Goovaerts, 1997).

TPG simulation methods build the ccdfs by considering the *n* neighbouring data one by one: for each **x**, a kriging system is solved based on two-point variogram values. Sequential Gaussian simulation (SGS) assumes a multi-Gaussian RF: the ccdfs are Gaussian-shaped and their mean and variances are derived from a (simple or ordinary) kriging system. Sequential indicator simulation (SISIM) determines the ccdfs by applying IK (the indicator formalism) (Goovaerts, 1997) and can thus be used to simulate categorical variables (Emery, 2004).

2.3 Multiple-point geostatistics

About a decade ago, Caers and Zhang (2004) published Figure 2.2. It shows three different geological patterns together with their variogram. The similarity of the variograms proves that a two-point statistic is not enough to characterize the geometry of different geological patterns. Therefore one should consider statistics that relate the same property at more than two locations at a time, or *multiple-point statistics*.

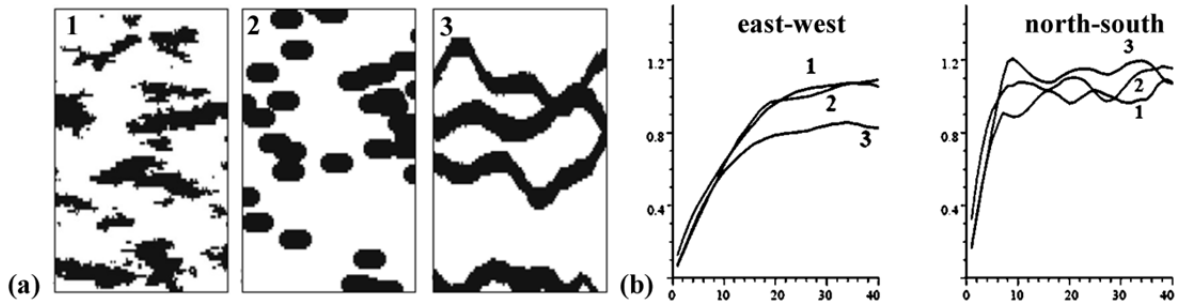


Figure 2.2 (a) Three different spatial patterns with (b) similar east-west and north-south variograms (Caers and Zhang, 2004).

2.3.1 The principle behind multiple-point geostatistics

In MPG the two-point variogram is replaced by a multiple-point TI as model of the spatial structure. A TI is a conceptual image of the expected spatial structure reflecting the spatial dependence between multiple points (Guardiano and Srivastava, 1993). Possible TIs are for instance hand drawings, model outputs or existing maps or photographs that are assumed to be analogues with the phenomenon under study. Alternatively, a TI can be constructed from more densely sampled zones (Goovaerts, 1997). A good TI reflects the prior structural concept (Hu and Chugunova, 2008), allowing the user to provide information about the shape of the structures and their degree of connectivity (Straubhaar et al., 2011). A TI does not need to carry any local information about the studied phenomenon (Strebelle et al., 2003). Remy et al. (2009) defined a TI as a representation of how the random variables $Z(\mathbf{x})$ are jointly distributed in space, or an unconditional realization of the joint probability distribution (Eq. 2-3). Figure 2.3 shows a –now famous– example of a binary TI representing a channel system (Strebelle, 2002), together with one MPG realization using this TI (Caers and Zhang, 2004).

The MPG toolbox consists of different simulation algorithms. Most of these are based on the sequential simulation principle as explained in 2.2.2. The main differences between TPG and MPG simulation algorithms are that the latter estimate the ccdf at each $\mathbf{x}$ by considering the n (typically between 20 and 100) neighbouring data jointly, instead of pairwise, and by scanning the TI, instead of solving a kriging system. Consequently, MPG algorithms aim at generating realizations that honour the conditioning data and the multiple-point statistics of the TI, instead of just the histogram (a one-point statistic) and the variogram (a two-point statistic). In other words, MPG simulation algorithms anchor the multiple-point patterns of the TI to the conditioning data (Caers and Zhang, 2004).

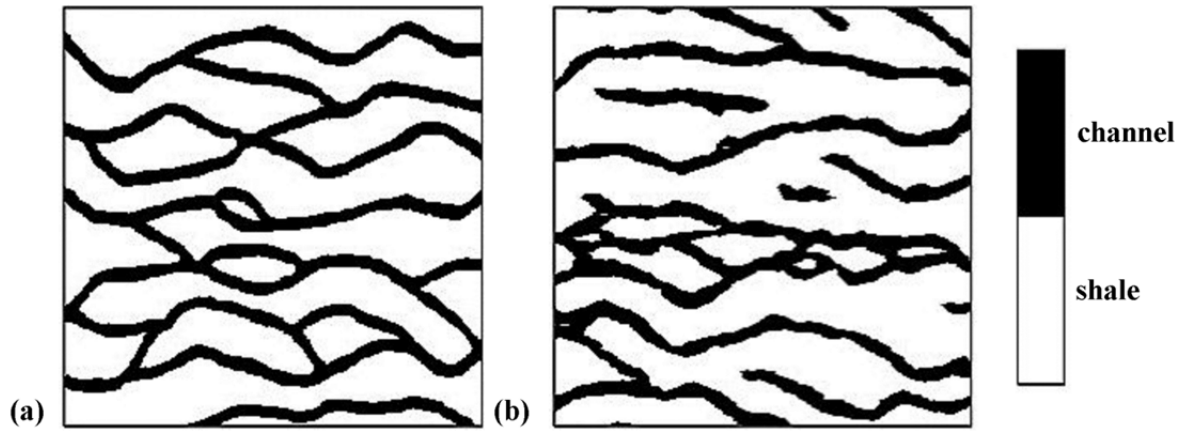


Figure 2.3 Example of (a) a binary TI representing a channel system (Strebelle, 2002), and (b) one MPG realization based on this TI (Caers and Zhang, 2004).

For each successively visited **x,** a data event $\mathbf{d}_n(\mathbf{x})$ of size *n* centred at location **x** is defined (Figure 2.4). This data event consists of the *n* neighbouring data values $z(\mathbf{x}+\mathbf{h}_\alpha)$ and the neighbouring data geometry, defined by the *n* vectors $\mathbf{h}_\alpha$ $(\alpha = 1,...,n)$.

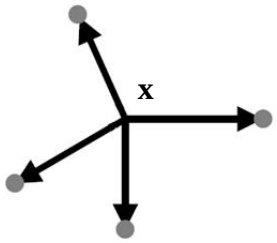


Figure 2.4 Example of a data event $\mathbf{d}_4(\mathbf{x})$ of size 4 centred at location $\mathbf{x}$.

Then, the TI is scanned for replicates of $\mathbf{d}_n(\mathbf{x})$. The TI scan is based on the principle that $p^*(\mathbf{x}; z_k|(n))$ corresponds to the ratio of the number of replicates with their central node value equal to z_k and the total number of replicates found, known as the Bayes relation of conditional probability (Strebelle, 2002).

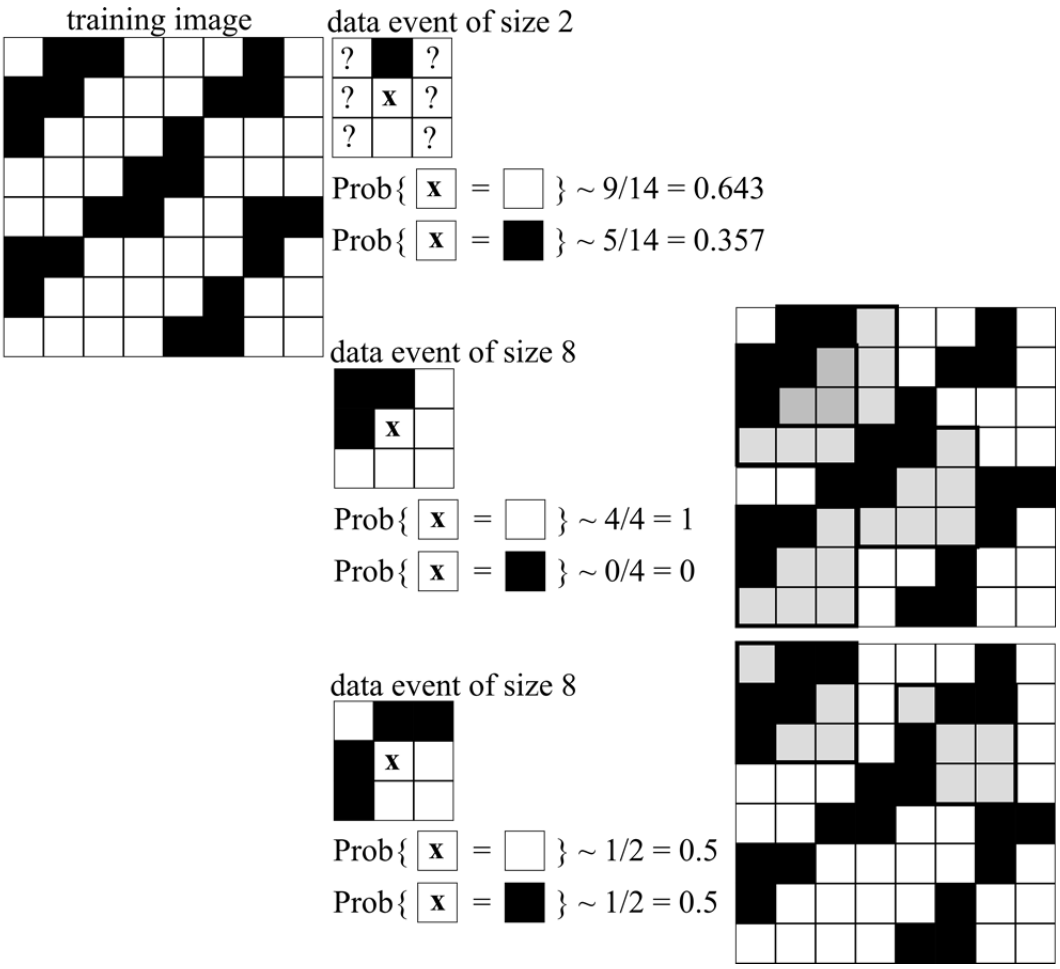


Figure 2.5 Principle of the TI scan: the training image (left), three different data events centred around **x** (centre) and the training image with the found replicates indicated in grey for the second and the third data event.

Figure 2.5 shows the principle of the TI scan. Here, the data event is limited by a rectangular search window and consists of two neighbouring data values in the first case, and eight in the second and the third case. For instance, the TI has 14 replicates of the data event of size 2. The probability to find a white pixel at **x** is thus 9/14 because 9 replicates have a white central node, and the probability to find a black pixel at **x** is 5/14 because 5 replicates have a black central node. The same principle holds for the other two data events.

MPG relies on a stationarity decision that allows one to derive multiple-point statistics from the chosen TI. This stationarity decision is not stronger than for TPG. Any mapping algorithm requires the same amount of multiple-point statistics. The often-used sequential Gaussian simulation algorithm, for instance, seems to require only a histogram and a variogram. However, the assumption that all higher-order statistics are multivariate Gaussian is included in the algorithm (Caers and Zhang, 2004). A limitation of this multi-Gaussian RF model is that it does not allow for any significant spatial correlation between small or large values (Goovaerts, 1997). This maximum disconnectivity of extreme values is known as the 'maximum entropy' property. As a consequence, MPG simulations strongly depend on the chosen TI (Strebelle, 2002), but this sensitivity is not stronger than the sensitivity to the combination of a variogram model and implicit high-order assumptions (Journel and Zhang, 2006).

Using a TI instead of a variogram enables thus to model more complex spatial patterns, such as curvilinear or connected features. This is particularly important to model flow and transport processes. A second advantage is that a TI can be constructed more straightforwardly from expert knowledge because it connects more closely to reality than a variogram function. Although attempts have been made to translate prior conceptual knowledge of the studied phenomenon into a variogram function (Truong et al., 2012), it is easier to translate these concepts into an image (Journel and Zhang, 2006).

2.3.2 Multiple-point geostatistical algorithms

Guardiano and Srivastava (1993), who proposed the basic idea of using multiple-point statistics, developed a first MPG sequential simulation algorithm for categorical variables, and called it ENESIM. ENESIM re-scans the entire TI at each **x** to predict $p^*(\mathbf{x}; z_k|(n))$. Unfortunately, this strategy is not practically implementable since it requires a long computing time.

Strebelle (2002) offered a practical solution with his SNESIM code, which is an improved version of the algorithm of Guaridano and Srivastava (1993). Prior to simulation, SNESIM scans the entire TI with a search template and stores all possible TI replicates in a dynamic data structure called a search tree. During simulation, this catalogue is used to compute $p(\mathbf{x}; z|(n))$ at each **x**.

A second improvement of SNESIM is its multi-grid approach that ensures a good reproduction of the patterns at different scales while keeping the size of the search template rather small (Strebelle, 2002). More information about the multi-grid approach can be found in chapter 9. To date, SNESIM is still a popular MPG algorithm (e.g. Huysmans and Dassargues, 2011; Le Coz et al., 2011; Ronayne et al., 2008) that is implemented in the freely distributed SGeMS software (Remy et al., 2009). Liu (2006) presents a practical guide to SNESIM together with a sensitivity analysis on its input parameters.

During the last decade, MPG has become an active research topic and several alternative MPG algorithms have been developed (Hu and Chugunova, 2008). We limit our discussion to two recent MPG algorithms that are further applied in this thesis: the IMPALA algorithm (Straubhaar et al., 2011) and the DS algorithm (Mariethoz et al., 2010).

The IMPALA algorithm stores the catalogue of possible TI replicates in lists instead of tree structures. The main advantage of the list approach is a significant reduction in memory usage (Straubhaar et al., 2011). The IMPALA algorithm is included in the commercial Isatis software (Bleinès et al., 2011). More details about the IMPALA code are given in Chapter 9.

Because all possible TI replicates are stored beforehand in a catalogue, both SNESIM and IMPALA can only be used to simulate categorical variables. The first MPG sequential simulation technique that enables simulating different variable types, including categorical, continuous and multivariate variables, is the Direct Sampling (DS) code (Mariethoz et al., 2010). DS re-scans the TI for each $\mathbf{x}$ during sequential simulation, as was first proposed by Guardiano and Srivastava (1993), but it directly samples the TI without explicitly modelling the ccdfs. Up to date, DS has not been implemented in a software package, but the code is available for academic purposes. Chapter 5 gives a detailed analysis of the possibilities offered by DS.

For completeness, we mention that MPG is not only an alternative to variogram-based geostatistics but also to Boolean object-based techniques (e.g. Deutsch and Wang, 1996). A main shortcoming of object-based techniques is that conditioning them to dense data sets is virtually impossible (Caers and Zhang, 2004). There is also an evolution in the development of pattern-based MPG techniques, such as SIMPAT (Arpat and Caers, 2007) and FILTERSIM (Zhang et al., 2006b). Pattern-based MPG techniques can be considered as a kind of object-based techniques (Hu and Chugunova, 2008). Neither pattern-based or object-based techniques are discussed in this thesis.

Chapter 3

Imaging a polygonal network of ice-wedge casts with proximal soil sensing

The content of this chapter is based on: Meerschman, E., Van Meirvenne, M., De Smedt, P., Saey, T., Islam, M.M., Meeuws, F., Van De Vijver, E. and Ghysels, G. 2011. Imaging a polygonal network of ice-wedge casts with an electromagnetic induction sensor. *Soil Science Society of America Journal* 75, 2095–2100.

This chapter explains how we collected a test data set of complex soil patterns. We selected an agricultural field with a polygonal network of ice-wedge casts in the subsoil, being a typical example of a geometrically complex soil pattern. The field was surveyed with a proximal soil sensor. The indirect observations were complemented with bore holes and an excavation.

3.1 Introduction

In many parts of the mid-latitudes of the northern hemisphere, the past existence of peri-glacial conditions is evidenced by the presence of ice-wedge casts and relic sand wedges (French, 2007). These cryogenic structures are the remnants of thermal contraction cracks formed in permafrost-affected soils (Kolstrup, 1986). Progressive infilling of these cracks with ice, sand or both, resulted in wedge-shaped bodies of ice, sand or sand-ice (French et al., 2003; Ghysels and Heyse, 2006; Murton and French, 1993; Vandenberghe and Pissart, 1993). When changing climatic conditions caused the permafrost to thaw, the

wedge-shaped cavities were filled with wind- and water-transported sediments resulting in their preservation as ice-wedge casts or ice-wedge pseudomorphs (Harry and Gozdzik, 1988). Consequently, the wedge filling has a different composition than the host material.

The surface expression of thermal contraction cracks is generally a network of polygons, still observable in modern periglacial environments at high latitudes (French, 2007). In central Europe and North-America, polygonal networks of ice-wedge casts were often covered by eolian or fluviatile loess or sand, so their pattern is rarely directly observable. However, the morphology of these polygonal networks provides valuable information about past environmental and climatic conditions, since their formation depend on many factors such as soil temperature gradients, mineral composition of the soil, moisture content and variations in air temperature (Dutilleul et al., 2009; Mackay and Burn, 2002; Plug and Werner, 2002; 2008; Romanovskij, 1973). Apart from imaging the ice-wedge casts for paleoclimatological reconstructions, characterizing their abrupt changes in soil composition can suit other purposes. Ice-wedge casts can have an impact on engineering projects (Morgan, 1971), preferential flow paths for leaching to groundwater (Dansart et al., 1999) and on crop yield calling for techniques known as precision agriculture.

Occasionally, near-surface networks of pseudomorphs show up on aerial photographs of cultivated fields due to color contrasts of the crop, called crop marks. These are caused by pedological differences between host material and wedge filling. However, the occurrence of crop marks is very sensitive to variations in soil type, soil moisture content, vegetation type, nutrient availability and meteorological conditions (Walters, 1994). Therefore, the time frame for such observations is often very narrow and the costs of obtaining them are large. In the particular situations where crop marks reveal the presence of ice-wedge casts, aerial photographs can be used to map the polygonal network morphology (Ghysels, 2008; Lusch et al., 2009).

Near-surface geophysical prospection methods are an alternative for mapping polygonal networks of ice-wedge casts. A few studies have shown the use of ground-penetrating radar (Dansart et al., 1999; Doolittle and Nelson, 2009) and electrical resistivity (Lusch et al., 2009) to detect relic ice-wedges. Cockx et al. (2006) were the first to map near-surface Pleistocene ice-wedge casts with an electromagnetic induction (EMI) sensor. However, their survey covered a small excavated area where the casts were visible at the surface.

3.2 Material and methods

3.2.1 Aerial photograph and test area

Figure 3.1a shows a part of an oblique aerial photograph of an agricultural field in Deinze, Belgium (central coordinates: 51°01'16"N, 3°29'41"E). The field is situated on the West Flanders plateau, a low-lying plateau (25 m above sea level) next to the Coastal Plain.

The photograph was taken on 4 August 1996 when sugar beets were cultivated on the field. Notice that the polygonal pattern was not visible on adjacent fields with a different crop. An aerial photograph of the same field but with a different crop taken one year later did not show the polygonal pattern, demonstrating the ephemeral character of crop marks (Ghysels, 2008). It is our experience that due to their deep rooting system, sugar beets often develop good crop marks. Besides the polygonal pattern, the aerial photograph also shows a former field track, crossing the field from north to southeast. The photograph was georeferenced and color stretched to enhance the contrasts, after which a test area of 0.63 ha was selected and clipped (Arcmap 9.3, ESRI) (Figure 3.1b).

3.2.2 Electromagnetic induction survey and data processing

The test area was surveyed with a Geonics EM38DD sensor which simultaneously measures the apparent electrical conductivity (ECa) in a horizontal (ECa-H) and vertical (ECa-V) dipole mode. With a fixed inter-coil spacing of 1 m, each coil pair has its own depth-response curve (McNeill, 1980). The depth of exploration (DOE), defined as the depth where 70 % of the response is obtained from the soil volume above this depth, is 0.76 m for ECa-H and 1.55 m for ECa-V (Saey et al., 2009a). Characteristic for ECa measurements is their strong relationship with soil texture in the absence of salinity (Cockx et al., 2007; Corwin and Lesch, 2005).

The sensor was mounted on a sled pulled by an all terrain vehicle, which drove along parallel lines with an inter-line distance of on average 0.75 m at a speed from 4 to 6 km h^{-1}. The ECa was measured with a frequency of 10 Hz and the data were recorded by a field computer. A Trimble AgGPS332, with Omnistar correction, was used to georeference the ECa measurements with a pass-to-pass accuracy of approximately 0.10 m (Saey et al., 2009a). The survey was conducted on 9 April 2010 during dry weather conditions on a bare field with no soil tillage since October 2009.

The ECa measurements were post-corrected for instrumental drift (Simpson et al., 2009) and standardized to a reference temperature of 25°C (Slavich and Petterson, 1990). A Gaussian low pass filter was applied to the data for noise removal using SGeMS (Remy et al., 2009).

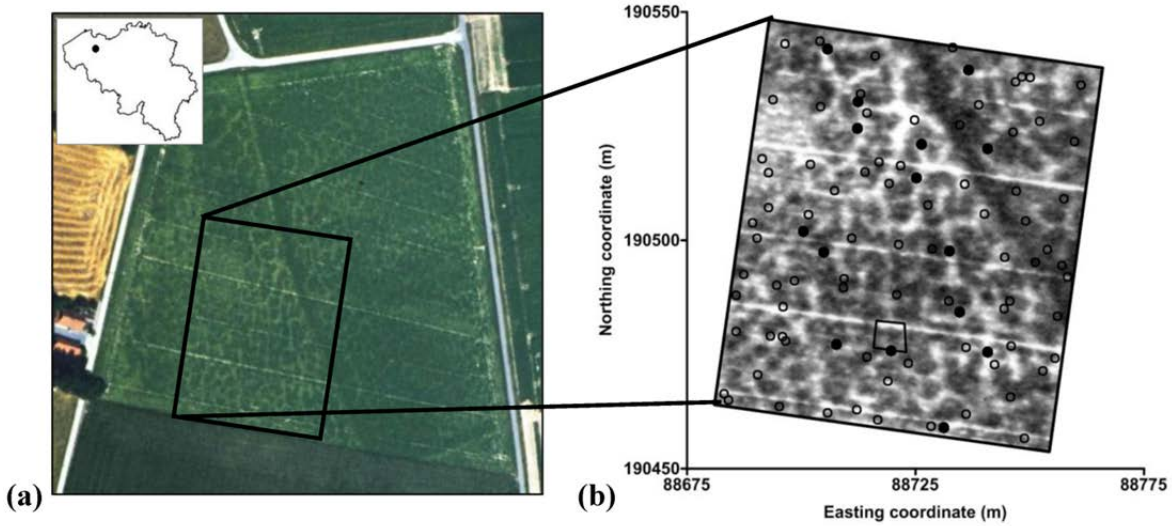


Figure 3.1 (a) Aerial photograph taken on 4 August 1996 showing polygonal crop marks and a former field track (north – southeast oriented) with delineation of the test area (large rectangle) (© J. Bourgeois, Department of Archaeology and Ancient History of Europe, Ghent University, Belgium, Photo: J. Semey) and (b) same aerial photograph after georeferencing, clipping and color stretching with delineation of the excavated area (small rectangle) and indication of the 94 augering locations (dots). Full dots represent a selection of 15 samples located on the polygonal crop marks. Coordinates are according to the Belgian metric Lambert-72 projection.

Because ECa values were generally larger at the former field track, we subtracted a moving spatial average (radius = 3 m) from each measurement to highlight the polygon boundaries. Finally, the residuals (ΔECa = ECa – moving average) were interpolated to a grid with a cell size of 0.1 m by 0.1 m using ordinary kriging (Surfer 9, Golden Software). Because of the larger data density in the direction of the measurements lines, we used an elliptical search window with a major axis of 2 m perpendicular to the measurement lines and a minor axis of 0.5 m parallel to the measurement lines.

3.2.3 Excavation

In a small part of the field (6 x 6-m) (Figure 3.1b) we exposed the polygonal pattern to investigate and characterize the network. An excavator crane systematically removed sediment layers of 0.3 m to a depth of 0.9 m. The horizontal exposure was photographed from a height of about 20-30 m using a remotely controlled camera attached to a kite. Afterwards, the photograph was georeferenced and color contrast enhanced (Arcmap 9.3, ESRI).

3.2.4 Soil Sampling

To characterize the textural variability of the subsoil, we took 94 subsoil samples within the test area according to a mixed systematic and random scheme. Half of the locations were sampled according to a grid to ensure equal coverage and the other half were randomly located (Figure 3.1b). As the casts extended downwards from a depth of 0.6 m (see further), samples were taken from the 0.6 - 0.8 m depth interval. The textural fractions were analyzed with the conventional sieve-pipette method.

The results of the texture analyses were classified into two groups by a fuzzy *k*-means algorithm with the FuzMe software (Minasny and McBratney, 2002). The multivariate classification was based on the clay and sand percentage using a Mahalanobis distance matrix and a fuzziness exponent φ of 1.6. The determination of φ was done following the scheme proposed by McBratney and Moore (1985). Each observation was assigned to the class for which it received the largest fuzzy membership.

3.3 Results and discussion

3.3.1 Excavated area

The aerial photograph, taken by the camera attached to the kite, of the 0.9 m deep excavation pit shows a more or less continuous part of a network of polygonal cells with a diameter of about 6 m (Figure 3.2). The structures suggest thermal contraction cracking in a permafrost environment, probably during the last part of the Weichselian (Buylaert et al., 2009). The wedge infillings comprised yellowish brown, structureless sandy sediments with dispersed gravel elements. The pseudomorphs extended down from the base of a 0.6 m thick silty-sandy Quaternary layer and penetrated sandy-clayey host material belonging to the Ypresian stage of the Eocene epoch (55.8 - 48.6 Ma).

The shape of the ice-wedge casts was irregular with wide (0.3 - 1.2 m) upper parts. Irregular shapes point to thaw modification as wedge ice melted, though the occurrence of sand in the original wedge filling cannot be excluded. Ghysels and Heyse (2006) described composite-wedge casts at other sites on the same plateau.

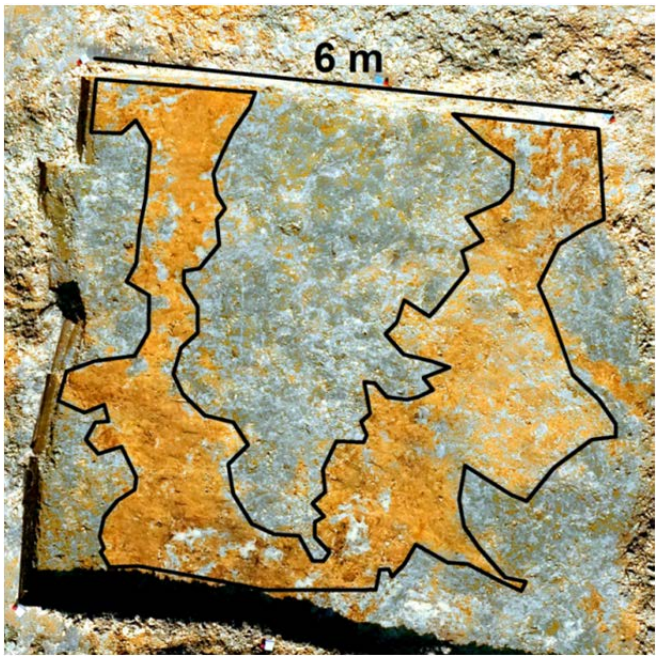


Figure 3.2 Kite aerial photograph of the 0.9 m deep excavation pit (6 x 6-m) with enhanced color contrast and indication of the outline of the polygonal network of ice-wedge casts.

3.3.2 Subsoil textural variability

The subsoil texture covers four USDA textural classes as shown in Figure 3.3. The average soil textural composition corresponds to a sandy loam texture class, but given the bimodal nature of the textural fractions in this field (see further), the average class is not representative. Table 3.1 gives the result of the 94 texture analyses of the 0.6 - 0.8 m subsoil samples. The coefficients of variation are 0.19 for the sand fraction, 0.52 for the silt fraction and 0.62 for the clay fraction. This large variability in subsoil texture contrasts strongly with the homogeneous topsoil (not analytically determined but this could clearly be observed in the field by hand feeling) and is responsible for the substantial variation in crop performance.

The fuzzy *k*-means classification resulted in a division of the 94 samples in almost two equal classes (Table 3.1): class I contained 51 samples with a centroid at 61.3 % sand, 17.4 % silt and 21.3 % clay (i.e. sandy clay loam), and class II contained 43 samples with a centroid at 85.9 % sand, 8.1 % silt and 5.9 % clay (i.e. loamy sand) (Figure 3.3). A Wilks' lambda test showed that the means of these classes are significantly different ($p < 0.001$). Based on the aerial photograph we selected 15 sampling locations which were clearly located on a crop mark polygon (Figure 3.1b). Since these 15 points were all classified as belonging to class II (Figure 3.3), we concluded that class II corresponds to the Quaternary wedge filling, and that the polygonal crop marks visible on the aerial photograph (Figure 3.1) represent the network of ice-wedge casts. Hence, class I represents the Tertiary host material.

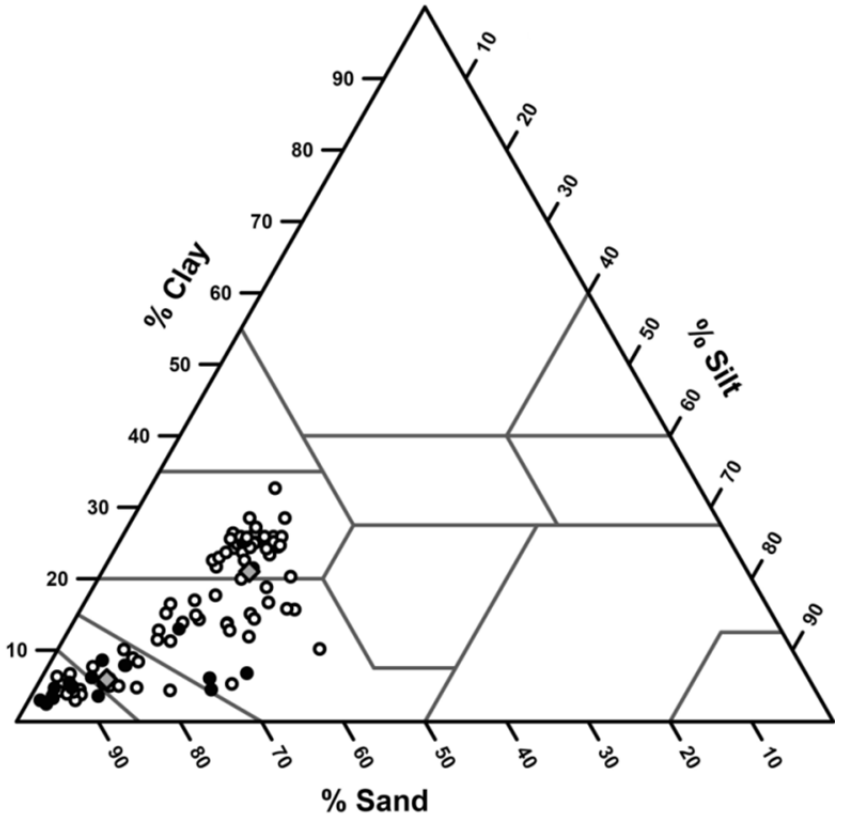


Figure 3.3 Results of the 94 texture analyses plotted on the USDA soil texture triangle. The full dots are the selection of 15 samples located on polygonal crop marks on Figure 3.1b. The grey diamonds represent the centroids of two classes created by a fuzzy *k*-means classification.

Table 3.1 Results of the texture analyses of the subsoil (0.6 - 0.8 m) samples taken within the test area.

	average (%)	min (%)	max (%)	variance $(\%)^2$
All samples (n = 94)				
sand	72.6	52.0	95.6	191.5
Silt	13.1	1.4	32.0	46.5
Clay	14.3	2.5	32.7	77.8
Class I (n = 51)				
Sand	61.3	52.0	74.1	28.0
Silt	17.4	10.6	32.0	16.6
Clay	21.3	10.2	32.7	28.2
Class II (n = 43)				
Sand	85.9	68.4	95.6	53.8
Silt	8.1	1.4	24.9	35.8
Clay	5.9	2.5	13.0	7.2

3.3.3 Image of the polygonal network

The average of the 82 770 ECa -V values was 41.0 mS m^{-1} and of the ECa-H values it was 32.3 mS m^{-1}. This indicates that the deeper soil layers have an overall larger ECa. The Pearson correlation coefficient between the ECa-V and ECa-H values was 0.88 and the pattern shown by both maps was very similar. However, the ΔECa-H map showed sharper contrasts revealing the polygons in more detail. So regardless of their smaller DOE, measurements taken in the horizontal dipole mode proved more appropriate to map the ice-wedge casts in the subsoil. A possible explanation is that the ECa-V measurements received a larger response from the Tertiary material underlying the ice-wedge casts, which masked the influence of their sandy infillings. Despite our experience (Cockx et al., 2006), a combination of both signals did not result in an improvement. Therefore, we continued with the ΔECa-H map shown in Figure 3.4a and further indicated as ΔECa map. Similar to the classification of the subsoil samples, a *k*-means classification of the ΔECa map resulted in two classes of approximately equal size meaning that both subsoil textures occur with about the same frequency.

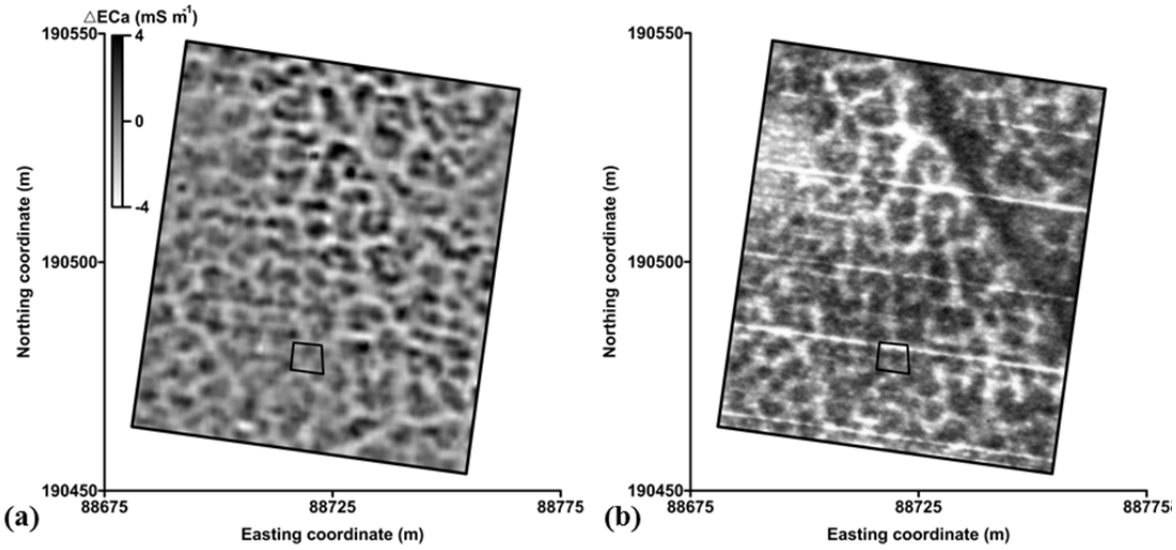


Figure 3.4 (a) ΔECa (mS m^{-1}) map and (b) georeferenced aerial photograph taken on 4 August 1996 with delineation of the excavated area on both figures (small rectangle).

The ΔECa map (Figure 3.4a) images the polygonal network of ice-wedge casts clearly due to the smaller EC of the wedge filling, caused by its smaller clay content. In general, positive ΔECa values correspond to the host material, whereas negative ΔECa values correspond to the wedge filling. For comparative reasons, Figure 3.4b shows the georeferenced aerial photograph. Although one image represents the variability in soil electrical conductivity and the other one in crop color, it can be observed that both are very similar. However, measuring ECa is much less dependent on external conditions than observing crop marks, asking for a particular combination of crop and climatic conditions.

The five parallel white horizontal lines on the aerial photograph are due to a non-uniform sowing density.

3.3.4 Verification

For each of the 94 sampled locations the ΔECa was extracted from the ΔECa map to investigate the relationship between soil texture and ΔECa. The Pearson correlation coefficient between ΔECa and the subsoil textural fractions was -0.68 for sand, 0.46 for silt and 0.71 for clay. So it is clear that ΔECa is a proxy for the subsoil clay and sand content. Figure 3.5 illustrates these relationships and confirms the existence of two distinctly different subsoil classes.

Figure 3.6 shows a detail of the ΔECa map around the excavated area with indication of the boundary of the exposed polygon (Figure 3.2). The differences between the wedge filling and the host material are clearly visible on the ΔECa map confirming the direct relationship between the processed ECa measurements and the presence of ice-wedge casts in the subsoil of this area.

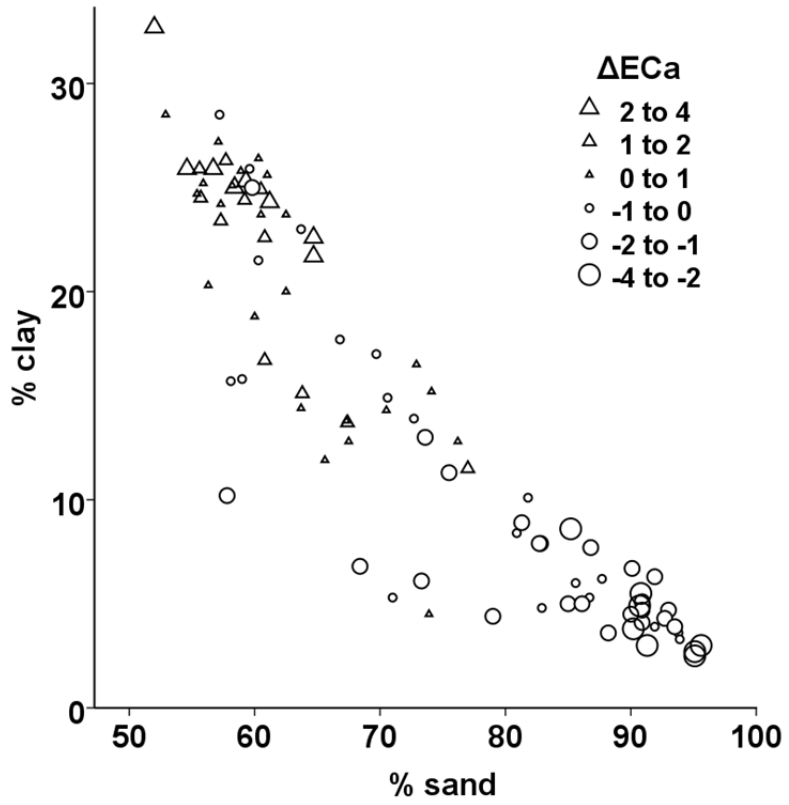


Figure 3.5 Subsoil (0.6 - 0.8 m) clay and sand content in relation to ΔECa (mS m^{-1}): the size and shape of the symbols correspond to the ΔECa value measured at the same location.

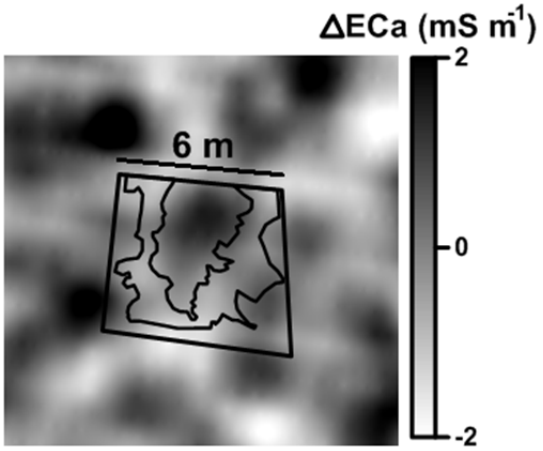


Figure 3.6 Detail of the ΔECa map with indication of the excavated area and the observed polygon outline (as shown in Figure 3.2).

3.4 Conclusions

A textural difference between host material and wedge filling is the key to successfully mapping polygonal networks of ice-wedge casts with EMI sensors. In contrast to being dependent on occasional aerial photographs of polygonal crop marks, the use of mobile EMI sensors offers a more generally applicable method to map ice-wedge pseudomorphs, even when these are covered by a topsoil layer of 0.6 m. This non-invasive, fast method offers detailed exhaustive information about the morphology of the cryogenic features.

Our study showed that the presence of ice-wedge casts in the subsoil can be responsible for a highly heterogeneous subsoil texture. About half of the test area has a sandy clay loam subsoil texture, as indicated on the 1/20 000 soil map of Belgium. The other half represents the ice-wedges which contain considerably more sand and consequently less clay. Because the presence of both textures is spatially structured and can be mapped, this situation can be considered as a challenge for managing the within field variability to optimize crop yield, i.e. precision agriculture.

The collected data set consisting of the soil knowledge provided by the excavation, the 94 soil samples and the proximal soil sensor image, is a good test case to evaluate the applicability of MPG to reconstruct complex soil patterns.

Chapter 4

A geometric random function model for the polygonal network

The content of this chapter is based on: Lark, R.M., Meerschman, E. and Van Meirvenne, M. A stochastic geometric model of the variability of soil formed in Pleistocene patterned ground. Submitted for publication in *Geoderma* (2013).

The spatial pattern of the collected ECa data (chapter 3) shows the connectivity of small ECa values (coarser soil material). Before practicing MPG, we first selected and fitted a geometric non-Gaussian RF model to the ECa data. This alternative RF model was inferred from soil knowledge. We then compared the geometric RF model with a trans-Gaussian (TG) model of the ECa data, i.e. a model fitted by conventional geostatistical analysis after the data have been transformed to approximate normality. Specifically we compared the models with respect to a criterion that summarizes the spatial connectivity of small ECa values, which might be relevant to simulations of transport processes in the soil. We then evaluated which model appeared best to represent the spatial pattern in the ECa data.

4.1 Introduction

'Mais surtout nous insisterons sur la nécessité d'incorporer au maximum la physique du problème et le contexte géologique de la zone ´etudiée.' Chilès and Guillen (1984).

In most geostatistical analyses of soil the data are assumed to be a realization of a multi-Gaussian RF, perhaps after they have been transformed so that their histogram

represents a Gaussian distribution. Furthermore, the RF commonly has a spatial covariance or variogram function drawn from a limited subset of models (Webster and Oliver, 2007), which are used because of their convenient mathematical properties (see section 2.1 and 2.2). In some of the earth sciences there has been progress in the development of RFs with parameters that are determined, or at least constrained, by parameters of underlying processes which have a physical meaning (e.g. Kolvos et al., 2004; Chilès and Guillen, 1984). This has advantages (Lark, 2012a), for example, the efficiency of spatial sampling to model the spatial covariance function could be improved if prior distributions for covariance parameters could be specified from process knowledge. However, this has not been achieved in soil science. Lark (2012a) suggested that this is probably because the variables that soil scientists study are commonly influenced by a more complex set of factors at more diverse spatial scales than is the case for the variables where it has proved possible to specify the covariance function from process information. For example, the covariance function for diffusion processes is well-established (Whittle, 1954; 1962), and diffusion is a source of spatial variation in the concentration of nutrients in soil, but it is just one of many sources of spatial variation, and is of limited importance at the spatial scales most generally studied for practical purposes.

Lark (2012a, 2012b) suggested that progress might be made by recognizing a number of distinct *modes* of soil variation, simple and generalizable rules that capture how the effects of factors of soil variation vary laterally, and which map naturally on to particular spatial RFs. For example, in conditions where soil variation is strongly determined by differences between discrete domains in the landscape (such as geological units, topographic units, fields etc.) then a subdivision of space into random sets such as Poisson Voronoi polygons may be appropriate (Lark, 2009) and properties of the spatial model (such as the mean chord length of the polygons) may be given a physical meaning.

Lark (2012b) proposed a mode of soil variation: continuous local trends. Under this mode of variation soil varies laterally in space, changing continuously rather than in a step-wise fashion; and these trends are local and repeating, so that they are essentially unpredictable (in contrast to a large-scale trend in a variable that might be observed across a study area). Examples of continuous local trends would be concentration gradients around the rhizosphere, or around individual plants, and catenary variation at landscape scale. Lark (2012b) proposed a general family of RFs to describe continuous local trends (CLT random functions). The value of a CLT variable at some location is given by a distance function, whose argument is the distance from the location of interest to the nearest event in a realization of a spatial point process. This makes the CLT a random function. The CLT variables considered by Lark (2012b), and in this chapter, are Poisson CLT (PCLT) variables because the spatial point process is completely spatially random. Lark (2012b) estimated parameters of a PCLT process from data on a soil variable. It was

also pointed out that the PCLT process might differ from a comparable Gaussian RF with respect to its multiple-point statistics. This raises the possibility that PCLT models, as well as mapping closely on to a particular mode of soil variation, might be practically useful for applications where spatial connectivity plays a major role controlling processes in soil and so the multiple-point statistics of the variable are important.

4.2 Initial data analysis

In chapter 3 it was shown that the ECa measurements clearly reflected the polygonal patterns: small ECa values indicated the former ice-wedges filled with lighter material. In addition to the short-range variation in ECa, there were large values of ECa near an old field track in the north-east of the surveyed region. To avoid any assumptions about the form of this trend we decided to restrict our analyses to the lower left quadrant of the surveyed area, a region of approximately 40 x 40-m with 17 792 observations. We used the processed ECa data, and not the residuals ΔECa (see section 3.2.2). Figure 4.1 shows a post-plot of these data. The coordinates were first rotated and then translated to have their origin in the lower left corner.

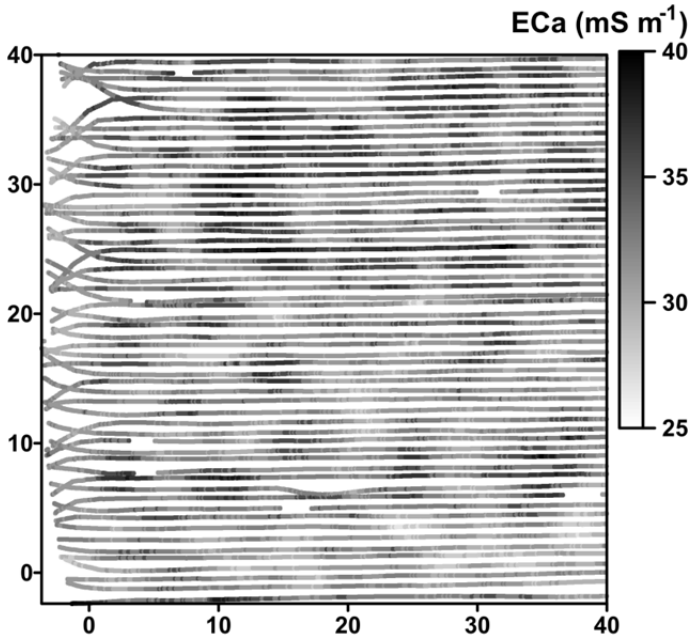


Figure 4.1 ECa data after coordinate transformation (coordinates are in metres relative to the the lower left corner).

Figure 4.2 shows the histogram of the data. Summary statistics are presented in Table 4.1. Note that the data are mildly skewed. In the analyses reported below the PCLT model was fitted in all cases to the raw data, and all analyses with the TG model were done with the data after a transformation.

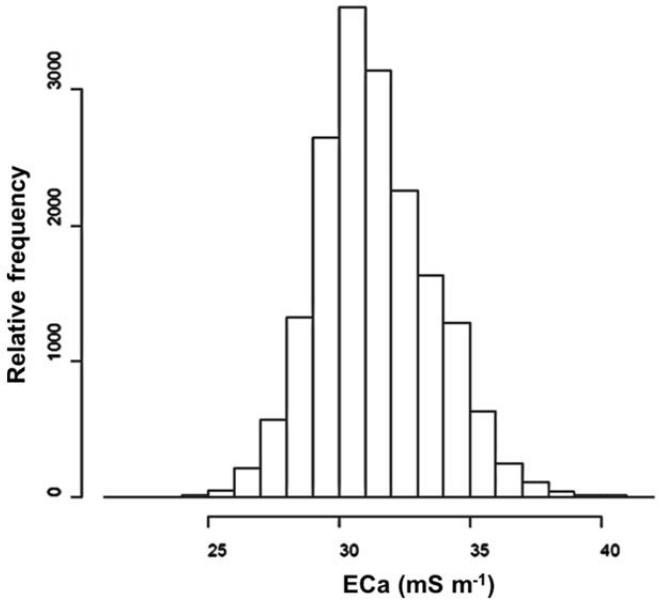


Figure 4.2 Histogram of ECa data.

Table 4.1 Summary statistics of the raw data on ECa (n = 17 792).

Statistic	mS m^{-1}
Average	31.37
Median	31.13
Standard deviation	2.2
Skewness	0.36
Quartile 1	29.9
Quartile 3	32.76
Octile 1	29.03
Octile 7	34.08

4.3 Trans-Gaussian model

The objective of the case study is to compare a continuous local trend (PCLT) model of the data with a trans-Gaussian (TG) model, as might be used in standard geostatistical analysis. We therefore used a Box-Cox transformation of the data to normality for the TG modelling:

$$\begin{aligned} y &= \frac{z^{\zeta}-1}{\zeta} \qquad \zeta \neq 1, \\ y &= \log_e(z) \qquad \zeta = 1, \end{aligned} \tag{4-1}$$

where z is a value on the original scale and y is a transformed value. We used the BOXCOX procedure from the MASS package (Venables and Ripley, 2002) for the R

platform (R Development Core Team, 2012) to find the likelihood profile of the ζ parameter. The data were then transformed with the maximum likelihood estimate of ζ (-0.57), substituted into Eq. 4-1 and then standardized to zero mean and unit variance. The summary statistics for the data after transformation, and standardization, are presented in Table 4.2.

Table 4.2 Summary statistics of the data on ECa after Box-Cox transformation and for the transformed data after standardization. Variogram parameters for the standardized data are also given.

Statistics	Transformed data	Transformed and standardized data
Average	1.508	0
Median	1.507	-0.056
Standard deviation	0.01	1
Skewness	0	0
Quartile 1	1.501	-0.646
Quartile 3	1.514	0.668
Octile 1	1.497	-1.085
Octile 7	1.52	1.216
Variogram parameters *		
C_0		0.12
C_1		0.84
a		1.91
κ		1.49

*Powered (stable) exponential model, see Eq. (4-2)

An isotropic empirical variogram of the transformed and standardized data was then computed using Matheron's method of moments estimator (Matheron, 1962) as implemented in the FVARIOGRAM directive in GenStat (Payne et al., 2009). An authorized model was then fitted to the estimated variogram by weighted least squares (Cressie, 1985) using the MVARIOGRAM procedure in GenStat (Harding et al., 2010). Alternative models were considered and the stable or powered exponential model was selected on the basis of the Akaike information criterion (McBratney and Webster, 1986). This variogram model takes the form:

$$\gamma(h) = C_0 + C_1(1 - \exp(-\left\{\frac{h}{a}\right\}^{\kappa}), \qquad (4\text{-}2)$$

where κ is a shape parameter ($0 < \kappa \leq 2$). The estimates of these parameters are presented in Table 4.2, and the estimates of the variogram of the TG variable, and the fitted model are shown in Figure 4.3.

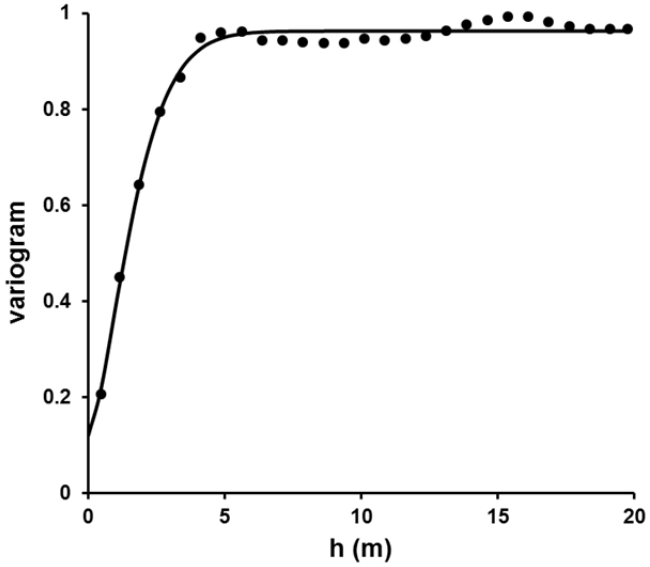


Figure 4.3 Empirical variogram of transformed and standardized ECa data with a fitted model.

4.4 Stochastic geometric model

Estimates of the isotropic variogram of the raw data on ECa were obtained using the method of moments estimator due to Matheron (1962) as previously described for the transformed data (these are the solid symbols in Figure 4.6). The identification and fitting of an appropriate stochastic geometric model for the soil variable will allow us to plot a continuous variogram function for these estimates.

When a TG model is fitted it is assumed that, after any transformation, the data $\mathbf{y} = \{y(\mathbf{x}_1), y(\mathbf{x}_2),..., y(\mathbf{x}_n)\}$ from the n locations $\mathbf{x}_1, \mathbf{x}_2,..., \mathbf{x}_n$ can be regarded as a realization of an n-variate Gaussian RF $\mathbf{Y}$. Under this assumption the variogram of $\mathbf{Y}$ entirely summarizes the information that the data contains about the spatial variability of $\mathbf{Y}$, and the task of estimating model parameters, under the assumption of a stationary mean, reduces to the task of estimating variogram parameters. This is not the case with models for random functions, such as the PCLT models, which have non-zero moments of order three or larger, and therefore are not Gaussian. The fitting of a PCLT model cannot, therefore, simply reduce to the computation of parameters which minimize the weighted sum of squared residuals between the empirical and fitted variogram.

In this study our approach to the selection and estimation of a PCLT model is to constrain it by soil knowledge. Soil knowledge consists of general understanding of the underlying processes that influence soil formation and so the variation of the target

variable, and also of general quantitative information about the variable in the study site or a homologous site, represented by summary statistics, empirical variograms or similar information. In the following sections we go through a semi-formal process of model identification based on inferences from soil knowledge and culminating in the estimation of parameters for an appropriate model. Each subsection is headed with a question, and with the general source of soil knowledge used to address it. The individual elements of soil knowledge are then summarized in brief labelled sentences, expanded in a short paragraph. Inferences from this soil knowledge are then set out.

4.4.1 Question: 'What mode of soil variation?' Soil knowledge about the underlying pedogenetic process.

SK1. *The dominant source of soil variation at metre scale in this landscape is the presence of Pleistocene ice-wedge polygons.* These are described in more detail in chapter 3. Ice-wedge polygons form in periglacial conditions on surfaces with slopes less than a critical value. Over much of central Europe ice-wedge polygons formed in periglacial conditions during the Quaternary, they are detectable at the study site from airphotography. It has been shown (Cresto Aleina et al., 2012) that the comparable polygonal patterns in ground of contemporary tundra can be modelled as a Poisson Voronoi Tessellation (PVT), that is to say one may postulate an underlying homogeneous spatial point process of completely spatially random seed points, and any one polygon consists of all locations nearest to one associated seed point than to any of the others. See Lark (2009) for a summary of some of the properties of PVT spatial processes and Okabe et al. (2000) for a more complete account. By analogy we infer that a PVT model would be a plausible descriptor of the ice-wedge polygons at the study site.

SK2. *We may expect more or less continuous variation in depth-integrated soil properties from the centre to the edge of any polygon.* Much of the polygonal patterned ground formed in Europe and North America during the Quaternary was covered by aeolian or glacio-fluvial sand or silty deposits. These have an important role in subsequent pedogenesis (Catt, 1979; Walters, 1994) imposing local lateral trends. At the centre of a polygon there is typically a relatively thin layer of sandy or silty superficial material over the host material in which the ice-wedges originally formed. After thawing, the space previously occupied by ice in the wedges that delineate the polygons was typically filled with the superficial material. Any depth-integrated soil property, such as ECa, can therefore be expected to vary laterally (although not necessarily linearly) from the centre of the polygon to its edge if there is a texture contrast between the host material and the superficial material. There is such a contrast at the Deinze study site where the overlying material is silty-sand Quaternary deposits, and the host material is Eocene sandy clay.

From these two elements of soil knowledge we may infer that the spatial variation of a depth integrated soil property such as ECa, in these conditions, can plausibly be regarded as a Poisson Continuous Local Trend random process as defined by Lark (2012b). In the next section we consider what distance function might be proposed.

4.4.2 Question: 'What type of distance function is plausible?' Soil knowledge about pedogenetic processes and summary statististics.

SK3. *We may expect ECa to decline from the polygon centre to the rim.* It is generally found that measurements of ECa made by electromagnetic induction are positively correlated with the clay content of the soil (e.g. Kachanoski et al., 2002; Saey et al., 2009b). For this reason we should expect ECa, as a depth-integrated variable, to decline from the polygon centre, where the thickness of sandy and silty material over the heavier host material is thinner, to the edge of the polygon where the former ice-wedge is filled with the lighter material. This was found to be the case for our test case (chapter 3).

SK4. *The data on ECa are mildly positively skewed.* This can be seen in Table 4.1.

The simplest PCLT model, as used by Lark (2012b), has a linear distance function $D(k) \propto k$. If the distance function has a positive slope, i.e. $\{k' > k\} \rightarrow \{D(k') > D(k)\}$, then it can be seen that the corresponding PCLT random function has a moderate positive skewness (about 0.65). A linear distance function with a negative slope, needed for consistency with SK3, would therefore give rise to a RF with a moderately negative skewness. This is not compatible with SK4.

Of the distance functions examined by Lark (2012b) one in which the distance function is proportional to the reciprocal of distance is compatible with SK3 and SK4. The reciprocal of distance declines with distance (SK3), and the example of such a RF given by Lark (2012b) has mild positive skewness (SK4). On this basis it was decided to proceed with further analysis on the assumption that the data on ECa could be regarded as realizations of a PCLT process with a distance function linearly proportional to

$$D(k) = \frac{1}{k + a} \qquad (4\text{-}3)$$

where k is the distance to the nearest event of the underlying spatial point process, and α is a parameter which must take some value $\alpha > 0$ to ensure that the distance function is defined for all positive k. We refer to this PCLT as the *inverse-distance PCLT* in the remainder of this chapter.

This distance function was selected because it was seen to be a simple function, at least potentially compatible with available soil knowledge. In due course its parameters are estimated and this gives some further indication of its plausibility, and in section 4.5 we evaluate statistics to compare its plausibility with the TG model.

We call the inverse-distance PCLT random function Z_{id}. We shall model the ECa data as a realization of a random function Z where

$$Z = Z_n + Z_{\mathrm{id}},$$

and Z_n is an is an independently and identically distributed Gaussian nugget component of mean zero. We now obtain the cumulative distribution and density functions of Z_{id}.

We first define the inverse of the distance function in (Eq. 4-3), $D'(z_{\mathrm{id}})$, such that

$$\left\{ z_{\mathrm{id}} = D(k) = \frac{1}{k+a} \right\} \leftrightarrow \left\{ D'(z_{\mathrm{id}}) = k \right\}.$$

Then

$$D'(z_{\mathrm{id}}) = \frac{1}{z_{\mathrm{id}}} - \alpha. \qquad (4\text{-}4)$$

Since $D(k)$ is monotonic and decreasing with increasing k for admissible (non-negative) values of k, the marginal cdf of Z_{id}, $F_{\mathrm{id}}(z)$ can be written as

$$F_{\mathrm{id}}(z_{\mathrm{id}}) = 1 - F_k(D'(z)), \qquad (4\text{-}5)$$

where $F_k(k)$ is the marginal cdf of k. In Eq. 14 of Lark (2012b) it is shown that, for a Poission point process in 2D with intensity λ,

$$F(k) = 1 - \exp\left\{ -\lambda\pi k^2 \right\}, \qquad (4\text{-}6)$$

and so

$$F_{\mathrm{id}}(z_{\mathrm{id}}) = \exp\left\{ -\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)^2 \right\}, \qquad (4\text{-}7)$$

which is defined for $0 \le z_{\mathrm{id}} \le 1/\alpha$, which shows that random function Z_{id} has an upper and a lower bound.

By differentiation of $F_{\mathrm{id}}(z_{\mathrm{id}})$ with respect to z_{id} we can obtain a pdf:

$$\begin{aligned} f_{\mathrm{id}}(z_{\mathrm{id}}) &= \frac{2\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)}{z_{\mathrm{id}}^2} \exp\left\{ -\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)^2 \right\}, \quad 0 \le z_{\mathrm{id}} \le \frac{1}{\alpha} \qquad (4\text{-}8) \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

A soil variable modelled as an inverse-distance PCLT random function is assumed to have a spatially correlated component that is linearly proportional to z_{id} for some values of the parameters α and λ. As noted above, the soil variable is assumed to be a realization of a random function Z that includes an independent Gaussian nugget component of mean

zero. If the pdf of the nugget component is denoted by $f_n(z_n)$, then the pdf of Z, $f(z)$, can be obtained by the convolution operation

$$f(z) = \int_{-\infty}^{\infty} f_n(x) f_{\mathrm{id}}(z-x)\mathrm{d}x, \qquad (4\text{-}9)$$

since Z_{id} and Z_n are independent random variables (Dudewicz and Mishra, 1988).

4.4.3 Question: 'What is a plausible range of values for λ, the intensity of the process?' Soil knowledge from field observations and an estimate of the proportion of variation of ECa that is attributable to the nugget component.

SK5. *Chapter 3 reports a detailed excavation of a polygonal cell with a diameter of about 6 m, which is regarded as typical from airphoto evidence.* If all cells have a diameter of d m then the average intensity of an underlying spatial point process is $4/\pi d^2$. On the basis of the information provided by the excavation it was decided to consider a range of possible values of λ for the spatial point process in the interval [0.02 m^{-2}, 0.08 m^{-2}] which corresponds to a range of polygon diameters from 4 to 8 m (i.e. 2 m either side of the value proposed as representative).

SK6. *The nugget variance of the (untransformed) ECa data is about 10% of the correlated variance.* This information is needed to allow us to calculate moments of the pdf in Eq. 4-9. To obtain it we fitted a powered exponential model (Eq. 4-2) to the empirical variogram of the ECa data (not shown here) using the MVARIOGRAM procedure in GenStat (Harding et al, 2010).

The mean and variance of Z_{id} for some values of the parameters α and λ was obtained from the pdf in Eq. 4-8, the QDAG algorithm in the IMSL library (Visual Numerics, 2006) was used for numerical integration. It was then possible to compute the variance of an independent Gaussian nugget component Z_n such that the variances of Z_{id} and Z_n were in the same ratio as SK6 suggests for the ECa data. The coefficient of skewness for the sum of these two random variables could then be calculated from moments obtained by numerical integration of the convolution of the distributions of Z_{id} and Z_n (Eq. 4-9).

Figure 4.4 is a plot of values of the skewness coefficient of Z_{id} for values of the parameters α and λ, the range for λ obtained from SK5. Note that over much of the range of values of λ it is α that has the strongest effect on the skewness. The two contours drawn on Figure 4.4 bound a region within which the skewness is in the interval [0.25, 0.5]. We regard this as mild positive skewness, compatible with SK4. Figure 4.4 shows that values of α less than 2 m seem unlikely to be compatible with SK4 since coefficients of skewness for such variables are larger than 0.5.

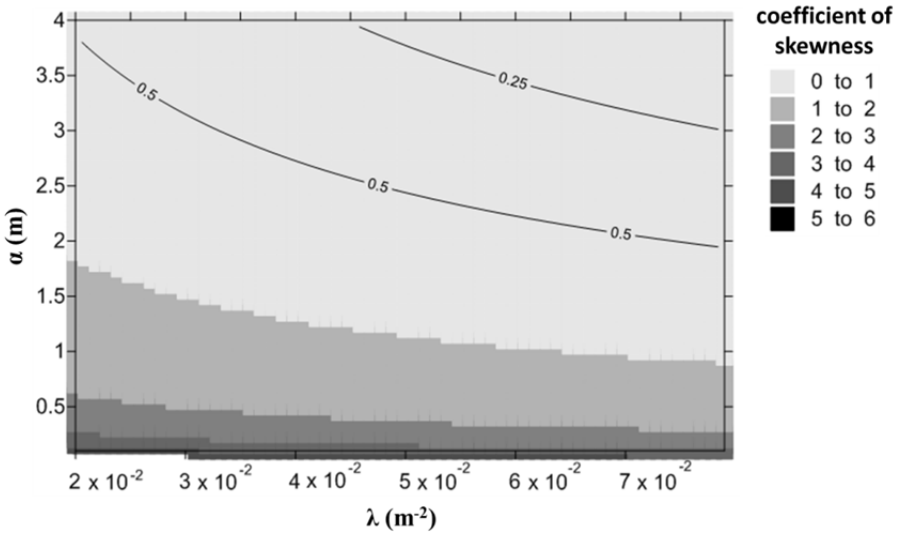


Figure 4.4 Values of the coefficient of skewness for an inverse-distance PCLT process with different values of the parameters λ and α. The two contours bound the region where we regard the variable as mildly positively skewed.

4.4.4 Model fitting given the soil knowledge

Estimates of the isotropic variogram of the raw data on ECa were obtained using Matheron's method of moments estimator (Matheron, 1962) as implemented in the FVARIOGRAM directive in GenStat (Payne et al., 2009). An inverse-distance PCLT model was then fitted to the estimates. This variogram was specified by:

$$\gamma_{\mathrm{id}}(h) = C_0 + C_1 g_{\mathrm{id}}(h|\alpha,\lambda), \qquad (4\text{-}10)$$

where $g_{\mathrm{id}}(r|\alpha,\lambda)$ is the standardized inverse-distance PCLT variogram:

$$g_{\mathrm{id}}(h|\alpha,\lambda) = 1 - \frac{C_{\mathrm{id}}(h|\alpha,\lambda)}{C_{\mathrm{id}}(0|\alpha,\lambda)}, \qquad (4\text{-}11)$$

where $C_{\mathrm{id}}(h|\alpha,\lambda)$ is the covariance function for lag h for an inverse-distance PCLT process with parameters α and λ. The covariance function for a variable in 2D is given by

$$C_{\mathrm{id}}(h|\alpha,\lambda) = \int_{R^2} \{S(k,k_r) + F(k) + F(k_r) - F(k)F(k_r) - 1\} \left\{-\frac{1}{(k+\alpha)^2}\right\} \mathrm{d}k \left\{-\frac{1}{(k_r+\alpha)^2}\right\} \mathrm{d}k_r, \qquad (4\text{-}12)$$

where $S(k,k_r)$ is the joint survival function for the underlying spatial point process, as defined by Lark (2012b). This equation is obtained directly from Eq. 20 of Lark (2012b) and the reader is referred to that paper for details.

To fit the inverse-distance PCTL variogram the value of α was first fixed. The parameter λ was then set to values over the range defined from SK5, and for each value the IMSL optimization subroutine BCPOL (Visual Numerics, 2006) was used to find estimates of C_0 and of C_1 that minimize the weighted sum of squares of deviations between the

fitted variogram model and the point estimates (Cressie, 1985). This produces a 'profile' plot of the weighted sum of squares against λ for fixed α.

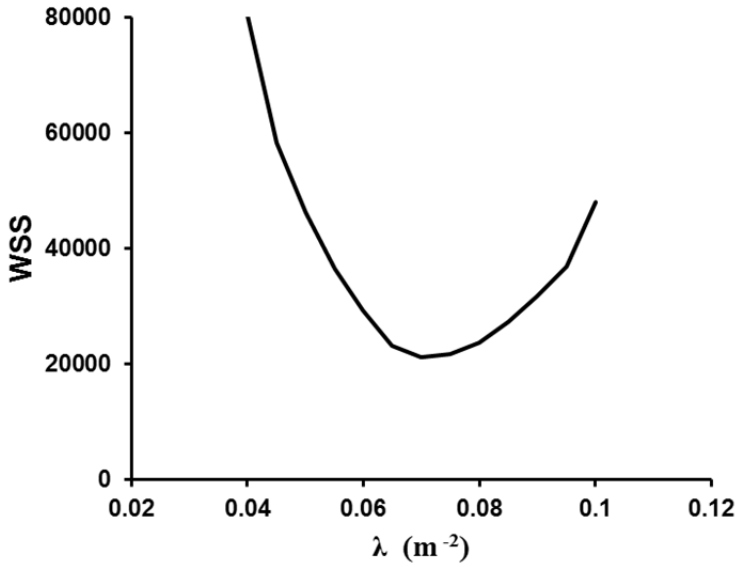


Figure 4.5 Profile plot of the weighted sum of squares (WSS) for the fit of the inverse-distance PCLT variogram function against λ, with α fixed at 2.5 m.

The procedure followed was to set α to discrete values ≥ 2.0 and to find the value of λ with the minimum weighted sum of squares on the profile plot. If the two values of the parameters fell within the region of mild positive skewness in Figure 4.4, then they were retained as possible estimates. The set of parameters was then selected for which the weighted sum of squares was smallest. The resulting values of α and λ were 2.5 m and 0.07 m^{-2} respectively. The estimated nugget and spatially correlated variance were 0.49 and 4.03 respectively. Figure 4.5 shows the profile plot of the weighted sum of squares (WSS) with α = 2.5 m and Figure 4.6 shows the empirical variogram for the untransformed data and the fitted inverse-distance PCLT model. In Figure 4.7 is shown the qq-plot of the standardized random function $Z = Z_{\mathrm{id}} + Z_n$ and the standardized ECa data. The theoretical and empirical distribution functions are in reasonable agreement, although the median of the former is slightly smaller than the latter.

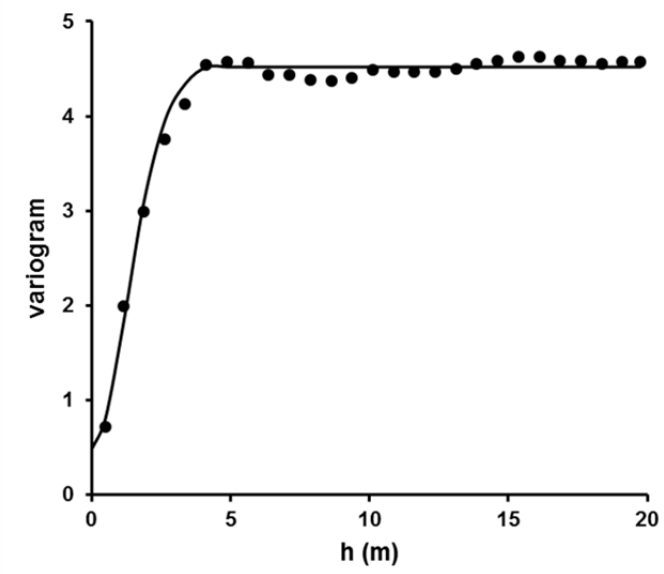


Figure 4.6 Empirical variogram of the untransformed ECa data with the fitted inverse-distance PCLT variogram.

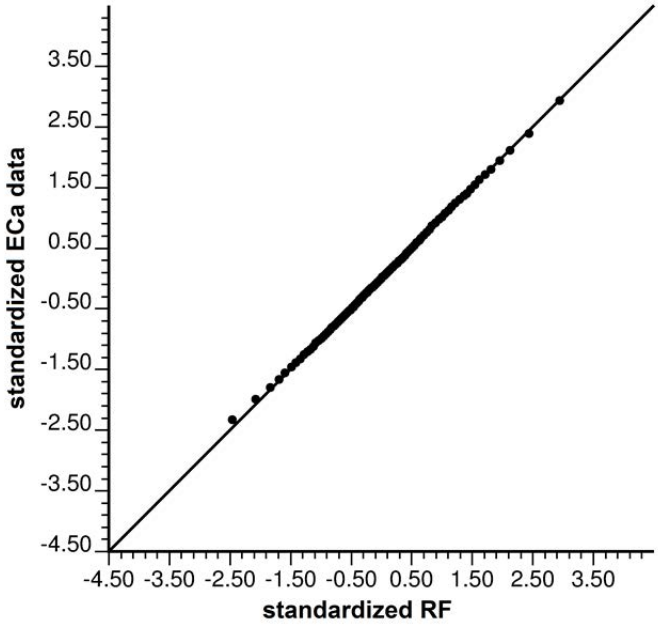


Figure 4.7 (a) qq-plot of the standardized inverse-distance PCLT random function with $\alpha = 2.5$ m and $\lambda = 0.07$ m^{-2} and the standardized ECa data.

4.5 Comparing the TG and PCLT models

It is well known that Gaussian (and trans-Gaussian) models of spatial variation, in which all information on variability is expressed by two-point statistics such as the covariance function, are not able to reproduce all important features of natural spatial fields, which must be represented by higher-order moments (e.g. Guardiano and Srivastava, 1993). This has been the motivation for the development of multiple-point

statistics. In this section we investigate whether the PCLT model characterizes the spatial structure of the ECa data better than the TG model.

One feature of the Gaussian and trans-Gaussian random variables that often limits their applicability is the fact that large values of the variable tend to be spatially isolated from other large values, the same holds for small values (e.g. Strebelle, 2002). In this case study we may consider locations with small values of ECa. These locations are likely to be dominated by lighter sandy and silty Quaternary material, rather than the heavier-textured Eocene host material, and so will have larger hydraulic conductivities than sites where the ECa is larger. The potential for rapid lateral transport of water-born contaminants through such a landscape may therefore be underestimated under a TG modelling framework if the TG model does not adequately represent the local connectivity of areas with small values of ECa. Figure 4.8 shows a realization of each of the fitted PCLT and TG models for ECa. The inverse-distance PCLT realization was generated directly following the procedure used by Lark (2012b). The TG realization was obtained by Sequential Gaussian Simulation using the SGSIM subroutine from the GSLIB library (Deutsch and Journel, 1997) modified to use the powered exponential variogram function. On visual inspection it can be seen that, while some large patches with smaller ECa values are seen in the TG realization, there are fewer isolated small patches with small ECa values in the inverse-distance PCLT realization, which has large and connected regions with small conductivity around the boundaries of the Voronoi cells of the underlying point process. However, this visual inspection is of limited usefulness and a more objective measure is needed.

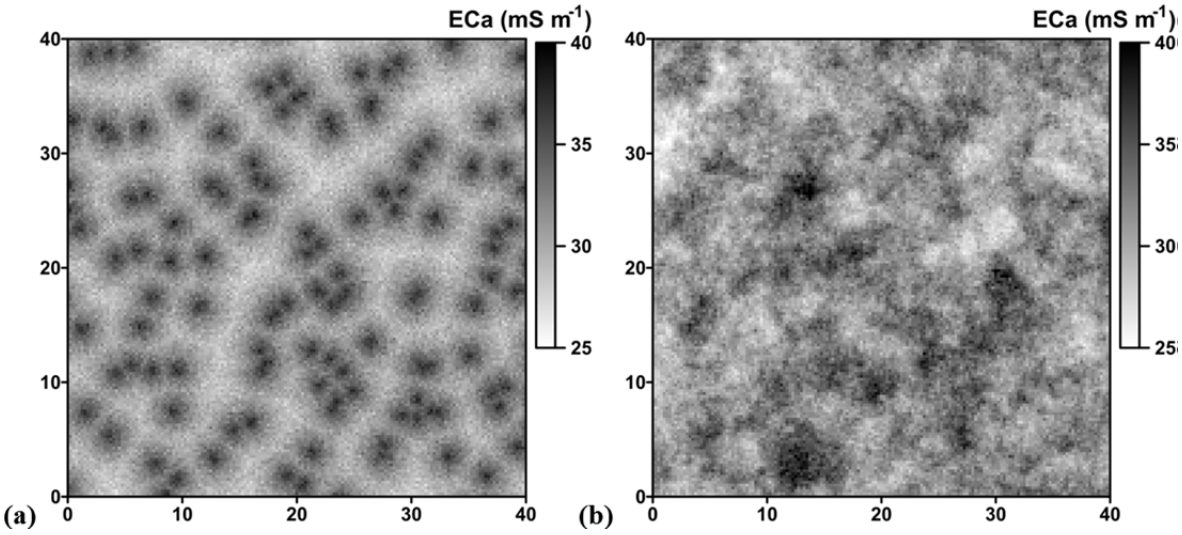


Figure 4.8 Realization of (a) the inverse-distance PCLT random function and (b) the TG random function (back transformed to original units) on a 0.25-m square grid.

To this end we consider a simple test criterion, which can be readily evaluated on the ECa data which are more or less regularly sampled but which do not constitute a comprehensively observed ‘image’. We define the criterion $P(\tau,\Delta)$ as the expected

proportion of observations within a square window of width Δ, centred at a randomly selected location **x** which are $\leq \tau$, conditional on the value at **x** being $\leq \tau$. We may expect these values to be smaller for a TG random function than for a function which better-represents the spatial structure of a variable in which small values tend to be spatially connected.

We estimated $P(\tau, \Delta)$ for the TG and PCLT random functions fitted to the ECa data by simulation. These are denoted by $P_{\mathrm{TG}}(\tau, \Delta)$ and $P_{\mathrm{PCLT}}(\tau, \Delta)$ respectively. We considered windows of width 2 m or larger (because approximately 40 ECa observations occur within a 2-m window). Each simulation program generated a single independent realization of the random function at 25 equally-spaced locations in a window of width Δ one of which was at the centre of the window. If the simulated value at the centre was $\leq \tau$, the conditioning criterion, then the realization was retained and $P(\tau, \Delta)$ was estimated as the proportion of the observations in the window for which $\leq \tau$. This was repeated until 10 000 independent realizations which met the criterion that the central value was $\leq \tau$ had been obtained. The PCLT realizations were generated using the procedure described by Lark (2012b). The TG realizations were obtained by LU decomposition (Goovaerts, 1997). The mean value of $P_{\mathrm{TG}}(\tau, \Delta)$ and the standard deviation of the 10 000 independent values, were computed for different values of Δ and for τ set to the median, first quartile and first octile of the ECa data. This was also done for $P_{\mathrm{PCLT}}(\tau, \Delta)$. The difference between the mean values of $P_{\mathrm{PCLT}}(\tau, \Delta)$ and $P_{\mathrm{TG}}(\tau, \Delta)$ for these different thresholds and for windows of different size, are plotted in Figure 4.9.

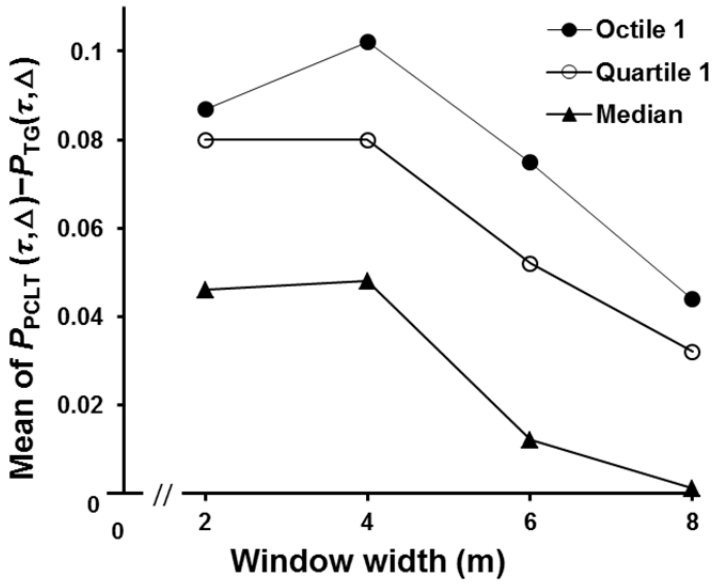


Figure 4.9 Plot of the difference between the mean of $P_{\mathrm{PCLT}}(\tau, \Delta)$ and that of $P_{\mathrm{TG}}(\tau, \Delta)$ for different window widths (Δ) and with τ set to the median, first quartile and first octile of the ECa data (mean for 10 000 realizations of each random function).

Figure 4.9 shows three things. First, the mean value of $P_{\mathrm{PCLT}}(\tau,\Delta)$ is larger than that of $P_{\mathrm{TG}}(\tau,\Delta)$ for given τ and Δ. That is to say, given that a value falls below a threshold, there is a larger proportion of neighbouring values which do so for the PCLT process than for the TG process. Second, the effect depends on the threshold, and increases as the threshold becomes more extreme relative to the overall distribution. Third, the effect depends on the window size. It is small for a large window, but it is also notable that the difference is larger for the window width 4 m than the window width 2 m. This reflects the spatial scale of the random function.

The $P(\tau,\Delta)$ statistic was then estimated from the ECa data for the same three threshold values used in the simulations, and for Δ = 4 m given that this window showed the largest differences between the two processes in the simulation. An independent random subsample of 250 observations for which ECa $\leq \tau$ was obtained, the proportion of ECa observations within a square window, width Δ about each of these observations was computed. The results are shown in Figure 4.10. The mean value of $P_{\mathrm{TG}}(\tau,\Delta)$ and $P_{\mathrm{PCLT}}(\tau,\Delta)$ from the simulations are plotted, and for each of these the 95 % confidence interval for the mean of a sample of 250 independent observations is also shown, based on the variances of the values obtained by simulation. The estimates from the ECa data are also plotted.

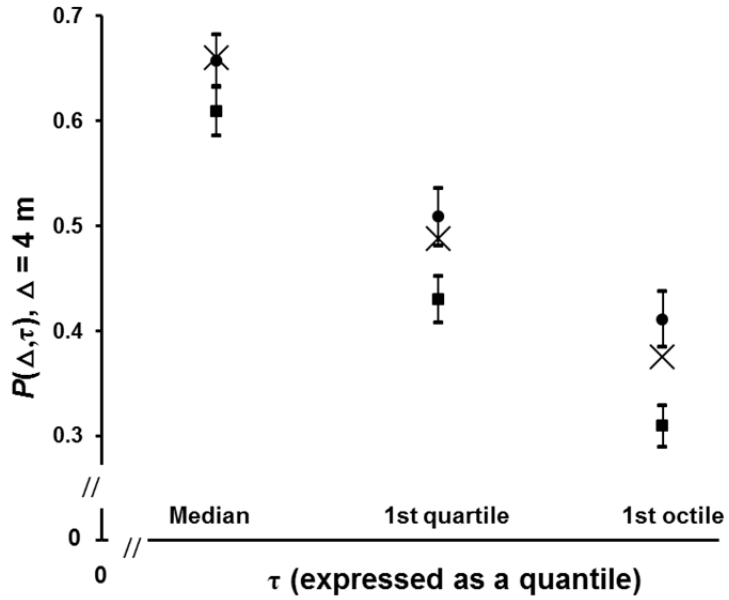


Figure 4.10 $P(\tau,\Delta)$ with Δ = 4 m plotted against τ set to the median, first quartile or first octile. The solid disc (●), is the mean value from 10 000 realizations of the PCLT random function, the solid square (■) is the mean value from 10 000 realizations of the TG random function. The horizontal bars show the 95 % confidence interval for the mean of based on 250 independently and randomly selected locations that mean the conditioning criteria. The crosses (x) show the mean values for 250 independently and randomly selected sites in the ECa data set.

Note that for all three thresholds the values of $P(\tau, \Delta)$ for the data are larger than the upper limit of the confidence interval for the TG process. For τ equal to the median and the first quartile the values from the data are within the confidence interval for the PCLT process, for the first octile the estimate is slightly smaller than the confidence interval for the PCLT process, but closer to the expected value for the PCLT process than it is for the TG process.

4.6 Discussion

The overall objective of this study was to identify a stochastic model for a soil property that varies according to some mode, and to base this identification as far as possible on knowledge of the underlying soil process and, at most, some simple descriptive statistics of the variable such as the empirical variogram and summary statistics. This was achieved in this study by employing general soil knowledge in a structured way. This is proposed as a framework for similar studies on soil variation in contrasting modes.

The particular value of this approach is shown by the fact that the inverse-distance PCLT model was better than the TG model in terms of the test statistic on the connectivity of values with small ECa. If one wanted to generate conditional simulations of the soil in this environment as a basis for computing, for example, distributions of upscaled processes such as pollutant transport across a block of land, then the inverse-distance PCLT model would produce superior representations of the connectivity of material with large conductivities, and so of preferential flow pathways.

There is considerable scope for further development of this approach. Other distance functions could be considered for this variable, and for others. In this study we looked for the simplest distance function that seemed to be compatible with soil knowledge, and there may be scope further to refine a framework for selecting a function. More specific soil knowledge could be used. For example, in the case study considered here, one could generate a simple conceptual 3D model of a polygon, with material with different dielectrical properties, and compute the expected trend function from models of the EM properties of the soil. While the objective of this particular study was to restrict the use of direct observations on the target variable to simple descriptive statistics, one might also conduct specific surveys at fine scale on transects across polygons in order to identify plausible distance functions for further studies.

The model-fitting framework in this study made combined use of point estimates of the variogram, and a weighted least squares criterion for parameter estimation, subject to constraints identified from soil knowledge. Ultimately a likelihood framework is required for estimation. It is not straightforward to develop this for such non-Gaussian variables, but more general consideration of soil knowledge might help us to identify a limited set of

distance functions that are of interest, and for which the appropriate likelihood framework could then be developed.

There is scope for further work on the comparison of realizations of the PCLT and TG processes with respect to multiple-point statistics and for weighing the evidence that one model rather than the other best represents particular data. We used a relatively simple criterion in this chapter, given that our data are not-quite regularly sampled and so do not constitute an image. However, it would be interesting to see how statistics developed for images (e.g. De Iaco and Maggio, 2011) might be adapted to irregularly sampled data. That said, the statistic which we used in this chapter was not a general measure of spatial structure but rather was focussed on a particular problem of direct interest (i.e. the connectedness of areas likely to have larger hydraulic conductivities). This is arguably more relevant than a generalized measure. It would be interesting to develop methods to quantify the spatial structure of RFs as this affects particular processes. For example, one might compare the outcomes of a process model for the dispersal of contaminant plumes when it is run with input data on conductivity or similar model parameters which are realizations of contrasting random processes.

Any PCLT model could be used in conventional spatial prediction by kriging since the variogram or, equivalently, the covariance function can be specified. However, since the PCLT covariance function is not available in closed form, it would generally be more efficient to use a standard variogram function for kriging. Furthermore, one might generally want to transform a non-Gaussian variable prior to kriging. The value of the PCLT model is not to provide an alternative form of the covariance function, but rather for spatial prediction of non-Gaussian variables whose multivariate distribution is not entirely characterized by the covariance function. Spatial prediction in such cases may be may be done by codes such as SNESIM (Strebelle, 2002) or the direct sampling method of Mariethoz et al. (2010) which allow one to obtain the distribution function of random variables at unsampled sites from multiple realizations of a non-Gaussian process. These procedures require TIs of the variables of interest, and the availability of sufficient TIs of adequate quality is a potential limitation on the use of MPG methods in soil science. For this reason Pyrcz et al. (2008) developed a library of training images for a particular geological setting (fluvial and deepwater reservoirs) by a combination of stochastic and object-based simulation methods. If an appropriate PCLT process could be identified for a particular soil variable, then it might be used similarly to generate TIs, either for a library or as required for a MPG simulation.

4.7 Conclusions

We have developed an alternative non-Gaussian RF that is able to model the polygonal pattern of the ECa measurements. The appropriate stochastic geometric model was fit through a structured use of soil knowledge. We have shown that this model appears to capture features of the spatial variation of our target variable better than the standard Gaussian model, even after transformation of the data to marginal normality. There is more work to be done in the development of this approach, and exploring its practical implications but we believe this case study shows that there is considerable potential.

In particular, realizations of PCLT processes may be better than standard TG simulations for predicting outcomes of non-linear processes such as contaminant transport, and for quantifying the uncertainty of such predictions. If PCLT models succeed in capturing the multiple-point behaviour of soil variables, then PCLT simulation could be used to provide an inexhaustible supply of TIs for existing MPG algorithms.

Chapter 5

Performing MPG simulations with the Direct Sampling algorithm

The content of this chapter is based on: Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M. and Renard, P. 2013. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. *Computers & Geosciences* 52, 307–324.

This chapter analyses the Direct Sampling (DS) algorithm, the MPG algorithm that we selected as the most appropriate to reconstruct the polygonal network (chapter 3), because it allows categorical, continuous and multivariate simulations (Mariethoz et al., 2010). It is important to understand precisely the capacities of DS and its sensitivity to the user-defined input parameters. However, DS was only recently developed, and there is little supporting material to help users getting started. Therefore, the first section of this chapter describes the DS workflow in general and the second section reports a more detailed explanation and a comprehensive sensitivity analysis on the DS input parameters. Two of the seven TIs used for the sensitivity analysis were ice-wedge polygonal network TIs.

5.1 Introduction

The Direct Sampling (DS) algorithm is a recent MPG algorithm (Mariethoz et al., 2010) that is the object of an international patent application (PCT/EP2008/009819). The code is available on demand for academic and research purposes. DS is implemented in the ANSI C language and all input and output files are in an ASCII SGeMS compatible format (Remy et al., 2009).

5.2 Theory

Figure 5.1 shows the workflow of the DS algorithm. If there are conditioning data available, these are first assigned to their closest grid nodes in the simulation grid. Conditioning data are generally point observations, that can be either categorical or continuous, but it can also be transect or (quasi) exhaustive samples. When no conditioning data are available, DS will generate unconditional simulations. Then, a path is defined through the (remaining) locations to be simulated **x**. This path is usually random, but the user has the option to define a unilateral path.

For each sequentially visited **x**, DS defines the data event $\mathbf{d}_n(\mathbf{x})$ consisting of the n closest neighbours (including conditioning data and previously simulated grid nodes) within the defined search area. The user defines the maximum number of neighbours n_{max} and the maximum search area. This search area can be defined by setting the parameters 'maximum search distance', i.e. the radius in the x-, y- and z- direction of a rectangular search area. Generally, it is advised to use a large search area by setting the radii to half the size of the simulation grid, corresponding to the maximum neighbourhood size. Making n_{max} the only limiting factor results in data events that cover a large part of the search area when the first unknown grid nodes are simulated, and a progressive decrease of the size of the area covered by $\mathbf{d}_n$ when the number of already simulated nodes increases. Consequently, DS ensures that patterns at different scales are present in the simulation, which is also the purpose of the multi-grid approach in SNESIM (Strebelle, 2002).

Next, a TI scan is performed, as illustrated in Figure 5.2. For a random location **y** in the TI grid, the TI pattern $\mathbf{d}_n(\mathbf{y})$ is defined that has the same data geometry as $\mathbf{d}_n(\mathbf{x})$. The dissimilarity between the data values of $\mathbf{d}_n(\mathbf{x})$ and $\mathbf{d}_n(\mathbf{y})$ is quantified by a distance measure $D\{\mathbf{d}_n(\mathbf{x}), \mathbf{d}_n(\mathbf{y})\}$. As soon as D is smaller than a user-defined acceptance threshold t, the value at the central node of this TI pattern $z(\mathbf{y})$ is assigned to $z(\mathbf{x})$ in the simulation grid. If D is larger than t, the TI scan continues. This acceptance threshold t needs to be defined because a TI pattern matching $\mathbf{d}_n(\mathbf{x})$ exactly is often not found, especially for continuous variables. For each variable type, one only has to select the appropriate dissimilarity distance D, making DS a flexible technique. The default distance type for categorical variables is the fraction of non-matching nodes between $\mathbf{d}_n(\mathbf{y})$ and $\mathbf{d}_n(\mathbf{x})$. For continuous variables, it is the sum of the absolute value of the differences between the corresponding data values in $\mathbf{d}_n(\mathbf{y})$ and $\mathbf{d}_n(\mathbf{x})$. The latter is normalised, so both dissimilarity distances range between zero (exact match) and one (no match).

The user can decide about the maximum fraction of the TI that is scanned for each **x** by setting parameter f, ranging between 0 (no scan) and 1 (scan full TI if necessary). If this maximum fraction has been scanned and still no TI pattern with $D < t$ has been found, DS assigns the central node of the TI pattern with the smallest D to $z(\mathbf{x})$. When no neighbour

is found for $\mathbf{x}$, for instance for the first node of an unconditional simulation, DS randomly samples a node $\mathbf{y}$ in the TI and assigns its value $z(\mathbf{y})$ to $z(\mathbf{x})$ in the simulation grid.

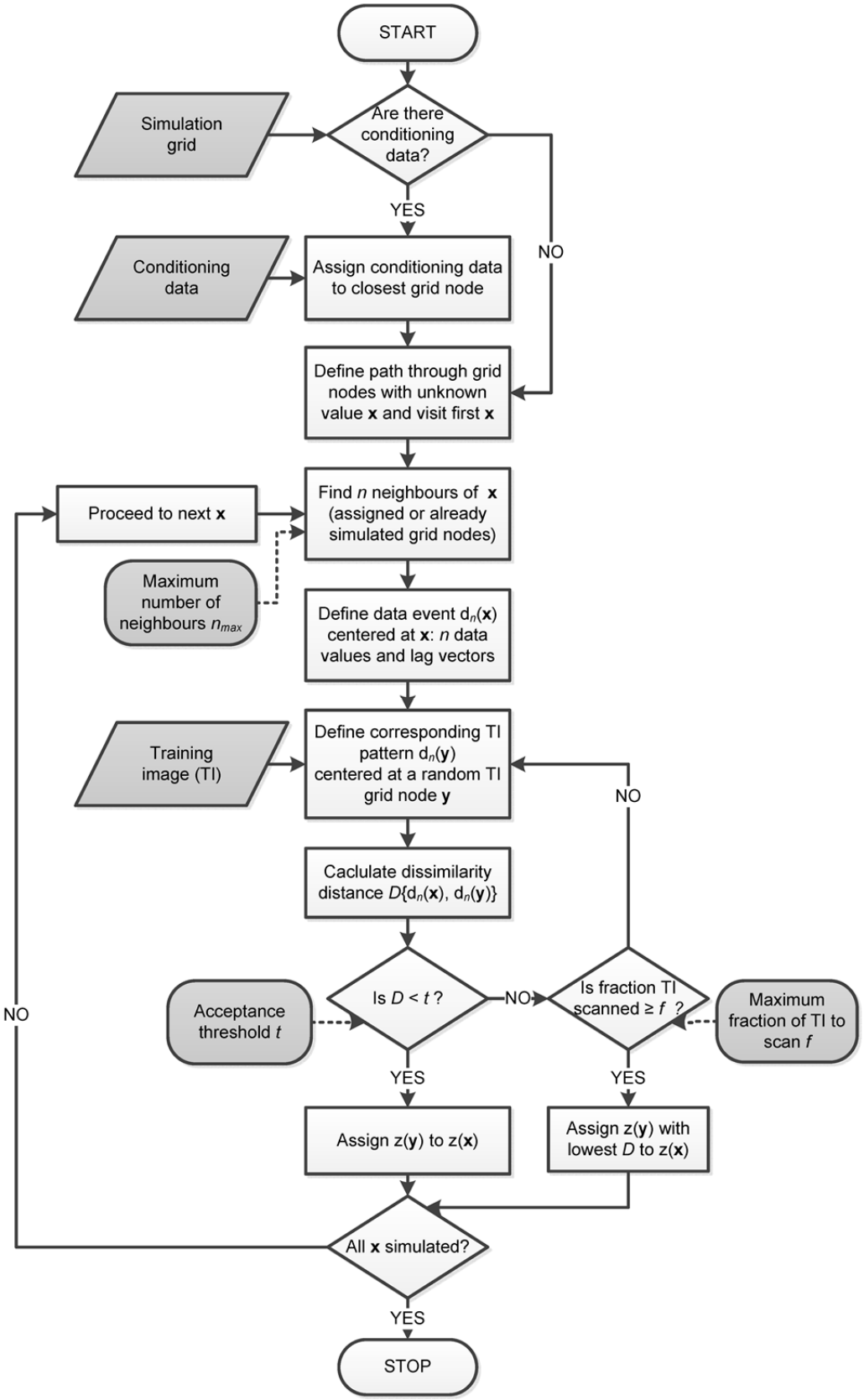


Figure 5.1 Workflow of the Direct Sampling algorithm.

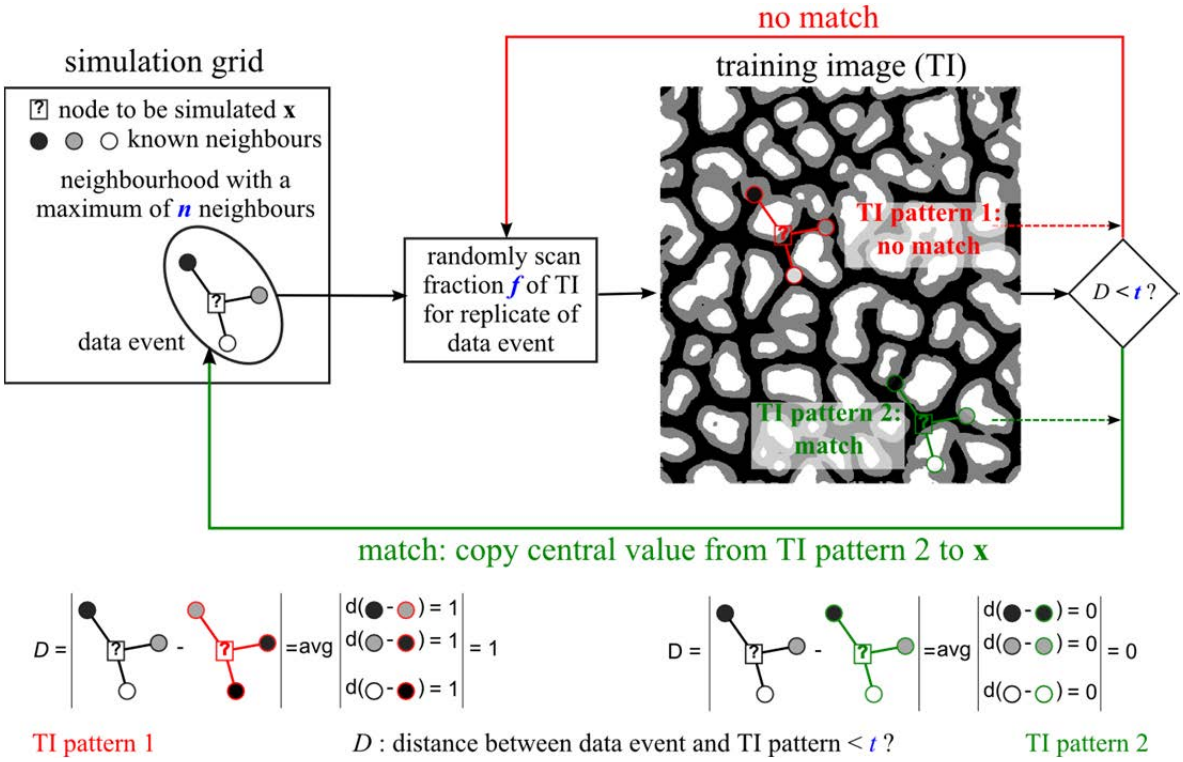


Figure 5.2 Illustration of how the Direct Sampling algorithm scans the TI focusing on the three main input parameters: *t*, *f* and *n*.

5.3 Sensitivity analysis on the Direct Sampling algorithm

Using DS requires the user to define some parameters: among them, the acceptance threshold *t*, the maximum fraction of TI to scan *f* and the maximum number of points in the neighbourhood *n* are the most important since they are balancing simulation quality and CPU time. For these three parameters, we report a detailed sensitivity analysis by generating non-conditional simulations for the entire 3D parameter space. Next to a visual inspection of the resulting simulations, we quantified the similarity between the simulations and the TI by means of simulation quality indicators (CASE 1). The same quality indicators were calculated for a 3D example (CASE 2). We also illustrated the potential of the post-processing option (CASE 3), the multivariate simulation option (CASE 4) and the data conditioning option (CASE 5) and discussed the corresponding user-defined input parameters. Table 5.1 summarizes the values of the parameter that we kept fixed and the range of values of the parameters that we varied.

Since Mariethoz (2010) already showed good performance of DS with as many as 54 processors, the parallelization option is not discussed here. For more information about the option to use transform-invariant distances we refer to Mariethoz and Kelly (2011).

Table 5.1 Fixed parameters with their default values chosen for this study (sorted according to their appearance in the parameter file) and parameters that were varied with their default values and range over which they were varied (sorted according to the case number in which they were studied).

Fixed parameters	
Name	Default
Simulation method	MPS
Number of realizations	10
Max search distance	125 125 0 (½ size simulation grid)
Anisotropy ratios in the search window (x,y,z)	1 1 1
Transformations	0 (no transformations)
Path type	0 (random path)
Type of variable	0 for categorical, 1 for continuous
Exponent of the distance function in the template	0
Syn-processing parameters (4)	0 0 0 0 (no syn-processing)
Initial seed	1350
Parameters reduction	1 (no parameters reduction)
Parallelization	1 (serial code, no parallelization)

Varied parameters			
Name	Default	Range	Case
Threshold position t	0.05	0.01 – 0.02 – 0.04 – 0.06 – 0.08 – 0.1 – 0.12 – 0.14 – 0.16 – 0.18 – 0.2 – 0.25 – 0.5 – 0.75 – 0.99	1
Max fraction of TI to scan f	0.5	0.05 – 0.1 – 0.15 – 0.2 – 0.3 – 0.4 – 0.5 – 0.6 – 0.75 – 1	1
Max number of points in neighbourhood n	50	1 – 5 – 10 – 15 – 20 – 30 – 50 – 80	1
Post-processing parameters	0	0 – 1 – 2	2
- Number of post-processing steps (p)	0	0 – 1 – 3	
- Post-processing factor (p_f)			
Number of variables to simulate jointly	1	1 – 2	3
Relative weight of each variable	1	0.1 0.9 – 0.25 0.75 – 0.5 0.5 – 0.75 0.25 – 0.9 0.1	3
Weight of conditioning data (δ)	1	0 – 1 – 5	4
Data conditioning	no	no – yes	4

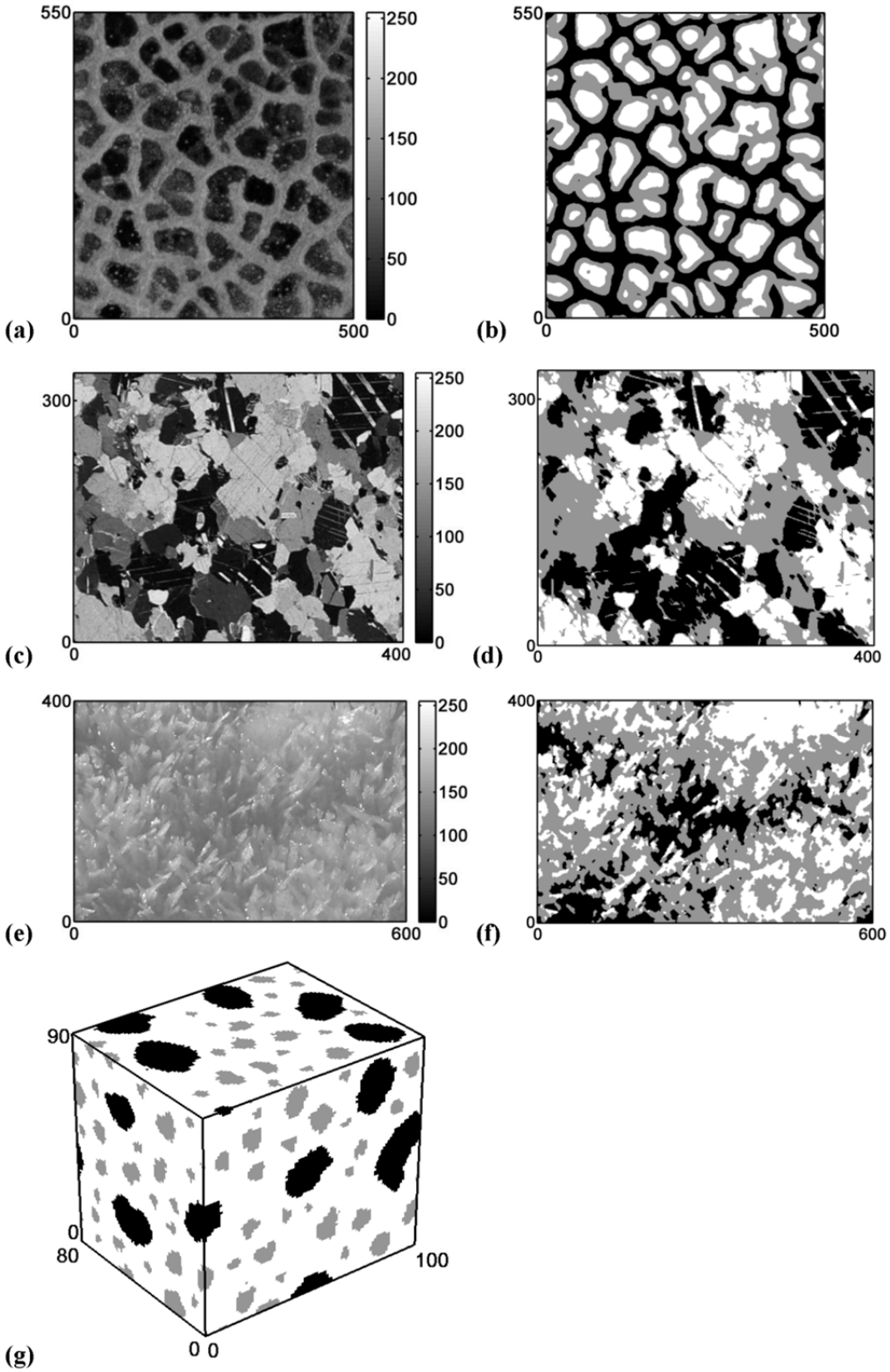


Figure 5.3 The seven training images (TIs) that were used for the sensitivity analyses: (a) continuous (photograph: Plug and Werner, 2002) and (b) categorical (k = 3) TI of ice-wedge polygons; (c) continuous and (d) categorical (k = 3) TI of a thin marble slice; (d) continuous and (e) categorical (k = 3) TI of snow crystals; (g) categorical (k = 3) 3D TI of a block of concrete The x-, y- and z-axes represent the number of pixels. The results of the sensitivity analyses for TIs (b), (c) and (g) are illustrated in this chapter; the results for TIs (a), (d), (e) and (f) can be found as supplementary material in Appendix.

Many previous studies have used only one TI with sinuous channels (Figure 2.3 – left). In contrast, we include a greater variety of patterns by performing sensitivity analyses on seven TIs: an image of ice-wedge polygons (Plug and Werner, 2002), a microscopic view of a thin marble slice, an image of snow crystals, all three as categorical and continuous images, and a categorical 3D image of concrete (Figure 5.3). The continuous 2D TIs are grayscale photographs with pixel values between 0 and 255; the categorical 2D TIs are derived from these by classifying them into three categories. The 3D TI is generated by sequentially simulating 2D slices constrained by conditioning data computed at the previous simulation steps (Comunian et al., 2012).

The figures shown in this chapter are the results for the categorical ice-wedge TI, the continuous marble TI and the 3D concrete TI (3D case). They are presented with the same color scale as the TIs in Figure 5.3. The results for the other TIs can be found as supplementary material in Appendix.

5.3.1 CASE 1:parameters balancing CPU time and simulation quality: *t*, *f* and *n*

It is clear that the larger n and the closer t to 0 and f to 1, the better the simulation quality will be. However, these settings will be very expensive in terms of CPU time. For the six 2D TIs, we simulated 10 unconditional realizations for each parameter combination of 15 t values, 10 f values and 8 n values (Table 5.1), resulting in 12 000 realizations for each TI.

- **CPU time**

Figure 5.4 shows the CPU time needed to simulate one unconditional simulation using the categorical ice-wedge TI and one using the continuous marble TI. First the influence of t and n is shown for f = 0.5 after which the influence of f is shown for different combinations of t and n.

Besides the fact that generating simulations based on the continuous TI generally takes longer, the results for the categorical and the continuous case show a similar behavior. Simulations with small t and large n require a long simulation time and decreasing f strongly reduces CPU time. Modifying one of the parameters t, f or n increases or decreases the CPU exponentially. The combined effect of relaxing all three parameters only slightly, can reduce CPU time significantly. For instance, generating one simulation for the categorical case with default parameters (t = 0.05, f = 0.5, n = 50) takes 163 s. Relaxing t to 0.1 only takes 44 s, relaxing all three parameters to t = 0.1, f = 0.3 and n = 30 only takes 13 s.

This behavior is related to the scanning algorithm. When t is close to 0, f close to 1 and n large, the algorithm scans the entire TI for a very good match with complex data events (large neighbourhoods). This takes a lot of time. In the opposite case, the algorithm finds very quickly a TI pattern that matches the criteria and the algorithm is fast. What is striking

in Figure 5.4 is that the evolution between these two cases is rather abrupt for some parameter values. When the parameter values are beyond such an abrupt boundary, DS is very fast whatever the parameter values, below this boundary CPU time increases.

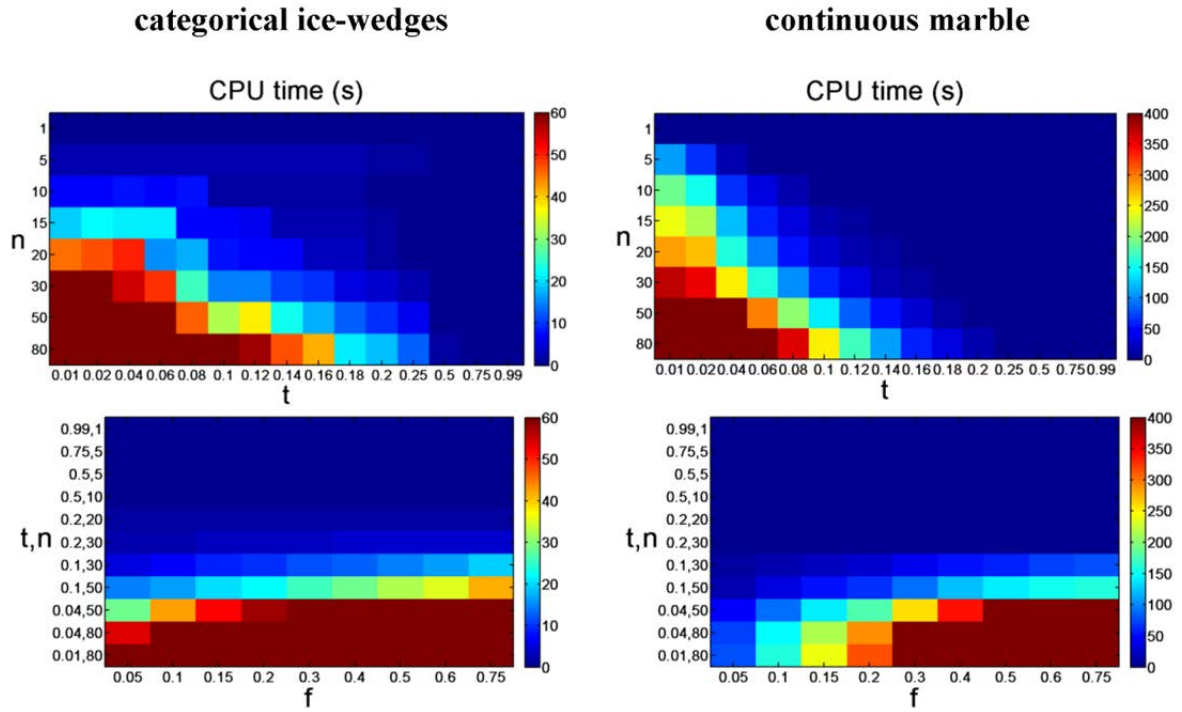


Figure 5.4 Influence of *t* and *n* (for *f* = 0.5) (top) and *f* (bottom) on the CPU time required to run one unconditional simulation.

- **Visual quality inspection**

Figure 5.5 shows the first out of 10 simulations for some combinations of *t*, *f* and *n* using the categorical ice-wedge TI and Figure 5.6 using the continuous marble TI. The results for the other TIs can be found as supplementary material. Similar as for Figure 5.4, first a sensitivity analysis on *t* and *n* is performed (for *f* = 0.5), after which the effect of *f* is illustrated for some combinations of *t* and *n*. We select simulations with different quality levels in order to illustrate the evolution of the simulation quality. As the quality steps depend on the TI, simulations with different *t* and *n* values are illustrated for each case.

For the categorical case, running DS with t > 0.5 or n ≤ 5 results in noisy images. This is not surprising since then the sampling is not selective enough: any $\mathbf{d}_n(\mathbf{y})$ can be accepted even if it is far away from $\mathbf{d}_n(\mathbf{x})$. This corresponds to situations in which the algorithm is very fast. For $t \leq 0.5$ and $n > 5$, the ice-wedge pattern is reasonably well reconstructed. For $t \leq 0.2$ and $n \geq 30$, the simulation quality is very good. Not only the pattern reconstruction, but also the appearance of noise and the fuzziness of the edges between different categories are influenced by *t* and *n* (CASE 3). For the categorical marble TI (Figure 5.3d) similar thresholds were found (Appendix - Figure B).

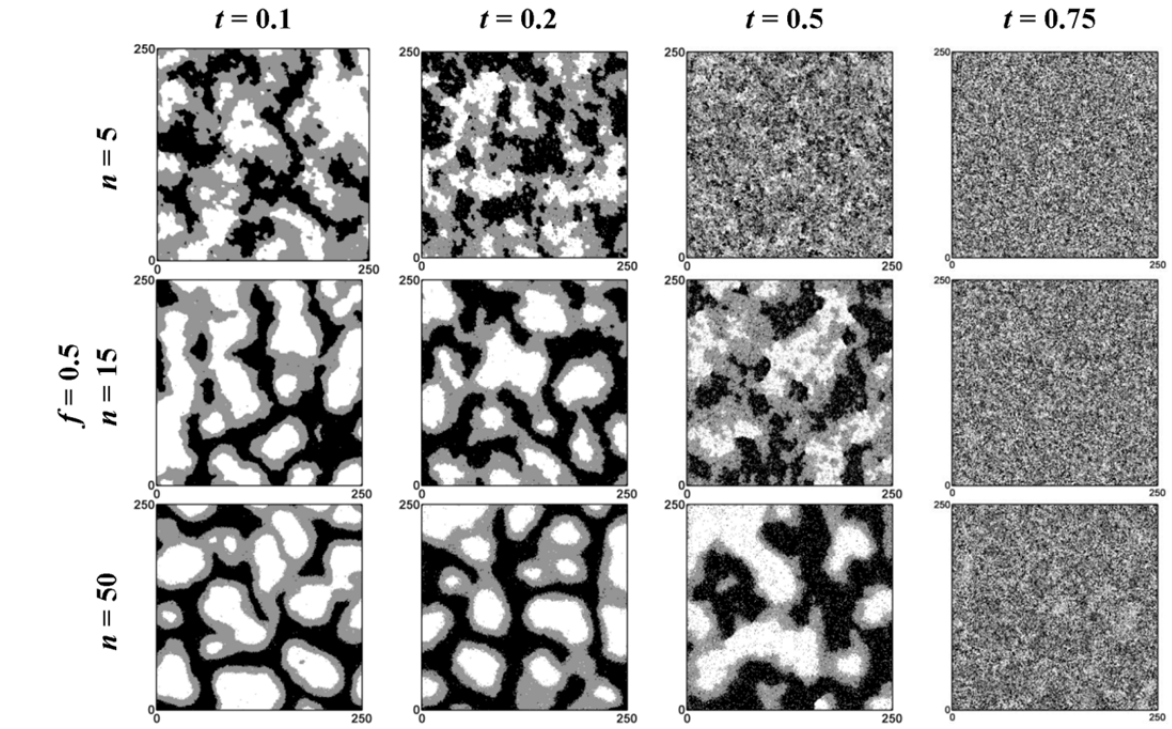


Figure 5.5 (a) First out of 10 unconditional simulations illustrating the effect of parameters *t* and *n* with *f* = 0.5 and (b) first out of 10 unconditional simulations illustrating the effect of *f* for constant *t* and *n* based on the categorical ice-wedge TI (Figure 5.3b).

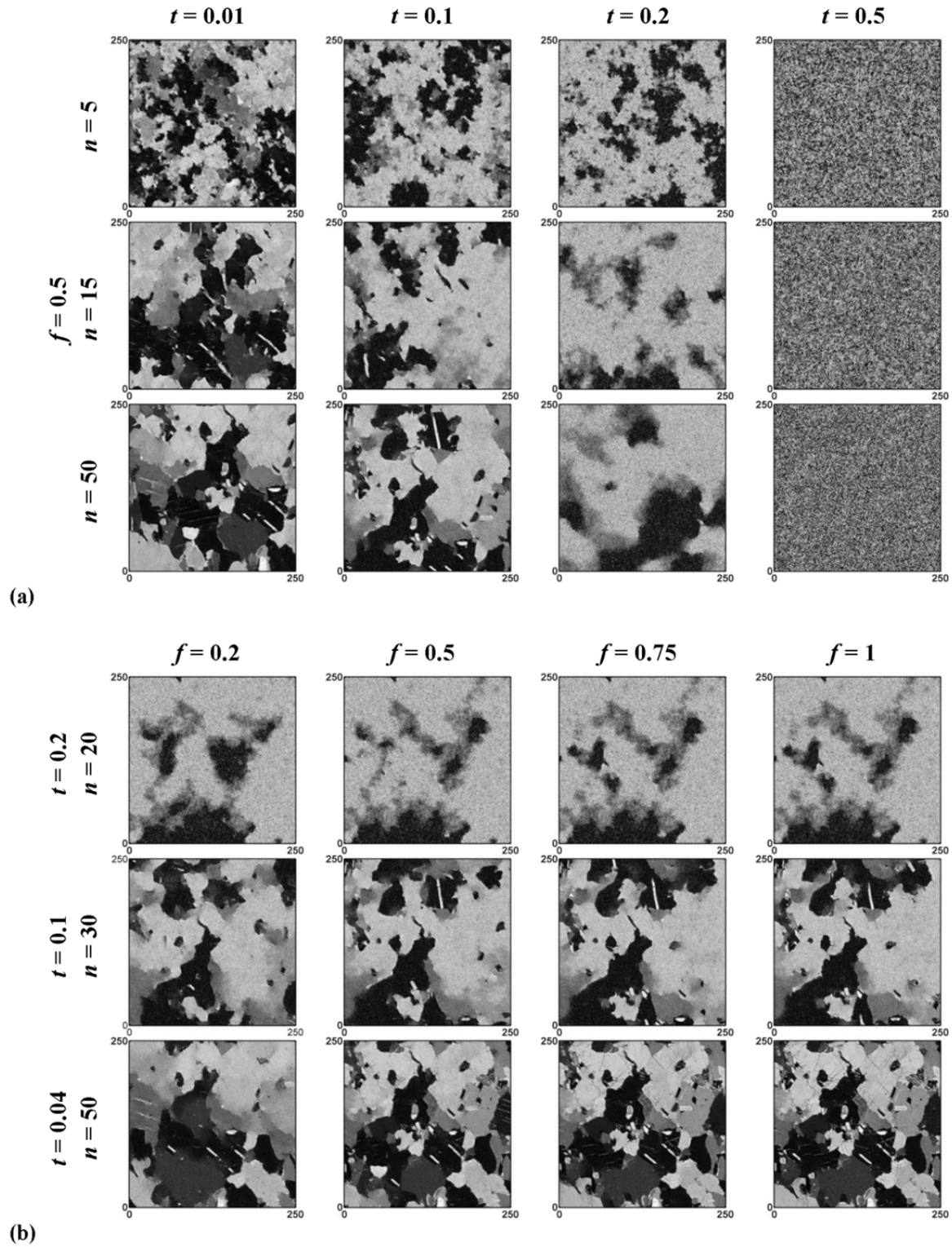


Figure 5.6 (a) First out of 10 unconditional simulations illustrating the effect of parameters t and n with $f = 0.5$ and (b) first out of 10 unconditional simulations illustrating the effect of f for constant t and n based on the continuous marble TI (Figure 5.3c).

For the categorical snow crystals TI (Figure 5.3f) the simulation quality is good for $t \leq 0.1$ and $n \geq 50$ (Appendix – Figure C). In contrast to the effect of t and n, f has a much

smaller effect on the simulation quality. Scanning a smaller part of the TI hardly results in a quality decrease (Figure 5.5b). The same conclusion can be drawn from the simulations using the other categorical TIs (Appendix – Figure B and C).

Figure 5.6 shows that generating continuous simulations generally requires stricter parameters (smaller *t*, larger *n* and *f*). Running DS with $t \geq 0.2$ results in noisy images. The simulation quality is good for $t \leq 0.1$ and $n \geq 30$. Simulations using the continuous ice-wedge TI (Figure 5.3a) show important changes in visual quality for the same values of *t* and *n* (Appendix – Figure A). The quality of the simulations using the continuous snow crystal TI (Figure 5.3e) is generally less good: only for $t \leq 0.1$ and $n \geq 50$ the simulation quality is moderate (Appendix – Figure D). For the continuous cases, it is observed that variations in *f* do not affect much on the simulation quality.

Especially for the continuous cases, it can be seen that some simulations are almost exact copies of parts of the TI. This phenomenon is called 'patching'. It is caused by copying each time the central node of the same best matching pattern. The issue of patching will be discussed further in this chapter.

- **Simulation quality indicators**

For each unconditional simulation we calculated several quality indicators by comparing the histogram, variogram and connectivity function of the TI and the simulations. The connectivity function $\tau(\mathbf{h})$ for a category *s* is defined as the probability that two points separated by a lag vector $\mathbf{h}$ are connected, denoted as $\mathbf{x} \leftrightarrow \mathbf{x}+\mathbf{h}$, by a continuous path of adjacent cells all belonging to *s*, conditioned to the fact that the two points belong to *s* (Boisvert et al., 2007; Emery and Ortiz, 2011; Renard et al., 2011; Renard and Allard, 2012):

$$\tau(\mathbf{h}) = \text{Prob}\left\{\mathbf{x} \leftrightarrow \mathbf{x}+\mathbf{h} \,\middle|\, s(\mathbf{x}) = s, s(\mathbf{x}+\mathbf{h}) = s\right\}. \qquad (5\text{-}1)$$

Figure 5.7 compares the histogram, variogram and connectivity function of the categorical ice-wedge TI (Figure 5.3b) with these of a good simulation ($t = 0.01$, $f = 0.5$, $n = 80$) and a bad simulation ($t = 0.5$, $f = 0.5$, $n = 15$ for categorical and $t = 0.2$, $f = 0.5$, $n = 5$ for continuous). Both the indicator variogram values $\gamma^k(\mathbf{h})$ and the connectivity function $\tau^k(\mathbf{h})$ for each category k are calculated for 20 lag classes $\mathbf{h}$ with a lag width of 5. The histograms (proportions of the three categories) are represented for both simulations. The indicator variograms and the connectivity functions are only reproduced for the good simulation, except for the intermediate material (grey), where the bad simulation partially reproduces the TI statistics.

Figure 5.8 illustrates the same for the continuous case. Here the standard variogram $\gamma(\mathbf{h})$ is calculated instead of the indicator variogram. To calculate the connectivity functions, the TI and the simulations were first divided into three categories based on two thresholds representing connectivity jumps in the TI (Renard and Allard, 2012). Similarly

to Figure 5.7, the histograms (represented as the cdf) were represented for both the good and the bad simulation, whereas the variogram and the connectivity functions were only reproduced by the good simulation.

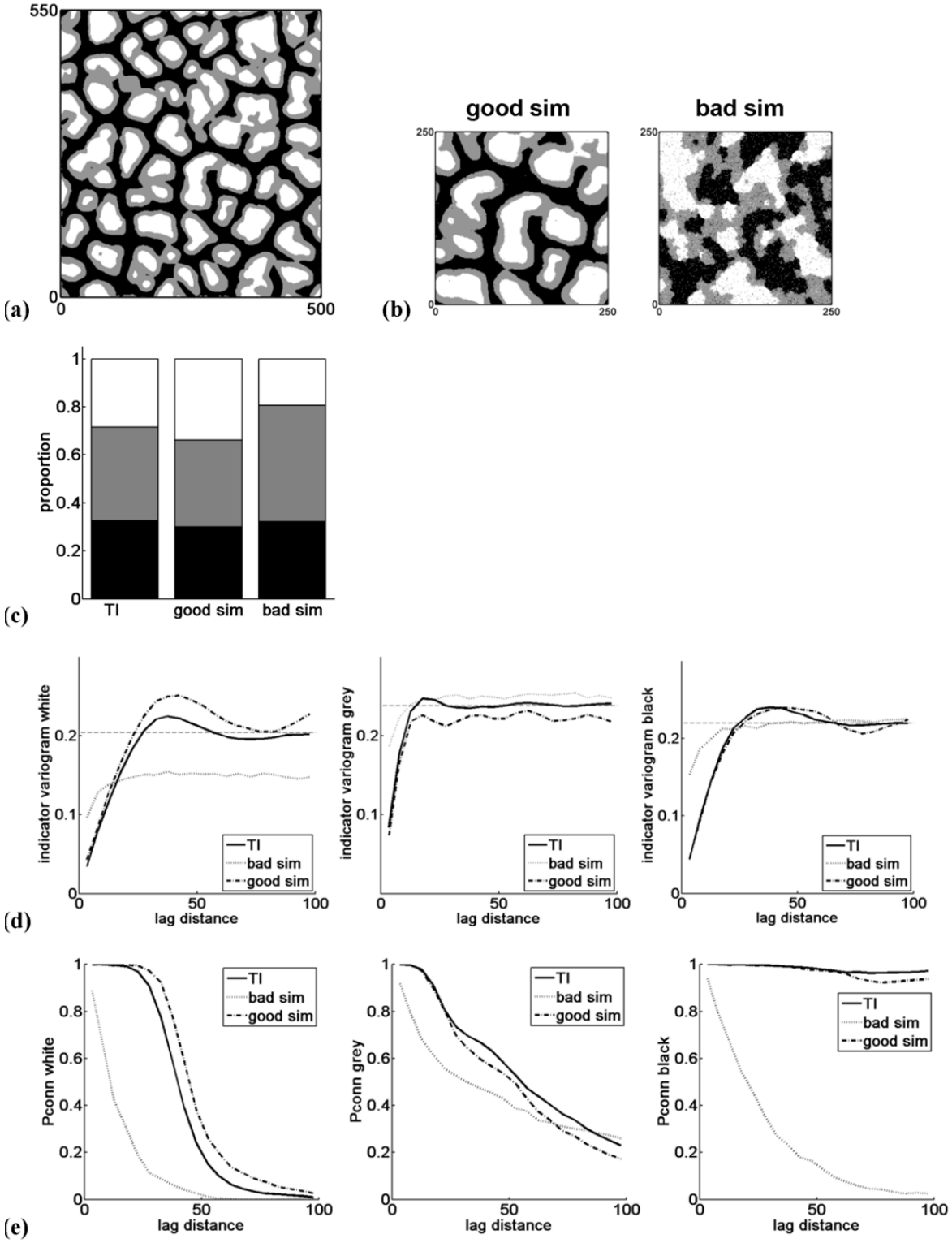


Figure 5.7 Reproduction of (a) the categorical ice-wedge TI statistics of (b) a good (t = 0.01, f = 0.5, n = 80) and a bad simulation (t = 0.5, f = 0.5, n = 15) with the reproduction of (c) the histogram, (d) the indicator variograms (the dotted lines represent the TI indicator variance) and (e) the connectivity functions.

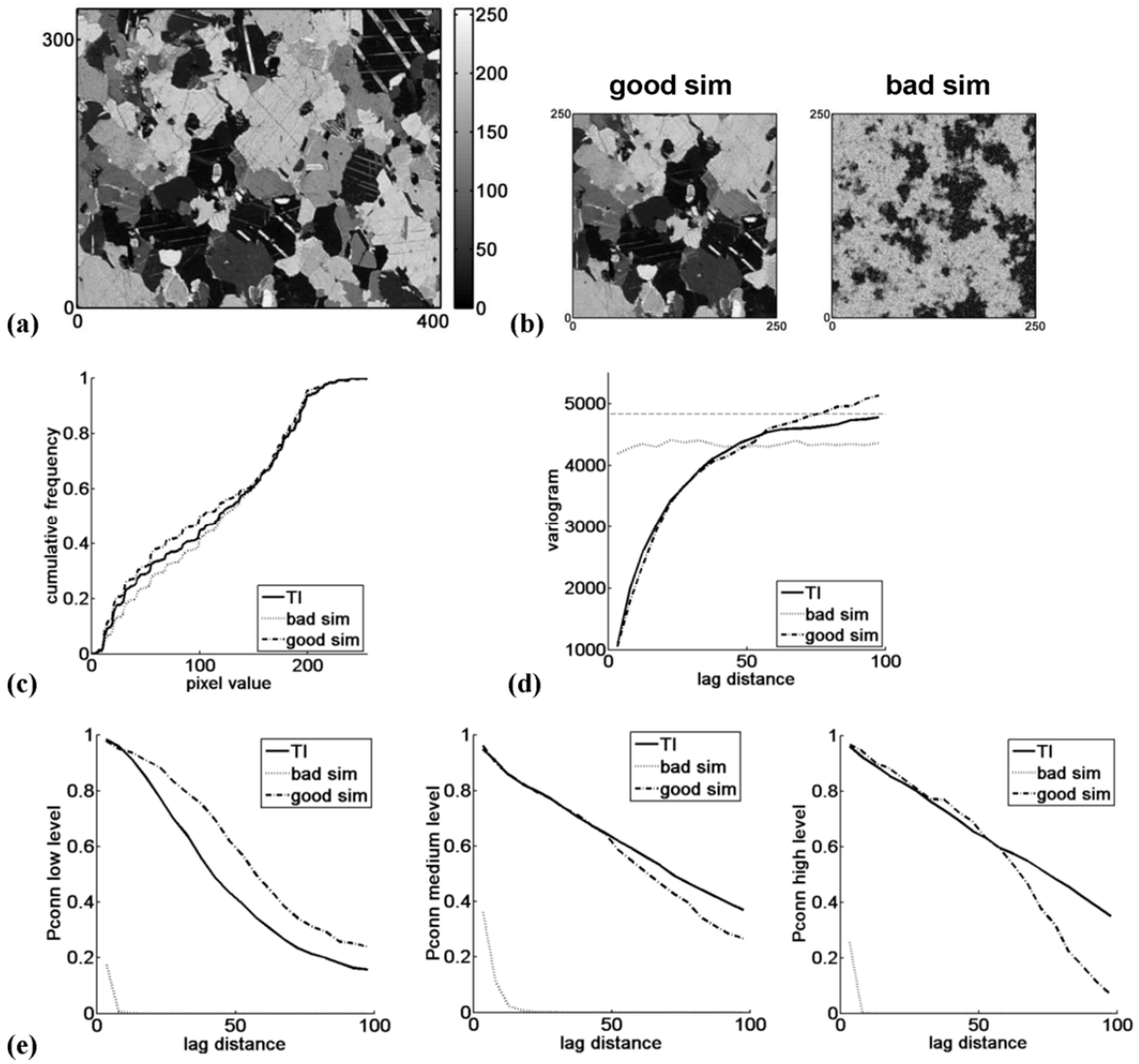


Figure 5.8 Reproduction of (a) the continuous marble TI statistics of (b) a good (t = 0.01, f = 0.5, n = 80) and a bad simulation (t = 0.2, f = 0.5, n = 5) with the reproduction of (c) the histogram, (d) the variogram (the dotted line represents the TI variance) and (e) the connectivity functions.

To quantify the dissimilarity between the simulations' statistics and those of the TI, we calculated three error indices for each simulation: a histogram error ε_{hist}, variogram error $\varepsilon_{\mathrm{var}}$ and connectivity error ε_{conn}. For the categorical simulations, ε_{hist}^{k} was defined as the absolute value of the difference between the proportion of k in the simulation grid f_{sim}^{k} and in the TI f_{TI}^{k}. For the continuous simulations, ε_{hist} is calculated as the Kullback–Leibler divergence D_{KL} (Kullback and Leibler, 1951):

$$\varepsilon_{hist} = D_{KL}(P\|Q) = \sum_{i} P(i)\log\frac{P(i)}{Q(i)} \qquad (5\text{-}2)$$

with $P(i)$ the probability distribution in the simulations and $Q(i)$ the probability distribution in the TI.

The variogram error ε_{var} for the continuous simulations is based on the weighted average difference between the variogram values of the simulations $\gamma_{sim}(\mathbf{h})$ and the TI $\gamma_{TI}(\mathbf{h})$ for 20 lag classes $\mathbf{h}_d$, and is calculated as

$$\varepsilon_{\text{var}} = \frac{\sum_{d=1}^{20} \frac{1}{\mathbf{h}_d} \frac{|\gamma_{sim}(\mathbf{h}_d) - \gamma_{TI}(\mathbf{h}_d)|}{\text{var}_{sim}}}{\sum_{d=1}^{20} \frac{1}{\mathbf{h}_d}} \qquad (5\text{-}3)$$

with var_{sim} the simulation variance used to standardize the absolute errors, so they range between 0 and 1. The term $1/\mathbf{h}_d$ was included to give larger weights to errors corresponding to small variogram lags. The variogram error $\varepsilon_{\text{var}}^{k}$ for the categorical case was calculated similarly using the indicator variogram values.

The connectivity error ε_{conn} was calculated as

$$\varepsilon_{conn} = \frac{\sum_{d=1}^{20} \left| \tau_{sim}^{k}(\mathbf{h}_d) - \tau_{TI}^{k}(\mathbf{h}_d) \right|}{20} \qquad (5\text{-}4)$$

and also ranges between 0 and 1.

- **Results and discussion**

Figure 5.9 and Figure 5.10 show the results of the simulation quality indicators for the categorical and the continuous case. The first part of the figures illustrates the effect of *t* and *n*, the second part the effect of *f*.

As was already concluded from Figure 5.7 and Figure 5.8, ε_{hist} behaves differently than ε_{var} and ε_{conn}. The histogram was generally well reproduced for all simulations. Noisy images reproduced the histogram the best, which was especially clear for the continuous case. This can be explained by considering the extreme case of $t = 1$. With such a setting, DS randomly samples values from the TI, resulting in a perfect reproduction of the marginal distribution ($\varepsilon_{hist} = 0$), but no reproduction of the spatial pattern (very large ε_{var} and ε_{conn}). For intermediate combinations of *t* and *n*, ε_{hist} is generally larger. In the areas with good simulation quality ($t \leq 0.2$ and $n \geq 30$) ε_{hist} behaves differently for the categorical and the continuous case. For the categorical case small *t* and large *n* guarantee both small ε_{hist} and good simulation quality (Figure 5.5, Figure 5.9). For the continuous case, the high quality simulations ($t \leq 0.2$ and $n \geq 30$) have larger ε_{hist} (Figure 5.5, Figure 5.10).

This counterintuitive result can be explained as follows: with small *t* and large *n*, the simulation has to honour very strong spatial constraints. When the structures are made of objects that are large with respect to the domain size, respecting such spatial constraints means to respect the connectivity of facies and the objects size, even if it contradicts the

target pdf. Because of a slight non-stationarity in the TI, the simulation can then follow the pdf of one specific part of the TI that is different than the rest. This can result in significant variability in the pdfs of the simulations. This is the opposite as the case of $t = 1$, where the TI distribution is honoured because there is no constraint on the spatial continuity. For the continuous ice-wedge TI (Figure 5.3a) and the snow crystal TI (Figure 5.3e), the histogram is well reproduced in the high quality simulations (Appendix – Figure E and H). Since certain applications require the histogram to be reproduced, this issue could be further addressed by the DS developers.

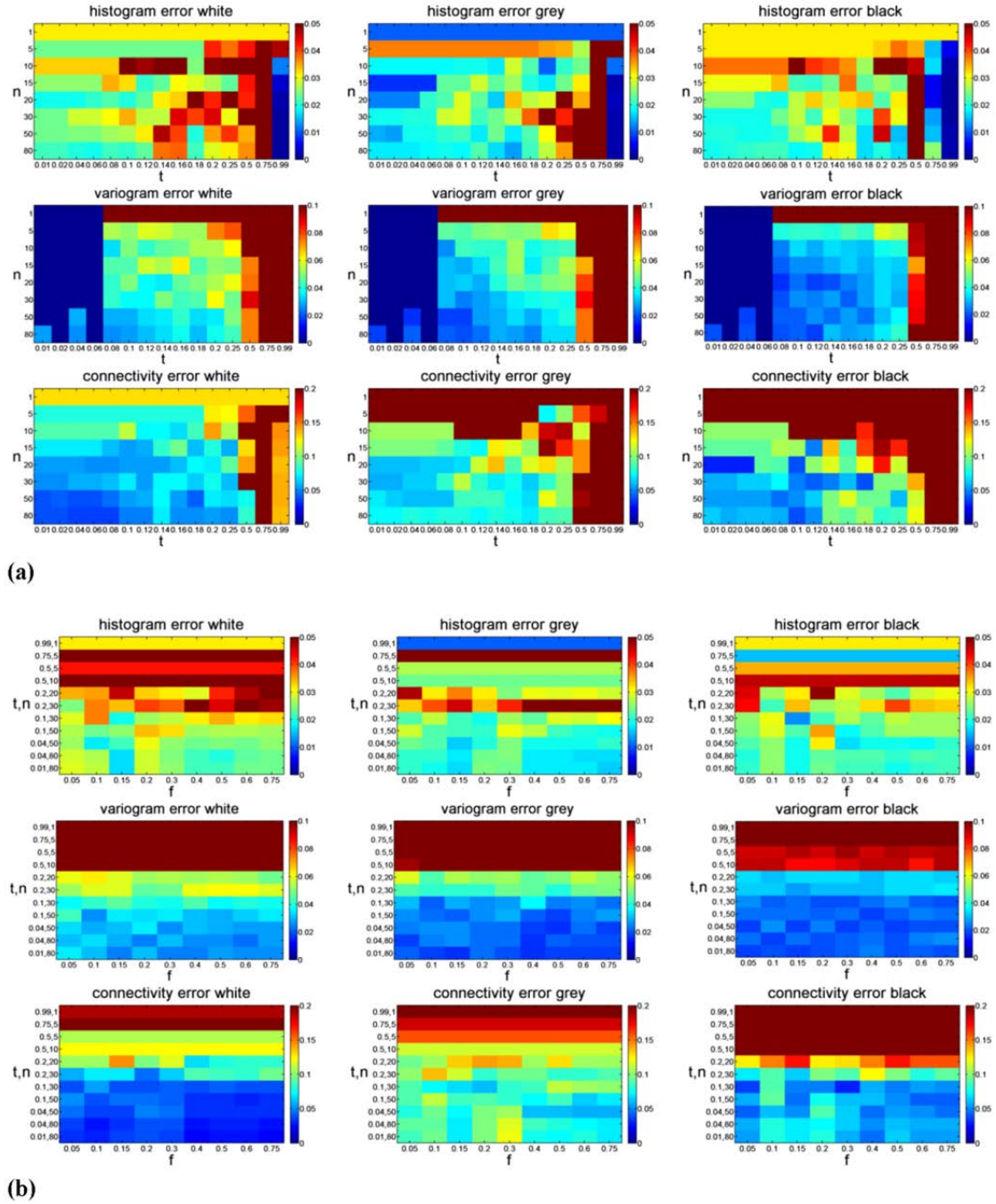


Figure 5.9 Influence of (a) t and n (for $f = 0.5$) and (b) f on the quality indicators based on the categorical ice-wedge TI (Figure 5.3b).

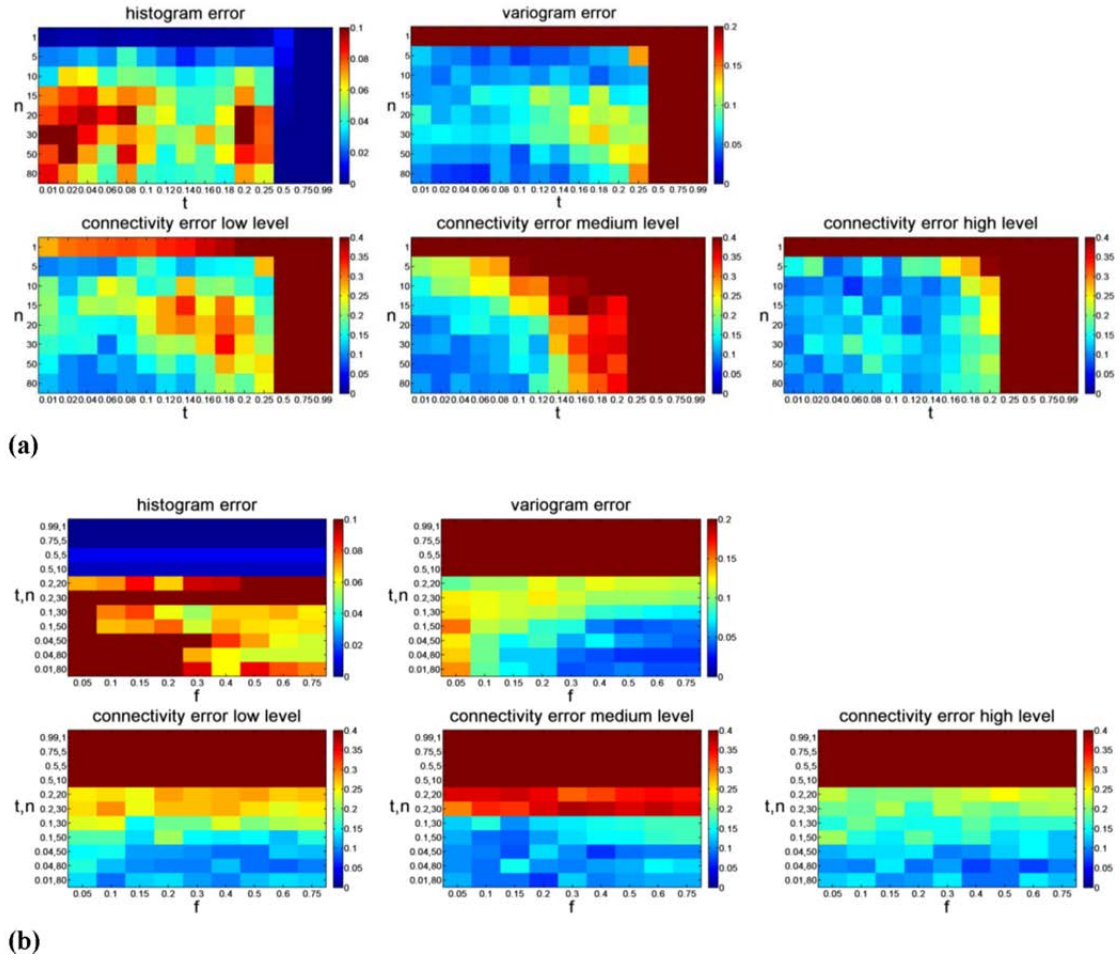


Figure 5.10 Influence of (a) *t* and *n* (for *f* = 0.5) and (b) *f* on the quality indicators based on the continuous marble TI (Figure 5.3c).

In contrast, ε_{var} and ε_{conn} increase for larger *t* and smaller *n*, which is a more intuitive behavior. Both errors show similar quality jumps as were derived from the visual inspection and therefore behave as stable simulation quality indicators.

These results allow us to derive some rules of thumb. Running DS using a categorical TI should result in good simulations for $t \leq 0.2$ and $n \geq 30$. Selecting $t \geq 0.5$ and $n < 15$ should be avoided. For continuous TI, it is advised to use $t \leq 0.1$, $n \geq 30$ and to avoid $t \geq 0.2$ and $n \leq 15$. The quality of intermediate combinations is hard to predict. It is obvious that the simulation quality steps strongly depend on the TI. The greater the pattern repeatability in the TI, the better the simulation quality will be.

Figure 5.9b and Figure 5.10b lead again to the conclusion that *f* does not have a strong influence on the simulation quality. This was confirmed by the other TIs (Appendix – Figure E,F,G and H). Only for certain situations, like for $f < 0.2$ in the continuous case (see results for ε_{var}), the pattern reproduction degraded with smaller *f* since the probability of finding a matching TI pattern was lower. Note also that using a small *f* value for TIs that contain insufficient diversity (Mirowski et al., 2009), might aggravate the statistical scarcity and lead to poor results. But generally decreasing *f* results in large computation

gains without a substantial decrease in simulation quality, which is an important conclusion for an efficient use of DS.

It is interesting to juxtapose the CPU time (Figure 5.4) with the corresponding quality indicators (Figure 5.9 and Figure 5.10). This reveals where the interesting boundaries are located in terms of quality over CPU time ratio. For instance, for the categorical case the quality is moderate from $n \geq 15$ and $t \leq 0.18$ ($f = 0.5$) (Figure 5.9a), whereas the CPU time really increases from $n \geq 30$ and $t \leq 0.1$ (Figure 5.4a). In between these boundaries, the simulation quality is good, as is confirmed by the visual inspection. In case CPU time is a limiting factor, users are recommended to investigate the quality over CPU time ratio for different parameter combinations running trial simulations on a small grid.

It is good practice to run DS initially with $f = 0.5$, t between 0.05 and 0.2 and n between 20 and 50. From this, the parameters need to be fine-tuned for every particular situation, knowing that decreasing t and increasing n and f should result in better simulation quality. However, one should keep in mind that using parameters which guarantee very good simulations has two drawbacks. First, these configurations will require large CPU times. Second, there is a risk of generating simulations which are all exact copies of (part of) the TI (patching effect or verbatim copy). This risk is higher when the TI does not show enough pattern repeatability (which is more often the case for continuous TIs) and when there are no conditioning data (CASE 5). Strategies to avoid patching are choosing $f < 1$, thus avoiding to pick each time the same best matching mode, slightly relaxing t and n, or using a smaller 'maximum search distance'.

5.3.2 CASE 2: 3D simulation

Similarly to two dimensions (CASE 1), DS can generate 3D simulations. To demonstrate this, we performed a limited sensitivity analysis using the 3D concrete TI (Figure 5.3g). We generated 10 unconditional simulations for each combination of eight t [0.01 – 0.05 – 0.1 – 0.15 – 0.2 – 0.25 – 0.3 – 0.5] and eight n [1 – 5 – 8 – 16 – 32 – 64 – 125 - 216] values, using a fixed value of 0.5 for f. The other parameters were set as indicated in Table 5.1, with exception of the maximum search distance that was defined as half of the search grid in three directions (x, y and z).

The CPU time as a function of t and n behaved very similar as for 2D (Figure 5.4). For instance, generating one simulation with $t = 0.1$ and $n = 32$ took 194 s, with $t = 0.05$ and $n = 32$ 1998 s and with $t = 0.05$ and $n = 64$ 6804 s.

Figure 5.11b shows the simulation quality indicators as a function of t and n. Overall, the results were analogous to those of CASE 1. The simulation quality generally improved with increasing n and decreasing t with a quality jump for $t = 0.2$ and $n = 32$, as can be seen from $\varepsilon_{conn}^{grey}$ and $\varepsilon_{conn}^{black}$, and the unconditional simulations shown in Figure 5.11a. Again, ε_{hist} was the smallest for noisy simulations with very small n.

Since the white category represented the background volume, $\varepsilon_{conn}^{white}$ was very small for all parameter combinations and hence not informative. With parameters producing noisy simulations ($t \geq 0.3$ and $n \leq 8$), $\varepsilon_{\mathrm{var}}^{grey}$ was again smaller. This can be explained by the small range of the grey indicator variogram, causing $\varepsilon_{\mathrm{var}}^{grey}$ to be small for pure nugget variograms reproducing the sill correctly.

5.3.3 CASE 3: Post-processing for noise removal

To further improve the simulation quality and more specifically to remove noise, DS includes a post-processing option. After having generated a DS realization, each node (except for the conditioning data locations) for which no TI pattern with $D < t$ had been found is re-simulated. Because all the unknown grid nodes had already been simulated, a data event in the post-processing step consists of the n closest grid nodes of $\mathbf{x}$. Whether or not this post-processing is performed should be decided by the user. Two post-processing parameters need to be defined: the number of post-processing steps p and the post-processing factor p_f. The latter is the factor by which f and n are divided aiming to save CPU time in the additional post-processing steps (Mariethoz, 2009). For example, if $p = 2$ and $p_f = 3$, all nodes are resimulated two times with parameters f and n 3 times smaller than their original values. Figure 5.12 illustrates the noise removal effect of post-processing for increasing t and varying p and p_f for the categorical ice-wedge TI .

The post-processing step proved to be valuable especially for intermediate t values (0.1 and 0.2), since the noise can be considered as entirely removed without substantially increasing CPU time. The simulations obtained with intermediate t after post-processing were similar to these obtained with small t, except for the boundaries which were less sharp. Furthermore, post-processing allowed for a significant reduction in CPU time. With $t = 0.1$ and one post-processing step, we obtained in 58 s realizations similar to when using $t = 0.05$ without post-processing, which took 163 s.

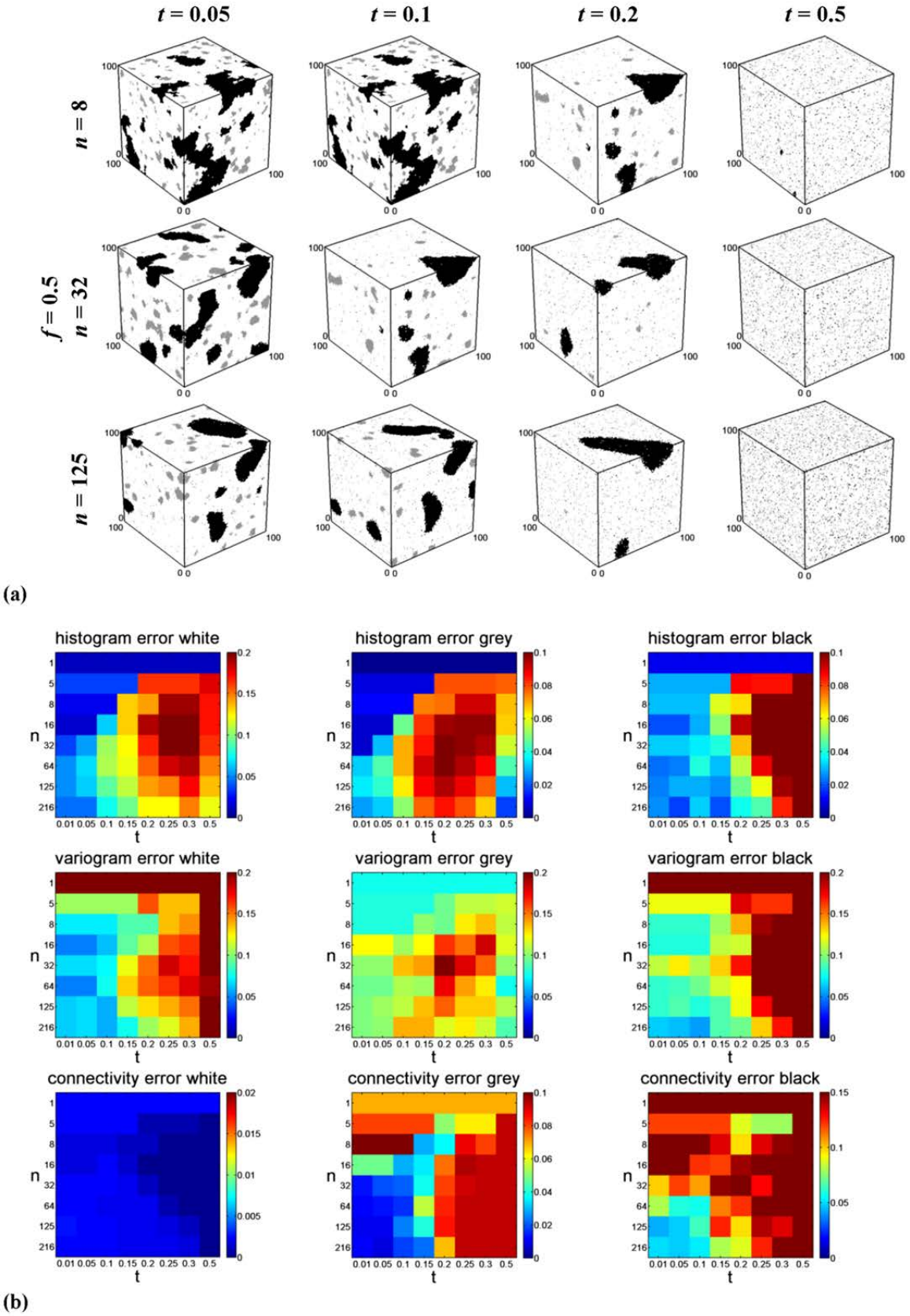


Figure 5.11 3D example with (a) first out of 10 unconditional simulations illustrating the effect of parameters t and n with $f = 0.5$ and (b) influence of t and n (for $f = 0.5$) on the quality indicators based on the 3D concrete TI (Figure 5.3g).

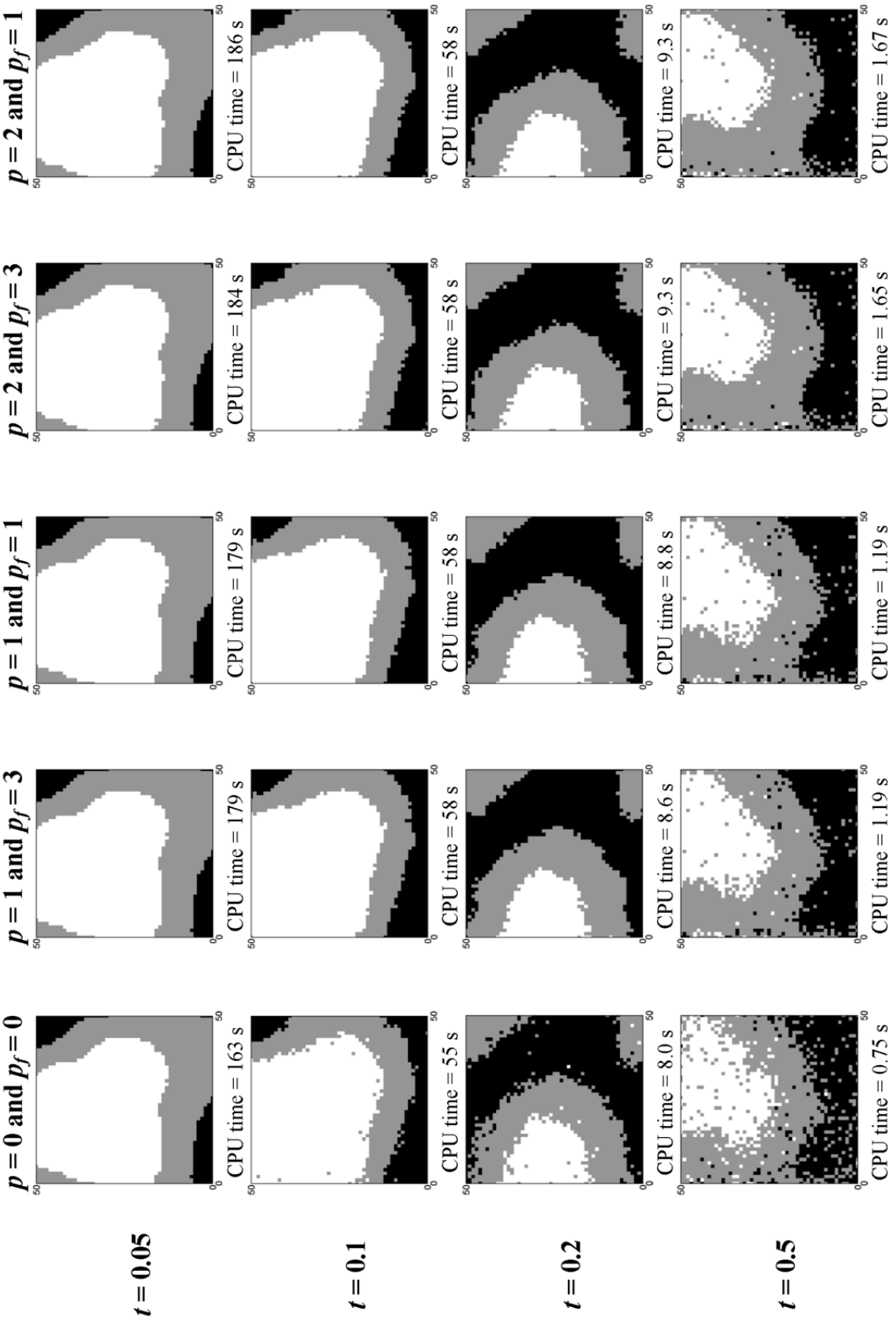


Figure 5.12 Illustration of the noise removal effect of post-processing using the categorical ice-wedge TI (Figure 5.3b) for increasing *t*, and sensitivity analysis for the number of post-processing steps (p) and the post-processing factor (p_f) showing the lower left corner of the simulation grid.

For small t (0.05) the post-processing step was not necessary because the simulation quality was already good without it. For large t (0.5) it was insufficient since post-processing only removes noise and does not improve structures at larger scale. Repeating the post-processing step did not result in significant quality improvements, and whether or not f and n were decreased in the post-processing step neither decreased the CPU time nor improved the simulation quality.

The effect of a post-processing step was less substantial for the continuous case than for the categorical case and the CPU cost was much higher. Hence, for continuous cases, the quality loss of selecting a large t, cannot be corrected with one or more additional post-processing steps.

Since p and p_f have to be chosen in advance, it can be considered as good practice to add one additional post-processing step when simulating categorical variables. When noise appears, it will be reduced and the extra CPU time needed is relatively low. For p_f a value of 1 can be selected, since adjusting p_f does not seem to have an effect. If the simulations still contain noise after post-processing, it is however advised to decrease t instead of adapting p and p_f.

5.3.4 CASE 4: Multivariate simulation

Among the MPG methods, only DS has demonstrated its potential to simulate m variables simultaneously based on m TIs. These variables can be continuous, categorical or a mixture of both since for each a different distance measure can be chosen (distance type parameter set to 0 for categorical and 1 for continuous). Several implementations have been tested. The one of Mariethoz et al. (2010) is used in this chapter. First, a path is defined that goes randomly through all the non-observed grid nodes $\mathbf{x}$ for each of the m variables. When one variable is simulated at one location, the other variable at the same location can be simulated later in the path. For each $\mathbf{x}$ a multivariate data event $\mathbf{d}_n(\mathbf{x})$ is defined that contains the neighbouring data for the m variables, which do not have to be collocated. For each variable the maximum number of neighbours n_m can be defined separately. Based on a weighted average of the m selected distance measures, the multivariate TI pattern, centred at the same node for each TI, is chosen that is most similar to the multivariate $\mathbf{d}_n(\mathbf{x})$ and the value at the central node of this TI pattern is pasted in the simulation grid at location $\mathbf{x}$. The weights used to define the multivariate distance measure w_m are user-defined. DS automatically normalizes their sum to one. If conditioning data are given for all or some of the m variables, they will be honoured by assigning them to the closest grid node prior to sequential simulation, as shown in CASE 5 (Mariethoz et al., 2010).

A potential application is a situation where one variable is (partially) known and the other(s) are to be simulated (the collocated simulation paradigm). DS becomes especially

interesting when the relationship between the variables is known via the training data set but not expressed as a simple mathematical relation. Applications can be found in Mariethoz et al. (2010; 2012) and Chapter 7 of this dissertation. As an illustration we showe five unconditional bivariate simulations using the categorical and continuous ice-wedge TI (Figure 5.3a and b) as bivariate TI and performed a sensitivity analysis on the weights given to both variables (Figure 5.13). For the other parameters we used the default values as given in Table 5.1: both n_{cat} and n_{cont} were 50.

Figure 5.13 shows that not only the spatial texture of each TI is reproduced, but also the multiple-point dependence between the TIs. The weights given to each variable strongly influenced the continuous variable. The larger w_{cont} , the better the quality of the continuous variable. The quality of the continuous variable decreased for smaller w_{cont} . In such cases, the bivariate relationship between both TIs was well respected, but the spatial continuity of the continuous variables was not strongly imposed. The quality of the categorical simulations was less affected by the choice of the weights. Note that for large w_{cont} the continuous simulation was almost an exact copy of the continuous TI (Figure 5.3a), which is again an example of the patching effect described in CASE 1.

These results suggest that it is often beneficial for the quality of the simulation of continuous variables to co-simulate a categorical variable that helps reproducing the continuity of the structures. This is a generally accepted technique in image processing in which the categorical variable is called ‘feature map’ (Lefebvre and Hoppe, 2006; Zhang et al., 2003).

5.3.5 CASE 5: Data conditioning

DS always honours conditioning data by assigning them to the closest grid node prior to simulation. Hence, local accuracy is guaranteed (the pixels at the data locations will have the correct values) but the simulated structures need to be coherent with the conditioning data. Therefore, one needs to check whether the fixed grid nodes are embedded in the spatial pattern or whether they appear as noise. The parameter that can be used to enforce pattern consistency in the neighbourhood of the conditioning data is the data conditioning weight δ. This parameter is used in the distance computation to weight differently data event nodes that correspond to conditioning data. If δ = 1, all the nodes in $\mathbf{d}_n(\mathbf{x})$ have the same importance. For δ > 1 higher weights are given to the data event nodes that are conditioning data, while for δ < 1 they are given lower weights (Mariethoz et al., 2010; Zhang et al., 2006a).

For both the categorical and continuous cases, one of the best unconditional simulations (t = 0.01, f = 1, n = 80) was used as reference image. To avoid using reference images that were copies of part of the TI due to patching, we first mirrored both

simulations horizontally and vertically, before sampling 100 conditioning data from each according to a stratified random sampling scheme (Figure 5.14 and Figure 5.15).

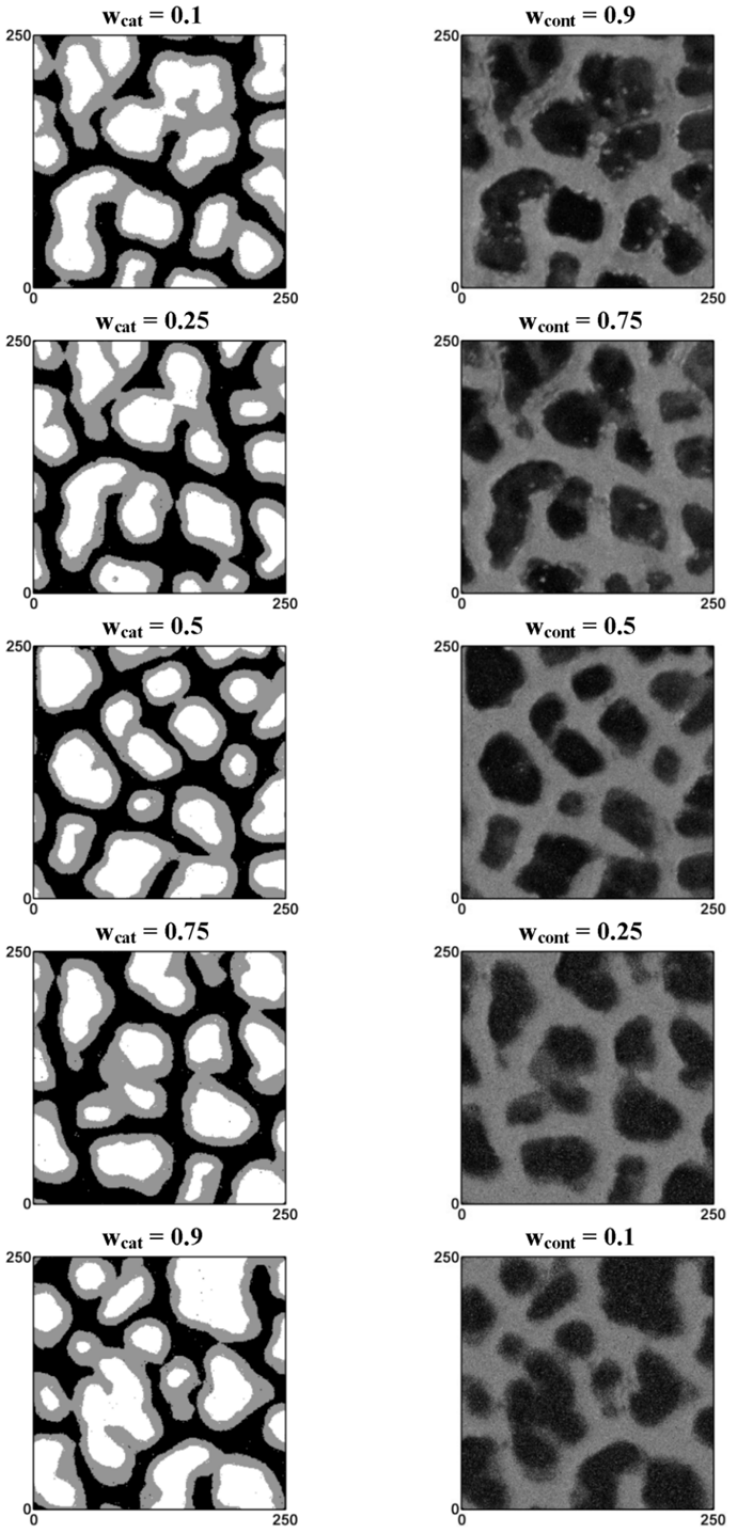


Figure 5.13 Illustration of the multivariate simulation option: five unconditional bivariate simulations using Figure 5.3a and b as bivariate TI, and sensitivity analysis for the weights given to the two variables (w_{cat} and w_{cont}). The left column represents the categorical variable for each simulation, and the right column represents the corresponding continuous variable.

Using the default parameters for t, f and n (Table 5.1), we run 50 simulations using these conditioning data and the corresponding TI. To remove noise, one post-processing step was performed with $p_f = 1$ (CASE 3). Conditioning data nodes were not re-simulated during post-processing. For $\delta = 0$, $\delta = 1$ and $\delta = 5$, the first conditional simulation was shown together with the conditional probabilities for category k in the categorical case (Figure 5.14), and the median over the 50 simulations for the continuous case (Figure 5.15).

It can be concluded that δ is an important parameter when conditioning data are available. For $\delta = 0$ the 50 simulations can be considered as unconditional simulations, since the conditioning data grid nodes were ignored in $\mathbf{d}_n(\mathbf{x})$. The simulation patterns were not at all consistent with the conditioning data and the large variation between the simulations resulted in non-informative summarizing images. For $\delta = 1$ the simulations showed patterns that were more or less consistent with the conditioning data. The remaining inconsistencies disappeared with $\delta = 5$, resulting in summarizing images that closely resembled the reference images. The better results for $\delta = 5$ were due to the high quality of the conditioning data, which perfectly represented the reference image without measurement errors. Generally, we advise to set δ to a value larger than or equal to 1. The smaller the expected uncertainty of the conditioning data, the larger δ can be chosen.

Note that for this example the conditioning data were sampled from a field with a spatial pattern that was very similar to the TI. When one expects that the underlying spatial pattern of the conditioning data deviates more from the TI, the use of transform-invariant distances can be beneficial. This option of DS increases the number of TI patterns by randomly scaling or rotating the patterns found in the TI (Mariethoz and Kelly, 2011).

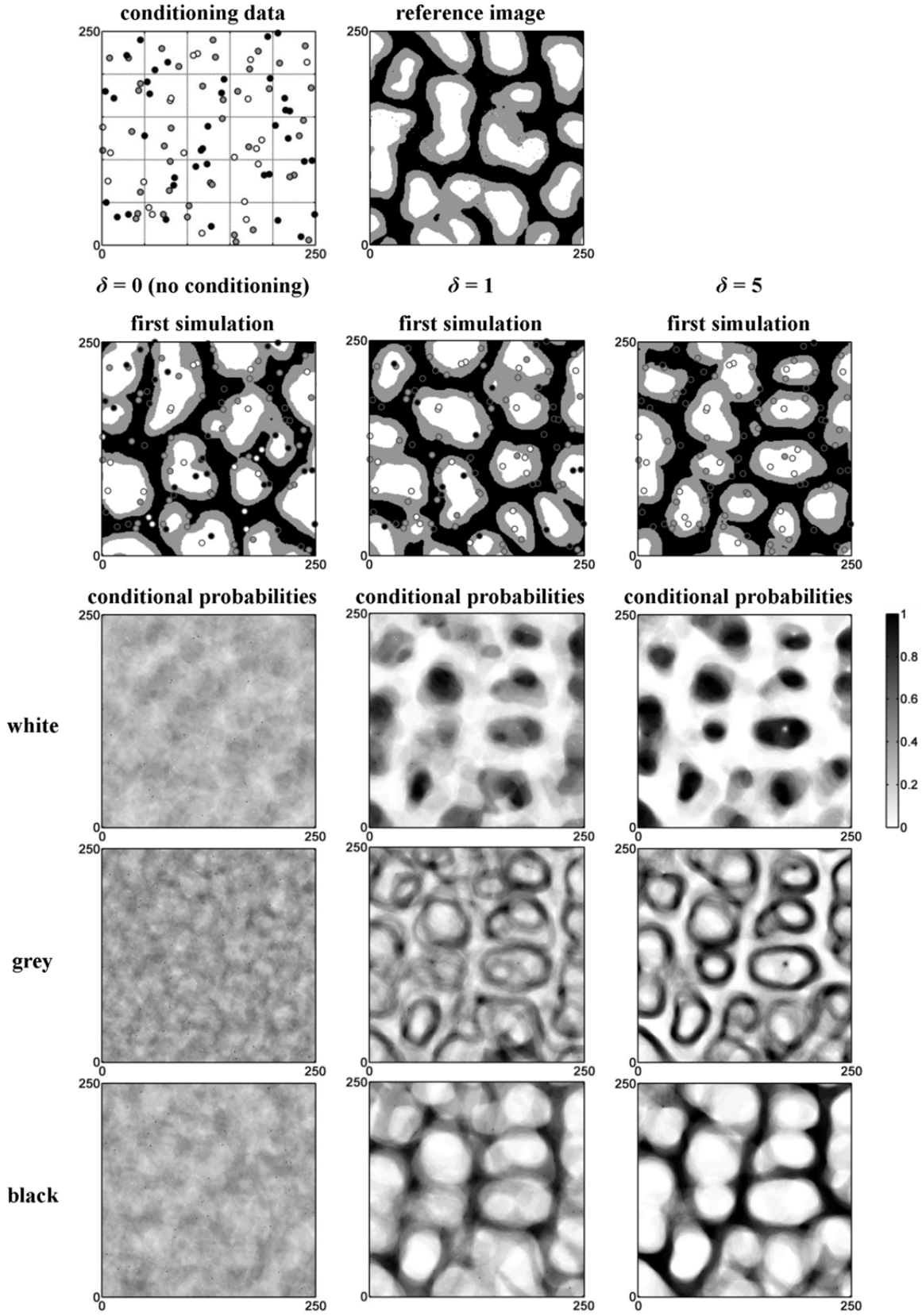


Figure 5.14 Illustration of data conditioning for the categorical ice-wedge TI (Figure 5.3b) based on 100 conditioning data. For δ = 0, δ = 1 and δ = 5 the first simulation is shown together with the conditional probabilities for each category summarizing 50 simulations.

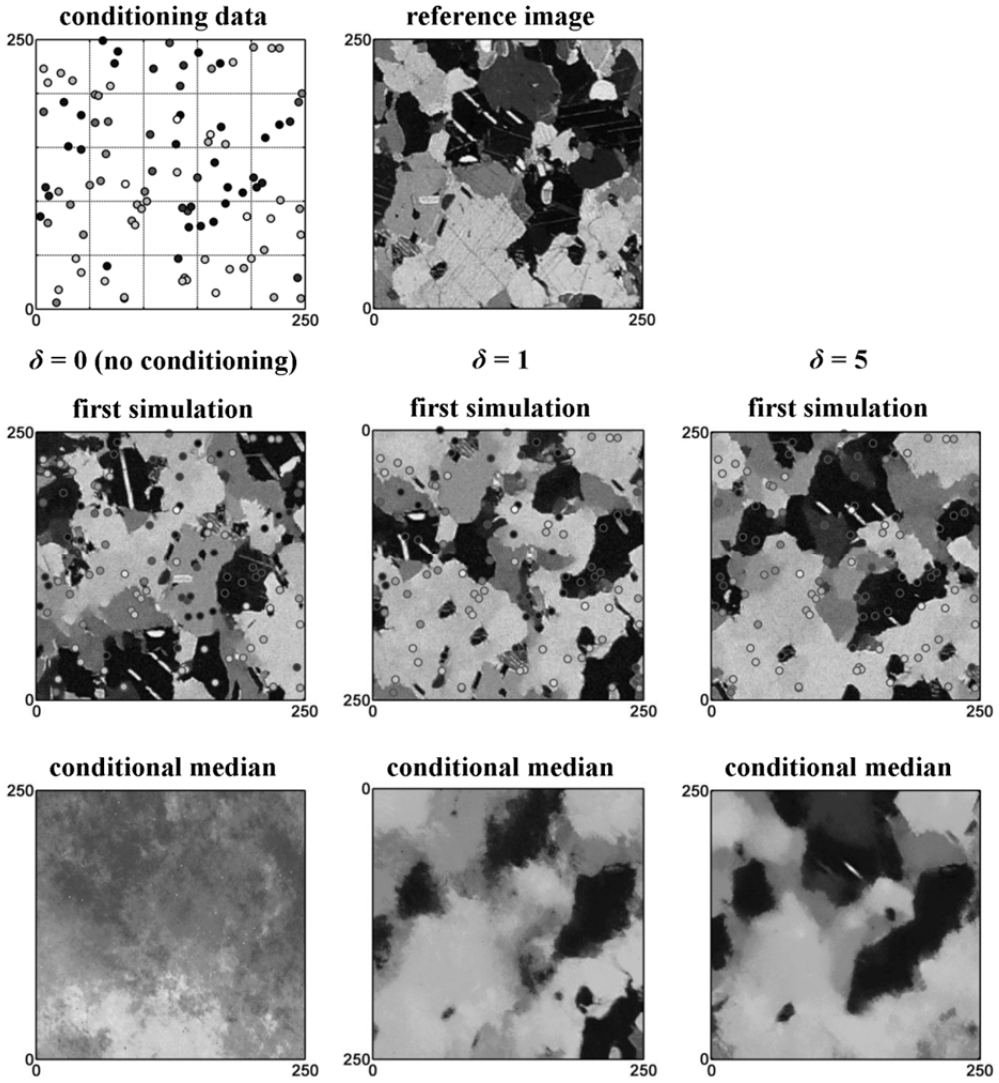


Figure 5.15 Illustration of data conditioning for the continuous marble TI (Figure 5.3c) based on 100 conditioning data. For δ = 0, δ = 1 and δ = 5 the first simulation is shown together with the conditional median for each category summarizing 50 simulations.

5.4 Conclusions

This chapter reported the first comprehensive sensitivity analysis for the DS algorithm, aiming to encourage users to benefit more efficiently from the potential of DS and its wide spectrum of applications. Given these results we provide the following general guidelines. For categorical TIs, choosing $t \leq 0.2$ and $n \geq 30$ will generally result in high quality simulations. Smaller t and larger n result in better simulation quality and a lower level of noise. However, this choice will also depend on the available CPU time. Furthermore, for small t and large n the user should check if there is still sufficient variability between the simulations. For continuous TIs, we advise to select $t \leq 0.1$ and $n \geq 30$. For continuous

cases, the selection of t and n is a delicate balance between ensuring good simulation quality and still guaranteeing sufficient variability between the simulations (avoiding patching). A good strategy to reduce both CPU time and the risk of patching is setting $f < 1$, and reducing the maximum search distance to a third the domain size or less, thus scanning a different fraction of the TI for each unknown grid node.

For categorical simulations in particular, it is advised to always add one post-processing step for noise removal. If the final simulations still contain (too much) noise, improvements should be sought by adapting t and n.

Simulating bivariate images is a very new and promising technique first offered by the DS algorithm. With the illustrative example in this chapter we have shown that the weights given to each variable clearly affect simulation quality. In case of continuous variable simulation, it is beneficial to add an auxiliary categorical variable that is co-simulated with a relative small weight. This generally improves the simulation of the continuous variable.

When conditioning data are available, it is interesting to put the weights given to the conditioning data (parameter δ) higher than the weights given to the already simulated nodes. This results in simulated patterns more consistent with the conditioning data.

Chapter 6

Categorical and continuous MPG reconstruction of the polygonal network

The content of this chapter is based on: Meerschman, E., Van Meirvenne, M., Van De Vijver, E., De Smedt, P., Islam, M.M. and Saey, T. 2013. Mapping complex soil patterns with multiple-point geostatistics. *European Journal of Soil Science* 64, 183–191.

After collecting an appropriate test data set (chapter 3) and studying the DS algorithm (chapter 5), in this chapter we applied a first MPG reconstruction of the polygonal network test data set.

6.1 Introduction

Most MPG applications can be found in petroleum and hydrogeological studies (Comunian et al., 2011; Huysmans and Dassargues, 2009; Le Coz et al., 2011; Ronayne et al., 2008; Strebelle et al., 2003; Zhang et al., 2006a). However, complex spatial patterns that are hard to model with traditional TPG also occur in soil science (chapter 1). This chapter provides a case study to demonstrate the applicability of MPG in soil science using the polygonal ice-wedge data (chapter 3).

The coordinates of the ΔECa map of the polygonal network of ice-wedge casts (Figure 3.4a) were first rotated and then translated to have their origin in the lower left corner. The lower left part of this map was used as the exhaustively known reference image (30 x 30-m). From this reference image we extracted a continuous and a categorical data set and evaluated the continuous and categorical MPG reconstruction of the polygonal network.

6.2 Continuous MPG reconstruction

6.2.1 Continuous reference image and conditioning data

The lower left part of the ΔECa image (30 x 30-m) (Figure 6.1) was used as the continuous reference image (Figure 6.2 – top right). Ten measurement lines within this area, having an inter-line distance of 3 m and a within-line distance of 0.4 m, were used as conditioning input data (655 data points) (Figure 6.2 – top left).

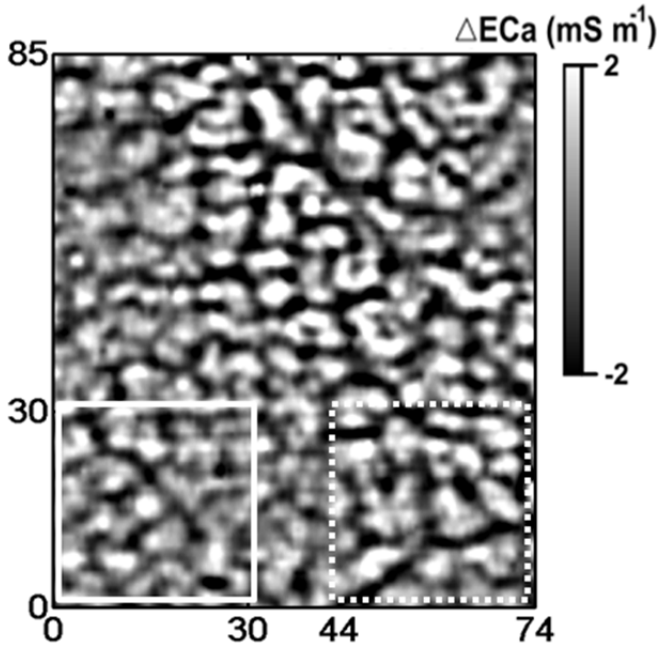


Figure 6.1 ΔECa image (mS m^{-1}) (Figure 3.4a) after coordinate transformation: the left bottom part of this image was used as continuous reference image (white rectangle) and the right bottom part as a continuous TI (white dashed rectangle).

6.2.2 Variogram modelling and mapping with traditional two-point geostatistics

Before reconstructing the image with MPG, let us first apply a standard TPG approach of fitting a variogram model to the experimental variogram and using this model to generate a prediction and simulation map (see section 2.2). The experimental variogram of the continuous data is given in Figure 6.2 (centre left). It shows no nugget effect and an almost linear increase to a sill which displays a hole effect, indicating a fairly regular repetition in the process (Webster and Oliver, 2007). To evaluate the contribution of considering the hole effect, we fitted both a non-periodic and a periodic variogram model. The non-periodic variogram model was a cubic function:

$$\gamma(h) = \begin{cases} C\left[7\left(\frac{h}{a}\right)^2 - 8.75\left(\frac{h}{a}\right)^3 + 3.5\left(\frac{h}{a}\right)^5 - 0.75\left(\frac{h}{a}\right)^7\right] & \text{for } h \le a \\ C & \text{for } h > a \end{cases} \qquad (6\text{-}1)$$

with C = 1.11 m and a = 4.1 m. Similar to a Gaussian function, a cubic function is a bounded model with reverse curvature near the origin (Webster and Oliver, 2007). We incorporated periodicity in the variogram model by fitting a combination of a cubic and cardinal sine variogram:

$$\gamma(h) = \begin{cases} C_1\left[7\left(\frac{h}{a_1}\right)^2 - 8.75\left(\frac{h}{a_1}\right)^3 + 3.5\left(\frac{h}{a_1}\right)^5 - 0.75\left(\frac{h}{a_1}\right)^7\right] + C_2\left[1 - \frac{\sin\left(\frac{h}{a_2}\right)}{\left(\frac{h}{a_2}\right)}\right] & \text{for} \quad h \le a_1 \\ C_1 + C_2\left[1 - \frac{\sin\left(\frac{h}{a_2}\right)}{\left(\frac{h}{a_2}\right)}\right] & \text{for} \quad h > a_1 \end{cases} \qquad (6\text{-}2)$$

with C_1 = 0.48, a_1 = 3.2 m, C_2 = 0.52 and a_2 = 0.9 m. The cardinal sine function is a simple periodic function that is valid in one, two and three dimensions (Webster and Oliver, 2007).

The prediction maps were created with ordinary kriging (OK) and the simulation maps with sequential Gaussian simulation (SGS) (Goovaerts, 1997) (Figure 6.2 – centre right). The ΔECa data distribution already followed a standard normal distribution, making a normal score transformation unnecessary. Both the variogram modelling and the two-point geostatistical mapping were performed with the Isatis software (Bleinès et al., 2011). The search area was set to the size of the study area (30 x 30-m) and the maximum number of neighbours to 50.

6.2.3 TI construction and mapping with multiple-point geostatistics

Whereas 655 observations are generally considered as sufficient to infer two-point statistics, this number is not enough to infer multiple-point statistics. Therefore, we applied an alternative strategy and constructed two continuous TIs. The first TI (Figure 6.2 – bottom left) was built by a histogram transformation of the unconditional simulation generated from the PCLT model (Figure 4.8b). We used the TRANS algorithm that is implemented in SGeMS that transforms a variable Z with a source cdf F_z into a variable Y with a target cdf F_Y: $Y = F_Y^{-1}(F_z(Z))$ (Remy et al., 2009, p. 216). We defined F_Y as the experimental cdf of the 655 observations. To get an idea about how sensitive the MPG maps are to the TI, we repeated the reconstruction with a second continuous TI, for which we used another part of the ΔECa image (Figure 6.1). This mimics a situation where a field is partially sampled at a very high resolution to infer multiple-point statistics and partially at a lower resolution (larger inter-line distance) for cost minimisation.

The prediction maps were obtained as the E-type (conditional mean) of 100 DS realizations and the simulation maps were obtained as the first DS realizations (Figure 6.2

– bottom right). We used the default distance type for continuous data. Parameter *t* was set to 0.02, *f* to 0.75, *n* to 50 and the maximum search area equal to the size of the study area (30 x 30-m).

6.2.4 Evaluation of the two-point and multiple-point maps

Both the TPG and MPG prediction maps corresponded reasonably well to the reference image (Figure 6.2). The mean absolute estimation error (MAEE) was 0.51 for the TPG map based on the cubic variogram, 0.52 for the TPG map based on the periodic variogram, 0.58 for the MPG map based on the PCLT model TI and 0.64 for the MPG map based on the densely sampled neighbourhood TI. However, pattern reconstruction was better for the MPG maps: the connectivity of the small values was better reproduced.

The TPG prediction maps were very similar ($r = 0.99$) and thus rather insensitive to the hole effect of the variogram model, whereas the two MPG prediction maps differed more ($r = 0.77$). This demonstrates that changing the TI has larger consequences than changing the variogram: a TI has greater control over the spatial structure (Boisvert et al., 2007). To correctly interpret this, one should revert to the fundamental distinction between two-point and multiple-point techniques. For MPG simulations, the user provides a prior multiple-point structural model, i.e. the TI. This model allows one to link the *n* neighbouring data jointly to $z^*(\mathbf{x})$. For TPG simulations, the user only provides a prior one-point, i.e. the histogram, and two-point structural model, i.e. the variogram, linking the *n* neighbouring data pairwise to $z^*(\mathbf{x})$. Beyond the histogram and variogram, TPG algorithms use their own intrinsic prior structural model that is beyond the control of the user. For SGS this model is multi-Gaussian: a model that imposes maximum entropy for the high-order statistics (Journel and Zhang, 2006) (see section 2.3.1).

The different multiple-point structural models are to some extent visualised in the simulation maps (Figure 6.2). The TPG simulation maps show a spatial distribution of higher entropy: the extremes are more spatially fragmented. The spatial patterns in the MPG simulations maps can be considered as better structured, and correspond more closely to these in the reference image. This is of course due to the lower entropy (spatially connected small values) prior multiple-point models we defined by means of the TIs. Note that the conditioning data strongly guide the pattern reconstruction. The differences between the TPG and MPG simulations maps would have been more profound for a smaller number of conditioning data.

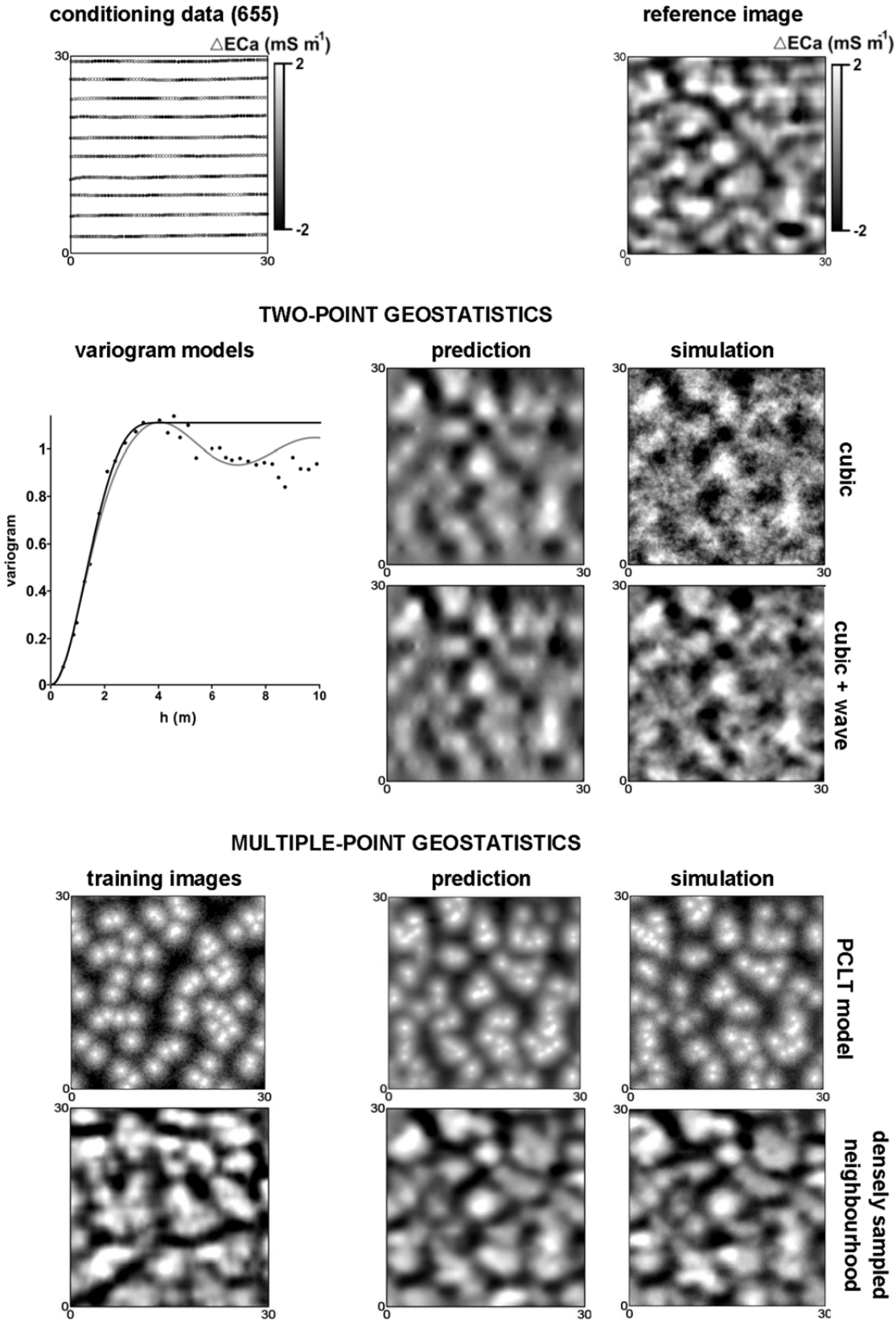


Figure 6.2 Reconstruction of the continuous reference image with two-point and multiple-point geostatistics starting from 655 conditioning data. The two-point prediction (OK) and simulation (first SGS realization) maps were based on two different variogram models. The two multiple-point prediction (E-type of 100 DS realizations) and simulation (first DS realization) maps were based on two different TIs.

6.3 Categorical MPG reconstruction

6.3.1 Categorical reference image and conditioning data

To obtain a categorical reference image, the continuous one was classified by a *k*-means classification after running a contrast enhancement filter (Figure 6.3 – top right). We chose $k = 3$ to have a strict categorical data set, and not a binary one. The three classes represent wedge material, host material and intermediate material. From this classified reference map 100 data points were extracted according to a stratified random sampling scheme, mimicking a soil sampling campaign where the subsoil textural class is observed by hand feeling (Figure 6.3 – top left).

6.3.2 TI construction and prediction with multiple-point geostatistics

The experimental set-up of the categorical case study requires an alternative approach to construct the TI. Assume that the only information we have about the spatial structure comes from the small excavated area and that there is no exhaustively measured neighbouring field to directly derive a TI from, as was done in the continuous case. Hence, we chose an existing photograph from literature and rescaled it based on the information we gathered from the excavation, because it is beneficial when the size of the TI patterns corresponds more or less with the true pattern size.

We selected a near-infrared aerial photograph of a present-day ice-wedge network in Alaska (Plug and Werner, 2002), assuming that a similar genetical process was at the basis of both ice-wedge patterns. The photograph was rescaled, equalling its average polygon size to this of the textural polygon that was observed in the Belgian field by excavation (Figure 6.4a). Then, we applied an adaptive low-pass filter for noise removal and classified the image into three classes with a *k*-means classification using Matlab (Mathworks, R2011a) (Figure 6.3b). We chose a photograph from Alaska, and not from Belgium, to better illustrate the strong concept of the TI as a database of representative patterns, independent of its origin.

The prediction map was the most probable category (conditional mode) of 100 DS realizations and the simulation map was the first DS realization. We used the default distance type and set t to 0.05, f to 0.5 and n_{max} to 50. The maximum search area was set equal to the size of the study area and after each simulation one post-processing step was performed for noise removal.

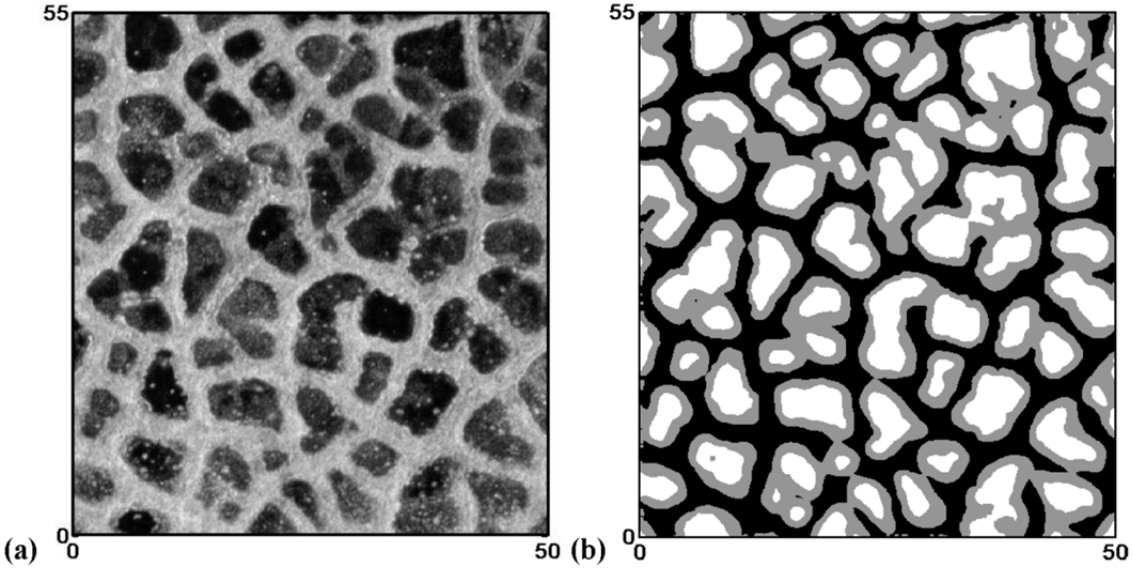


Figure 6.3 (a) The original, rescaled near-infrared aerial photograph of an ice-wedge network on the floor of a drained lake near Espenberg, northwest Alaska (Plug and Werner, 2002) used to construct (b) the categorical TI.

6.3.3 Evaluation of the multiple-point maps

Figure 6.4 compares the multiple-point categorical reconstruction with the categorical reference image. Most of the polygons were correctly identified. The categorical prediction map had a correct classification rate of 54.0 %. The largest source of error between the prediction map and the reference image was due to the difference in spatial pattern between the chosen TI and the reference image. The TI contains polygons with smoother boundaries, and slightly overestimates the connectivity of the wedge material and underestimates the connectivity of the intermediate material.

Note that we did not made the comparison with TPG here, because the number of observations was too small to reveal the spatial pattern, and the patterns themselves – especially the pattern of the intermediate material- were too complex to be modelled with a two-point variogram function.

6.4 Conclusions

We successfully applied MPG to reconstruct complex soil patterns. Soil scientists frequently face periodic, connected or curvilinear patterns. We believe that MPG is a promising and accessible technique to model these complex soil patterns and that it should be added to the pedometrician's toolbox.

Both the strength and the bottleneck of the approach is the construction of an appropriate TI. A TI allows one to reconstruct complex soil patterns that cannot be modelled with a variogram, but should be chosen carefully because it strongly influences the MPG simulations.

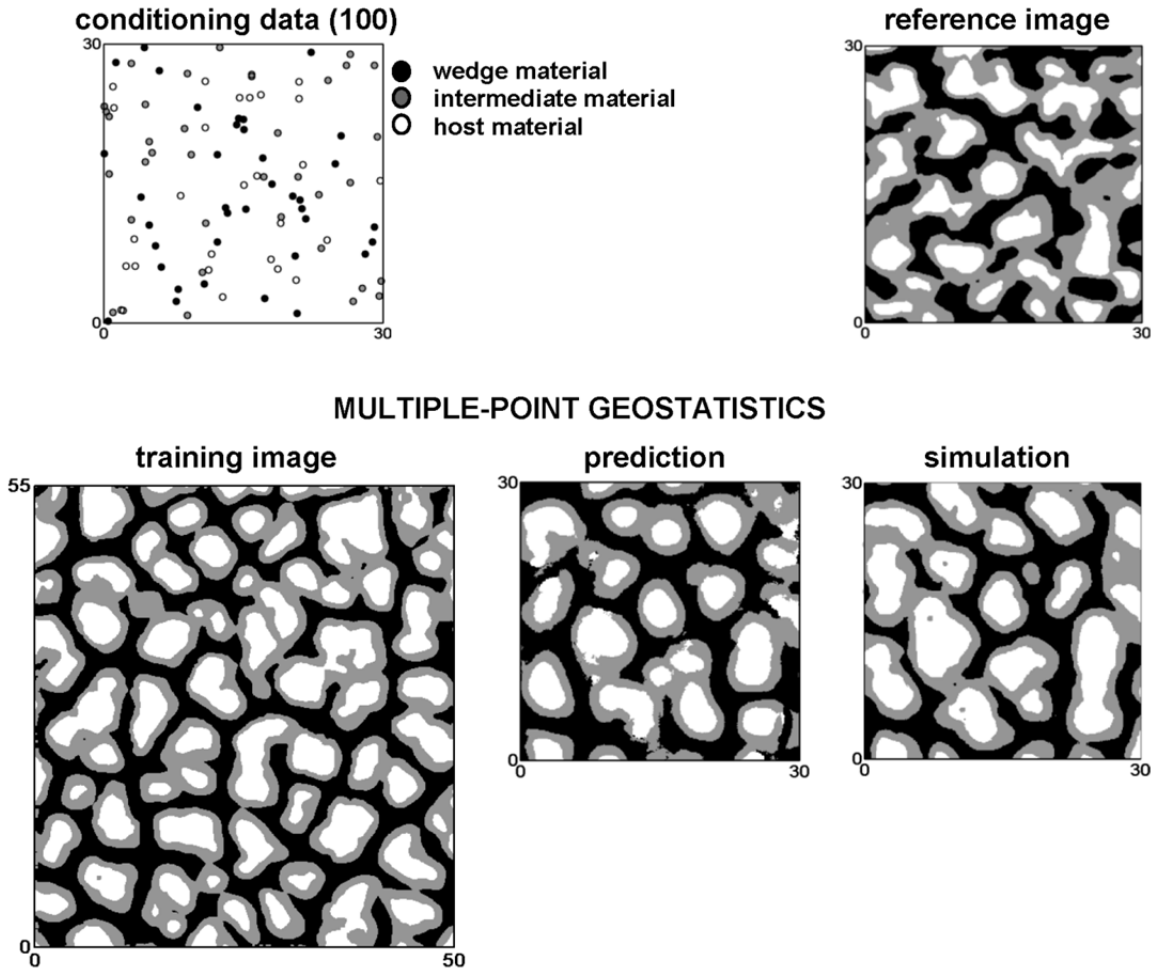


Figure 6.4 Reconstruction of the categorical reference image with multiple-point geostatistics starting from 100 conditioning data. The multiple-point prediction (conditional mode of 100 DS realizations) and simulation (first DS realization) maps were based on a categorical TI.

We applied three different strategies to construct a TI. In contrast to variogram modelling that is mostly data-driven, TI construction usually requires extra information. This extra information can consist of a geometric model that was fit to a particular situation (such as the PCLT model) or a photograph that is deemed to be representative for the studied phenomenon, as was used in the categorical case. Obviously, this selection strongly depends on the user's judgement and requires sufficient knowledge of the studied phenomenon. Creating a TI from neighbouring proximal soil sensor data and use it to interpolate less densely sampled areas, as illustrated in the continuous case, is an elegant approach, especially with today's growing availability of high density measurements.

Chapter 7

MPG reconstruction in inaccessible areas using neighbouring densely sampled areas as training data

A grid interpolated from densely sampled measurements can serve as an appropriate and easy-to-build TI (chapter 6). This chapter expands on this principle and uses it to reconstruct proximal soil sensor values in inaccessible areas showing up as gaps in a proximal soil sensor image. The neighbouring densely sampled areas are then used as both conditioning data and as TI, and are called ‘training data’. This is the first of two chapters that analyses the potential of MPG to improve the processing of proximal soil sensor data.

7.1 Introduction

Proximal soil sensing is an increasingly used data source for soil inventory (McBratney et al., 2000). Using proximal soil sensors in a mobile setup allows one to rapidly collect indirect observations of the subsoil in a non-destructive way (Adamchuk et al., 2004), as was demonstrated in section 3.3.3. It is typically done with a sensor attached to a vehicle taking measurements at fixed intervals while driving along parallel lines (Figure 7.1a). With the instruments available today, the within-line distance is generally small. The inter-line distance, on the other hand, largely affects the costs of a field survey. A good sampling strategy requires the inter-line distance to be chosen based on the expected scale of the soil features to be mapped, known as the Nyquist sampling theorem (Nyquist, 1928).

Even though proximal soil sensor data are considered as high resolution data, interpolating the data to a regular grid remains a crucial processing step. The data need to

be interpolated between the measurement lines, but possibly also in areas that were inaccessible to the mobile proximal soil sensor (Figure 7.1). Examples of the latter are areas with a dense (wooden) vegetation, stony areas, building areas or field boundaries. Nearby power lines or metal fences can also disturb the sensing system (Reynolds, 1997).

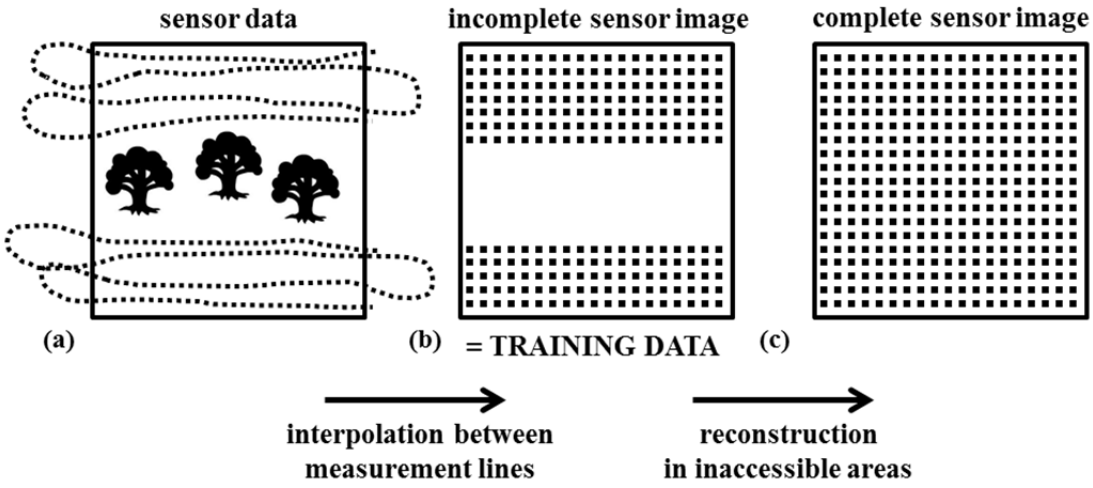


Figure 7.1 Schematic overview of the processing steps needed to interpolate proximal soil sensor data to a regular grid with (a) proximal soil sensor data with their typical sampling scheme and indication of an inaccessible area (trees); (b) an incomplete sensor image obtained by interpolating the sensor data between the measurement lines and (c) a complete sensor image obtained by reconstructing the sensor values in the inaccessible area.

Our experience has shown that OK is a successful method to interpolate the sensor values between the measurement lines (first step in Figure 7.1), on the condition that the inter-line distance is not too large. Whether a TPG approach is also successful to reconstruct the sensor values in inaccessible areas (second step in Figure 7.1) depends on the scale (represented by the range of the variogram) and the geometrical complexity of the soil features to be mapped. Assume for instance a field that has an inaccessible area of 20 x 20-m. When the sensor measurements reflect a smoothly varying texture pattern and their variogram has a range of 50 m, TPG should be able to reconstruct the sensor values in the inaccessible area. But, reconstructing sensor measurements that reflect clay lenses with an average radius of 4 m will not be possible. The size of the inaccessible area will prohibit an accurate prediction, and the geometrical complexity of the clay lenses will even prohibit a TPG simulation, such as SGS, to reconstruct the spatial patterns without local accuracy.

We suggest a MPG approach to reconstruct proximal soil sensor values in inaccessible areas (second step in Figure 7.1). We used the DS approach and built the TI from the neighbouring densely sampled areas (Mariethoz and Renard, 2010). The technique was evaluated on two different test cases: a proximal soil sensor image of a polygonal network of ice-wedge casts and a proximal soil sensor image of a buried tidal channel. We

systematically blanked zones from these proximal soil sensor images, and reconstructed the sensor values in the blanked zones with MPG.

With the introduction of the MPG approach, we target to reconstruct more geometrically complex soil patterns. However, just as for TPG, the potential of the MPG approach will also depend on the scale of soil features relative to the size of the inaccessible area (Mariethoz and Renard, 2010). Whereas the ice-wedge polygons of the first test case were small relative to the size of the blanked area, the buried tidal channel was large relative to the size of the blanked area. Therefore, we targeted good simulations of the spatial pattern (without local accuracy) in the first test case, and locally accurate prediction maps in the second case.

7.2 Material and methods

7.2.1 Test cases

The first proximal soil sensor image was the ΔECa image representing a polygonal network of ice-wedge casts (chapter 3) with transformed coordinates and a resolution of 0.25x0.25-m. We blanked an east-west oriented 30 m-wide area (Figure 7.2).

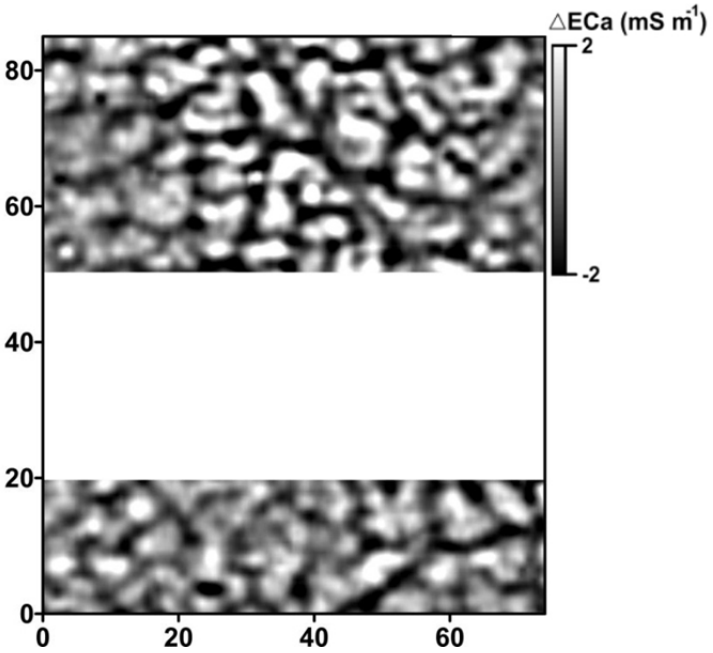


Figure 7.2 Test case 1: ΔECa image representing a polygonal network of ice-wedge casts (Figure 6.1) with a 30 m-wide blanked area (coordinates are in metres relative to the lower left corner).

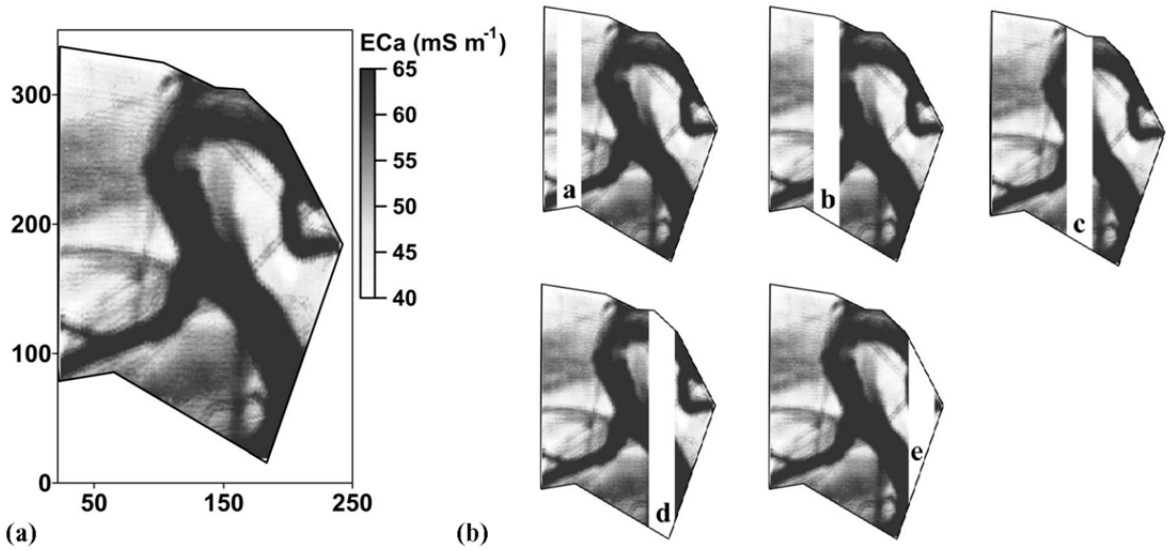


Figure 7.3 Test case 2: (a) ECa image representing a buried tidal channel (Saey et al., 2012) and (b) indication of five successively blanked 30 m-wide areas (coordinates are in metres relative to the lower left corner).

The second proximal soil sensor image was an ECa image representing a buried tidal channel (Saey et al., 2012) (Figure 7.3a). This ECa image resulted from surveying a 6.5 ha agricultural field located in the western part of the coastal plain of Belgium (central coordinates: 51°06'45"N and 2°42'04"E) with an EMI sensor (DUALEM-21S). The inter-line distance was 1 m and the within-line distance was 1.7 m, which was sufficiently small relative to the size of the buried tidal channel (Figure 7.3a). From the four ECa measurements simultaneously collected by the DUALEM-21S sensor, we used these measured with a horizontal coplanar coil orientation and a coil separation of 2 m (Simpson et al., 2009). The 39 326 data were interpolated to a regular grid with a resolution of 0.5x0.5-m using OK. More details about this test case can be found in Saey et al. (2012). We sequentially blanked five north-south oriented 30 m-wide areas from the ECa image, as indicated in Figure 7.3b.

7.2.2 DS using training data

The Direct Sampling algorithm is capable of using the conditioning data themselves as *training data*, as is described in Mariethoz and Renard (2010). This means that multiple-point statistics can be inferred directly from the conditioning data when their sampling density is sufficiently high. The conditioning data file is then the only input file needed.

Although it is theoretically possible to use the irregularly spaced proximal soil sensor data before interpolation (Figure 7.1a) as training data, our experience has shown that it was then more difficult to find replicates. Therefore, we first interpolated the proximal soil sensor data between the measurement lines with OK and used the incomplete sensor image

as training data (Figure 7.1b). For the two test case, the training data are shown in Figure 7.2 and Figure 7.3b.

The core of the DS algorithm, as explained in chapter 4, remains unchanged if a data set has to be scanned for replicates of the data event instead of a TI (Mariethoz and Renard, 2010). When the user sets the TI file parameter to 'none', DS automatically uses the conditioning data as training data, and scans these for replicates of the data event.

All simulation maps in this chapter were generated by using the default distance type for continuous variables. For the first test case, *t* was set to 0.02, *f* to 0.5 and *n* to 30, and for the second test case *t* was set to 0.05, *f* to 0.75 and *n* to 30. When using the training data, we set the weight given to the conditioning data two times larger than the weight given to the already simulated grid nodes, aiming to enforce pattern consistency at the boundaries of the blanked areas (see section 5.3.4). When extra point observations within the blanked area were added to the conditioning data (test case 2), we increased the data conditioning weight to five.

7.2.3 Evaluation

To evaluate the pattern reconstruction of the MPG simulations in test case 1, we calculated the connectivity function for both the training data and ten realizations. These were first classified in two categories, i.e. the smallest (< median) and largest (≥ median) values. For each category, the connectivity functions were calculated in the east-west direction using the MATLAB function ConnectFct.m accompanying the DS algorithm (Mariethoz, 2009).

The local accuracy of the MPG predictions maps in test case 2 was evaluated by calculating the mean absolute estimation error (MAEE), the root mean square estimation error (RMSEE) and the Pearson's correlation coefficient *r* between the predicted and the true ECa values within each blanked area.

7.3 Results and discussion

7.3.1 Test case 1

Figure 7.4a shows the first DS realization reconstructing the ΔECa values in the blanked area. Note that the difference between this MPG simulation and the one from section 6.2.3 is that the neighbouring densely sampled areas are here also used as conditioning data, resulting in the pattern consistency at the boundaries of the blanked area. It took DS 44 s (on a 3GHz Windows PC) to generate this realization. The spread of the simulated values was slightly smaller than for the training data: the standard deviation of the ΔECa values in the blanked area was 0.95, whereas it was 1.26 for the training data.

The polygonal patterns were well reconstructed which was illustrated by the similarity between the connectivity functions (see section 5.3.1) of the training data and ten MPG reconstructions (Figure 7.4b). The added value of completing the blanked ΔECa image with a MPG simulation is that the full image can be easily used as an input map for, for instance, preferential flow path modelling.

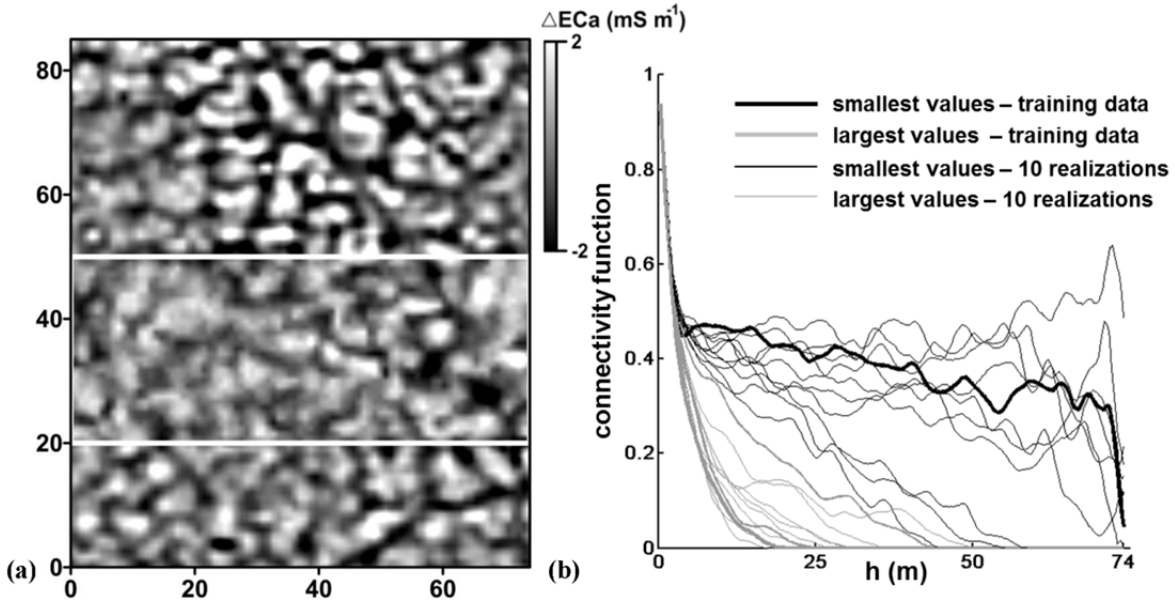


Figure 7.4 (a) A MPG simulation of ΔECa (mS m^{-1}) within the blanked area (white rectangle) using the densely sampled areas as training data and (b) connectivity functions (calculated in the east-west direction) of the smallest (< median) and largest ΔECa values (≥ median) of the training data and ten MPG realizations.

Creating a map that predicts the location of the ice-wedge casts in the blanked area is impossible because the size of the blanked area is large relative to the size of ice-wedge polygons. Different MPG realizations all reconstructed the polygonal pattern well, but each realization showed polygons at different locations, making the E-type map non-informative. To ensure local accuracy in the blanked area, one should collect additional point observations within the blanked area and add these to the conditioning data set, as illustrated for test case 2.

7.3.2 Test case 2

Figure 7.5 shows the E-type (conditional mean of 50 simulations) maps predicting ECa values within the five successively blanked areas. The E-type maps could be used as prediction maps because all realizations were more or less similar, in contrast with test case 1. Comparing the MPG prediction maps with the true ECa values in the blanked areas (Figure 7.3a) shows that the predictions were accurate. The buried tidal channel was well reconstructed. The predictions were consistent with the training data at the boundaries of the blanked area and the connectivity of the large ECa values was well preserved.

Just as for test case 1, the prediction values had a slightly smaller spread than the true values. For instance, the true values in area a had a standard deviation of 5.68 mS m^{-1}, whereas the predicted values had a standard deviation of 4.44 mS m^{-1}. Generating one E-type map took about 7.5 min, or 9 s per realization (on a 3GHz Windows PC).

The high prediction quality was confirmed by the validation results as shown in Table 7.1. The Pearson's correlation coefficient was large for all areas and ranged between 0.80 for area e and 0.95 for area a. The MAEE and RMSEE were also small relative to the magnitude of the ECa values (the mean of the reference image was 55 mS m^{-1}).

Table 7.1 Validation indices (MAEE = mean absolute estimation error, RMSEE = root mean square estimation error and *r* = Pearson's correlation coefficient) for the E-type maps (Figure 7.5) of the five blanked areas.

	area a	area b	area c	area d	area e	area e + 26 samples
MAEE	1.49	2.96	2.92	5.14	5.27	3.10
RMSEE	2.06	3.96	4.11	6.78	7.12	4.35
r	0.95	0.82	0.87	0.81	0.80	0.93

To improve the prediction quality of the area with the worst validation results, i.e. area e, we added additional samples from within this area. As has been shown for TPG simulation techniques (Meerschman et al., 2011; Van Meirvenne and Goovaerts, 2001), it is interesting to select additional samples at locations **x** where the conditional coefficient of variation (CV) is large:

$$\mathrm{CV}(\mathbf{x}) = \frac{\sqrt{\mathrm{var}(\mathbf{x})}}{z_E^*(\mathbf{x})}, \qquad (7\text{-}1)$$

where $\mathrm{var}(\mathbf{x})$ is the conditional variance and $z_E^*(\mathbf{x})$ the E-type prediction calculated from 50 realizations. We selected the 26 locations in area e with the largest $\mathrm{CV}(\mathbf{x})$ and a mutual distance of at least 5 m (Figure 7.6a), and sampled the true image (Figure 7.3a) at these locations. Figure 7.6d shows the new prediction map (E-type of 50 realizations) using the training data complemented with the 26 additional point observations.

The shape of the buried channel was better reproduced when the extra conditioning data were added (Figure 7.6d). The validation indices also improved remarkably (Table 7.1).

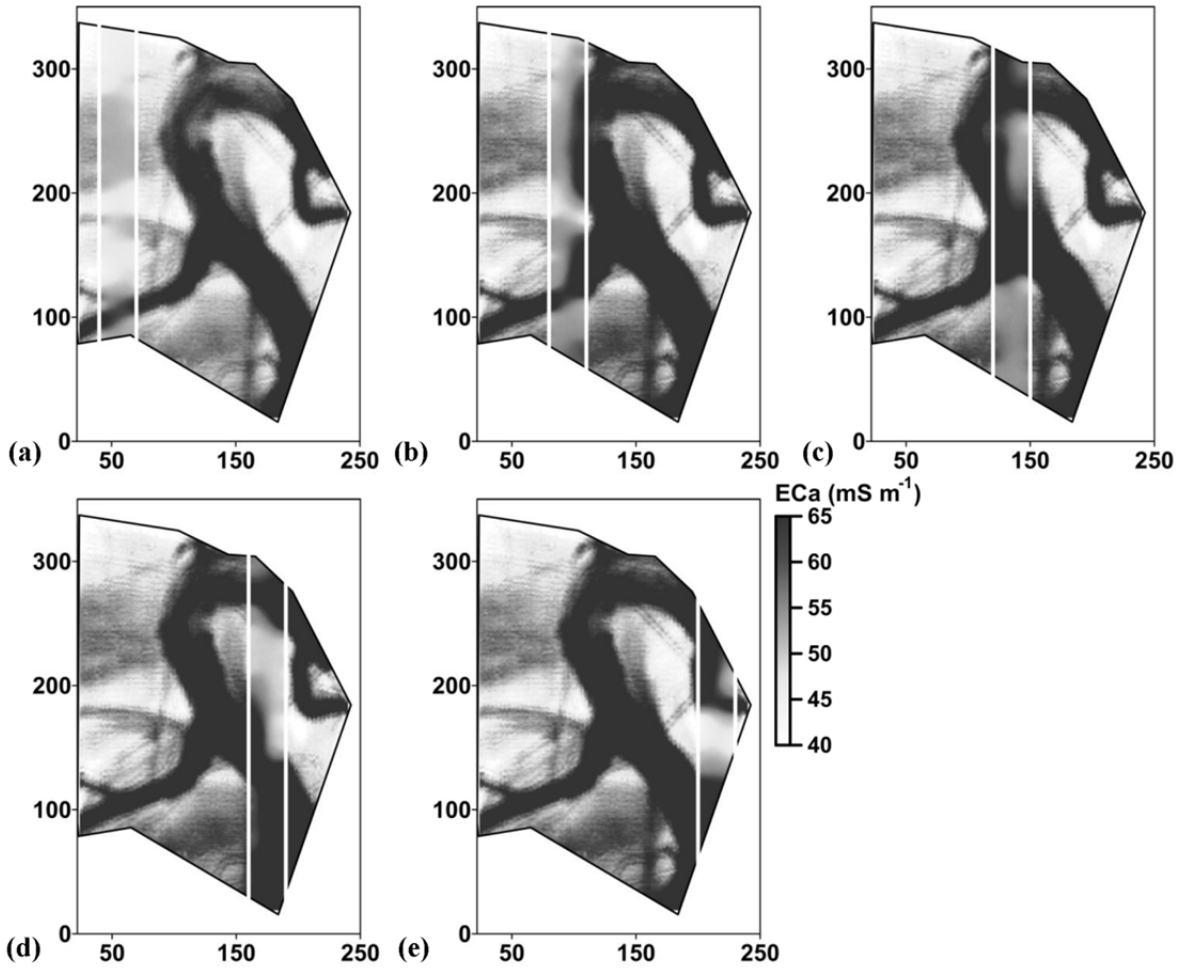


Figure 7.5 MPG maps (E-type of 50 simulations) predicting ECa within the five blanked areas (white rectangles) using the remaining densely sampled areas as training data.

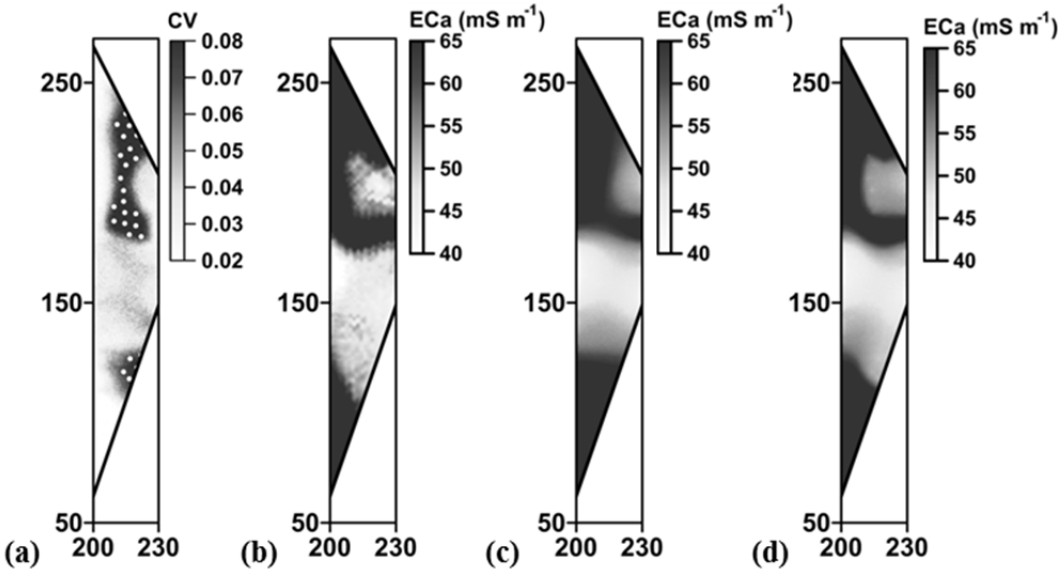


Figure 7.6 Detail of area e with (a) the conditional coefficient of variation (CV) and indication of the 26 selected locations to be sampled; (b) true ECa image (Figure 7.2b); (c) prediction map without using the 26 additional samples (Figure 7.5e) and (d) prediction map using the 26 additional samples.

The above presented methodology is very promising to be used in practice. For example, proximal soil sensor surveys are frequently interrupted along field boundaries, resulting in small elongated gaps in proximal soil sensor images. These incomplete images can be rapidly complemented by running the DS algorithm with the incomplete image as its only input file. This new approach is not only very straightforward to implement, it can also simulate geometrically complex spatial patterns that cannot be simulated with TPG.

The main condition of the approach is obviously a stationarity assumption, as for all prediction techniques. The presented approach will only work when you can assume that the multiple-point statistics in the inaccessible area are inferable from these in the densely sampled neighbourhood, or in other words, that their spatial patterns are similar.

When the prediction results are not satisfying, one can easily add additional point observations from within the 'inaccessible' area to the training data. Areas that are inaccessible to a mobile proximal soil sensor, such as areas with a dense vegetation, can often be manually surveyed. A mobile ECa survey can for instance be complemented with manual measurements with the same ECa sensor or with an EC probe. It is also possible to use point observations of a different (correlated) soil variable from within the inaccessible area, such as clay percentages. In the latter case, a bivariate MPG reconstruction is required, which is explained in the next chapter.

7.4 Conclusions

We suggested a MPG approach to reconstruct proximal soil sensor images in inaccessible areas. Multiple-point statistics were directly inferred from the neighbouring densely sampled areas, that were used both as TI and as conditioning data. Their role as TI ensured that the simulated sensor values had similar spatial characteristics as the neighbouring areas, and their role as conditioning data ensured pattern consistency along the boundaries of the inaccessible area. When the inaccessible area was small relative to the size of the soil features, the conditional mean of different MPG simulations served as an accurate prediction map. The conditional coefficient of variation was used as a guide to determine the location of extra point observations to further improve the prediction quality.

Chapter 8

Bivariate MPG to interpolate proximal soil sensor data and predict a target variable

The content of this chapter is based on: Meerschman, E., Van Meirvenne, M., Mariethoz, G., Islam, M.M., De Smedt, P., Van De Vijver, E. and Saey, T. 2013. Using bivariate multiple-point statistics and proximal soil sensor data to map fossil ice-wedge polygons. *Geoderma, in press* (DOI: 10/1016/j.geoderma.2013.01.016).

In this chapter we investigated if bivariate MPG can be used to simulate the ECa data while simultaneously predicting a categorical target variable. This chapter is the second of two chapters that answer the research question whether MPG can be used for the processing of proximal soil sensor data.

8.1 Introduction

Proximal soil sensor measurements are often considered as indirect observations that are used to predict the soil variable of interest, i.e. the target variable. Processing proximal soil sensor data then includes two steps: first the sensor data need to be interpolated to a regular grid (chapter 7) and then this map can be used as a proxy to predict the target variable (de Gruijter et al., 2010). Figure 8.1 schematically represents these steps. For this study, we assume that there are no inaccessible areas and that the data only need to be interpolated between the measurement lines.

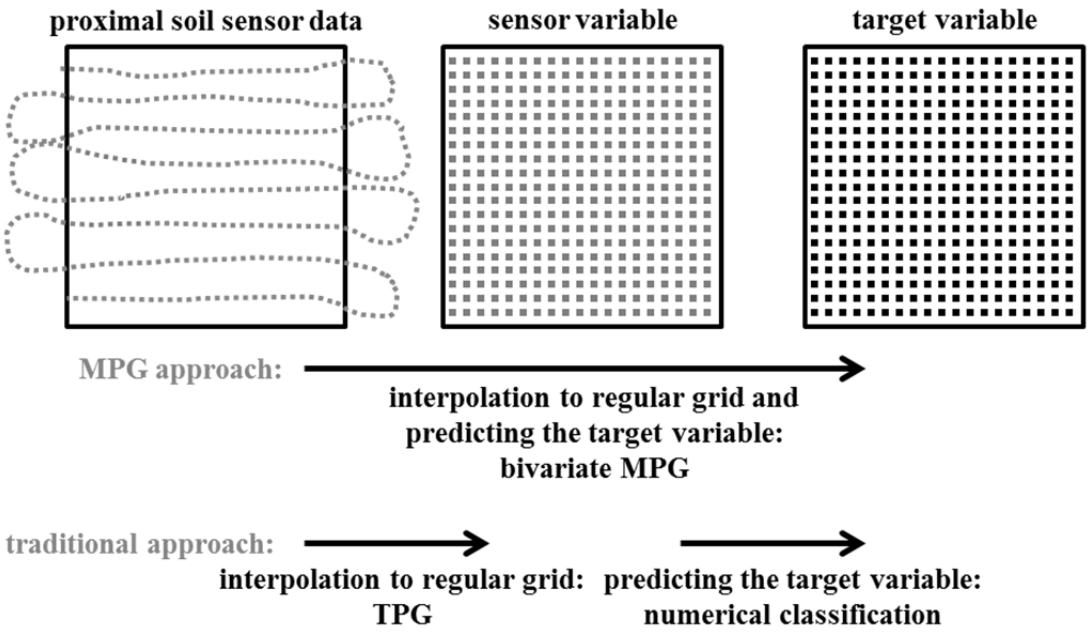


Figure 8.1 Schematic overview of the processing steps needed to predict a target variable from proximal soil sensor data

To date, ordinary kriging (OK) is an often-used method to interpolate sensor data because of its declustering ability (Goovaerts, 1997). In our experience, OK is a successful method to interpolate sensor data. However, when the sensor data reflect subsoil phenomena that have a complex spatial pattern, the two-point variogram is no longer sufficient. In practice, problems arise when the inter-line distance is large compared to the scale of the investigated soil features. Hence, it is worth investigating whether MPG can serve as a more suited interpolation technique for these situations.

If the sensor variable differs from the target variable (i.e. the soil variable of interest), a model is needed to predict the target variable from the sensor variable, which then serves as an ancillary or secondary variable (de Gruijter et al., 2010). For example, if the sensed variable is electrical resistivity and the variable of interest is porosity, the modelling of the relationship between these two attributes is critical. Depending on the specific situation and the type of target variable, a variety of pedometrical techniques can be used for this aim, ranging from numerical classification to CLORPT and hybrid techniques (McBratney et al., 2000). For instance, fuzzy *k*-means is an often-used predictive classification technique to delineate zones with homogeneous soil properties based on proximal soil sensor data (Cockx et al., 2006; 2007; Islam et al., 2011; Vitharana et al., 2008b). Examples of CLORPT techniques are predicting the depth to contrasting soil layers from proximal soil sensor data with inverse modelling techniques (Saey et al., 2008; 2009a; De Smedt et al., 2011) or predicting the soil clay content based on neural network approaches (Cockx et al., 2009). Vitharana et al. (2008a) used regression kriging to predict the depth to

clay substratum and Triantafilis et al. (2001) compared different hybrid techniques to predict soil salinity from proximal soil sensor data.

Multivariate MPG is promising for both the interpolation of sensor data and the prediction of the target variable. This technique is mainly developed for situations where one variable is (partially) known and the other is to be simulated (the collocated simulation paradigm). Using a bivariate TI is especially interesting when the relationship between the variables is known through training data but cannot simply be expressed as a mathematical relationship (Mariethoz et al., 2010; Meerschman et al., 2013). To investigate the use of multivariate MPG, we applied it to a case study aiming to predict the location of fossil ice-wedge polygons in the subsoil based on electromagnetic induction (EMI) data.

In this chapter, we applied bivariate MPG to interpolate the proximal soil sensor data to a regular grid and to simultaneously derive a map estimating the location of the fossil ice-wedge polygons in the subsoil. To set a comprehensive framework for the evaluation of the new method's prediction performance, we compared it with the often-applied procedure of interpolating the sensor data with OK and then performing a fuzzy *k*-means classification to derive the possibility of finding wedge material in the subsoil (Figure 8.1).

8.2 Material and methods

8.2.1 Study area and data collection

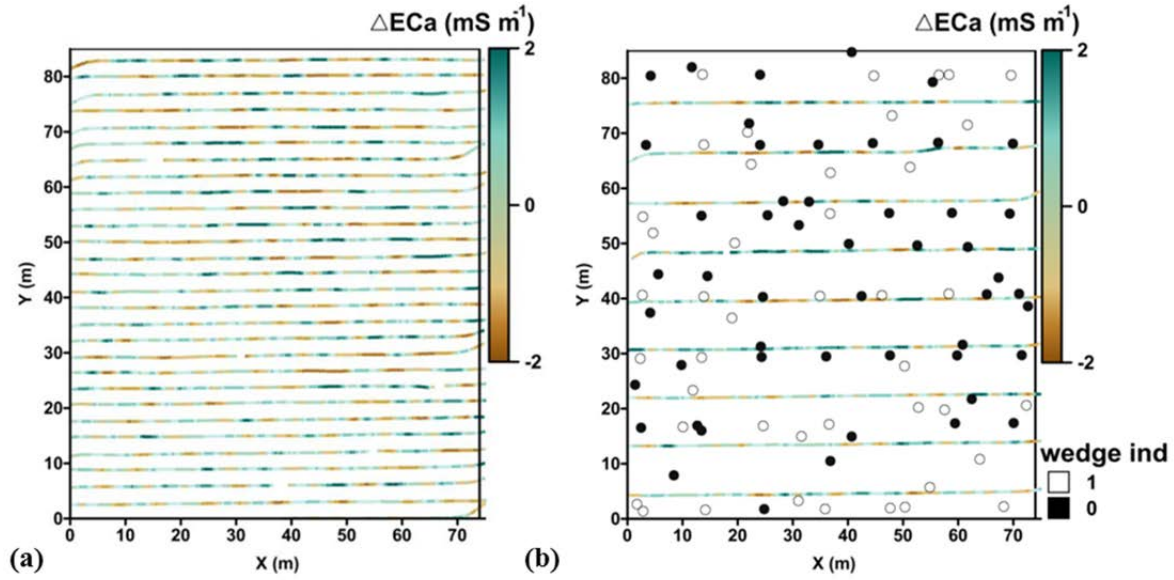


Figure 8.2 Overview of the ΔECa data used for this study with (a) the conditioning data: 28 measurement lines (within-line distance 0.4 m) (mS m^{-1}) and (b) the validation data: 9 measurement lines (within-line distance 0.4 m) (mS m^{-1}) and 94 classified bore hole samples indicating the presence (wedge indicator 1) or absence (wedge indicator 0) of wedge material in the subsoil (0.6 - 0.8 m).

In this study, we used the entire study area as surveyed in chapter 3, but we only used a subset of the ΔECa data having an inter-line measurement distance of approximately 3 m and a within-line distance of 0.4 m. Figure 8.2a shows the selected data set, referred to hereafter as 'sensor data'.

To validate the interpolated sensor data maps we selected nine other measurement lines with an inter-line distance of approximately 9 m and a within-line distance of 0.4 m (Figure 8.2b). The lines were positioned in the middle of two conditioning data lines. To validate the prediction of wedge material in the subsoil, we used the 94 classified soil samples (0.6 m - 0.8 m): 43 samples were classified as wedge material and 51 as host material (Figure 8.2b) (section 3.3.2).

8.2.2 Two-point geostatistics and predictive classification

First, the sensor data were interpolated to a regular grid (cell size 0.1 x 0.1-m) with OK using a spherical variogram model (C_0 = 0, C_1 = 1.7, a = 4.3 m) (Figure 8.3) (Goovaerts, 1997). The model was fit to the experimental variogram considering only data pairs in the direction of the driving lines. This directional variogram was more stable than the omnidirectional one which showed a jump at lag distances around 3 m, corresponding to the inter-line distance. This strategy could be applied since we assumed that the anisotropy shown by the experimental variograms was caused by the sampling configuration, whereas the spatial process being studied was assumed isotropic. We defined an elliptical search window with the longest radius perpendicular to the driving direction to ensure that neighbours from different measurement lines were selected.

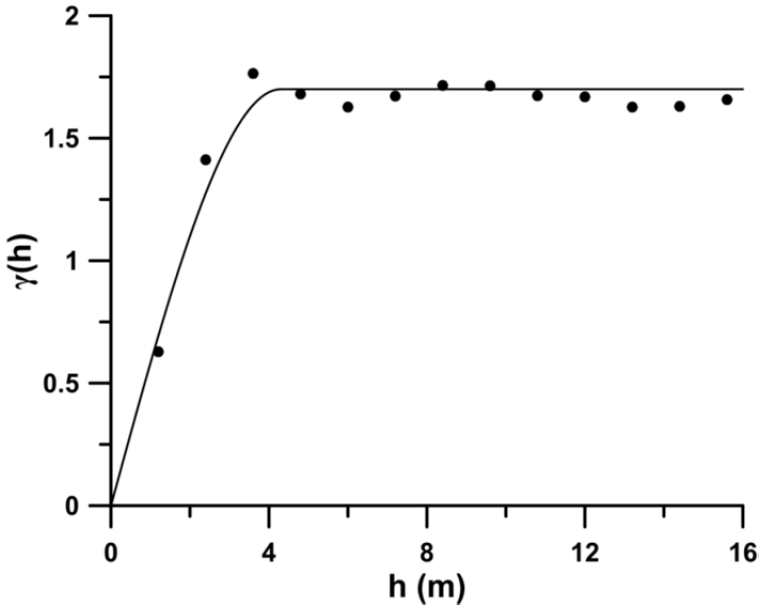


Figure 8.3 Spherical variogram model with C_0 = 0, C_1 = 1.7 and a = 4.3 m used to interpolate the proximal soil sensor data with traditional two-point geostatistics (OK).

Then, we performed a fuzzy *k*-means classification of the interpolated sensor data. Since fuzzy-set theory can deal with uncertainty especially due to imprecise boundaries between categories (McBratney and Odeh, 1997), this technique was appropriate to classify the soil into two classes: one area with host material and one with wedge material in the subsoil (from 0.6 to 0.8 m depth). Although theoretically required for predictive classification (de Gruijter and McBratney, 1988), we did not add an extragrade class here since this would complicate comparison with the MPG probability map. We used the FuzME software (Minasny and McBratney, 2002) and set the fuzziness exponent φ at 2.1 following the scheme proposed by McBratney and Moore (1985). Parameter φ controls the degree of fuzziness of the classification and has a value between 1 (hard classification) and ∞. The resulting fuzzy membership map for the wedge material class was interpreted as the possibility to find wedge material in the subsoil.

8.2.3 Multiple-point geostatistics

Bivariate MPG requires the construction of a bivariate TI. For this case study the TI needed to consist of a categorical image of the target variable (TI1), i.e. an indicator for the presence of wedge material in the subsoil, and a continuous image of the ancillary variable (TI2), i.e. the sensor data. Both TI1 and TI2 needed to represent the expected spatial structure of the corresponding variable and the bivariate image needed to represent the expected relationship between both variables. We built this bivariate TI based on our physical knowledge of the crack formation and the sensor measurements on the one hand, and the information we gathered during the field work on the other hand, i.e. the excavation and the prediction sensor data (Figure 8.2a).

TI1 was built from a binary image of a polygonal network of desiccation cracks in a Mexico silt loam, that we selected from literature (Baer et al., 2009). We resized the image to an image of 700 pixels high and 700 pixels wide (bicubic interpolation), each pixel representing an area of 0.01 m^2. Then, we dilated the wedges considering the width of the excavated polygon (see section 3.3.1). Figure 8.4a shows the resulting image that was used as TI1. TI2 was obtained by a forward modelling procedure predicting the corresponding sensor data starting from TI1. We spatially filtered TI1 with a kernel (11x11 pixels) representing the depth-response curve of the EM38DD soil sensor (McNeill, 1980). This filtered image was histogram transformed targeting the histogram of the sensor data (Figure 8.2a). The continuous TI (TI2) is shown in Figure 8.4b. The image processing steps were performed in Matlab (Mathworks, R2011a).

We used the Direct Sampling (DS) code to generate bivariate multiple-point simulations (Mariethoz et al., 2010). In this chapter we used the fraction of non-matching nodes for the categorical variable and the mean absolute error for the continuous variable. The continuous variable was given a weight three times larger than the categorical

variable. The weight given to the conditioning data was set five times larger than the weight given to the already simulated grid nodes (see section 5.3.4).

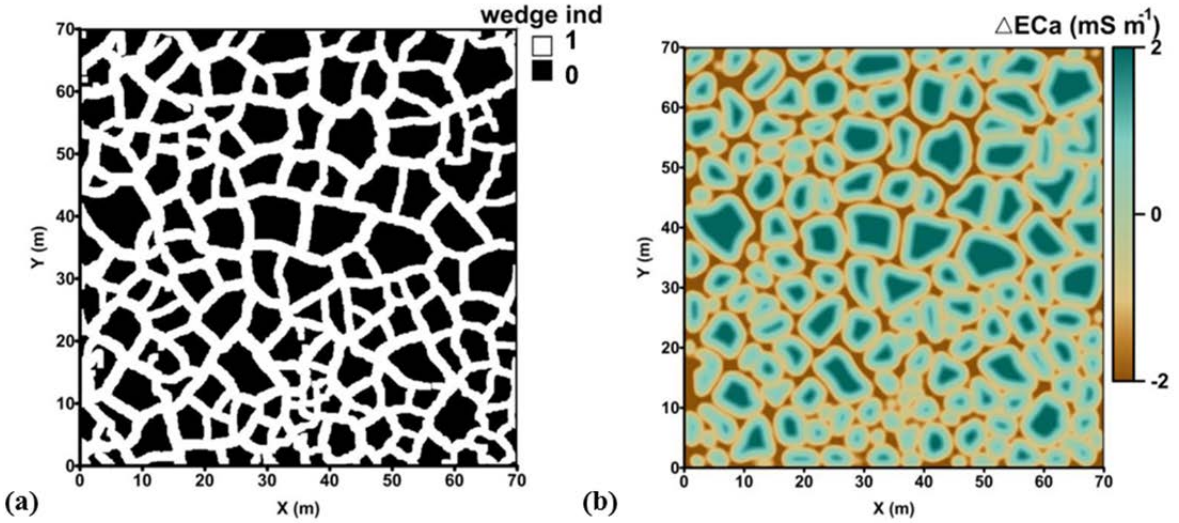


Figure 8.4 Bivariate TI used to interpolate the proximal soil sensor data and predict the target variable with multiple-point geostatistics with (a) the categorical image representing the spatial pattern of the wedge indicator (TI1) and (b) the continuous image representing the spatial pattern of the sensor data (TI2).

We ran 50 bivariate simulations with the constructed bivariate TI (Figure 8.4) and the sensor data as continuous conditioning data (Figure 8.2a). The resulting E-type for the continuous variable served as an interpolated sensor data map and the E-type for the categorical (binary) variable served as a probability map for the presence of wedge material in the subsoil.

8.2.4 Validation

The interpolated sensor data maps were validated by comparing the measured sensor values in the independent measurement lines (Figure 8.2b) with the estimated values at the closest grid node. For both the map interpolated with two-point geostatistics and the one interpolated with multiple-point geostatistics, we made a scatterplot and calculated five validation indices: the mean estimation error (MEE), the root mean square estimation error (RMSEE), the mean absolute estimation error (MAEE), the Pearson's correlation coefficient (r) and the Spearman's rank correlation coefficient (r_R).

Based on the 94 classified bore hole samples (Figure 8.2b), we validated the two maps predicting the presence of wedge material in the subsoil by calculating their receiver-operating characteristic (ROC) curve (Pontius and Schneider, 2001). This method was chosen since a ROC curve evaluates the two-class prediction performance of the maps independent of the chosen decision threshold. This is important to compare the fuzzy membership value map more objectively with the probability map derived with MPG. The effect of the degree of fuzziness, as is defined by φ, will not influence the comparison. The

ROC space is defined by the 1-specificity (false positive rate) and the sensitivity (true positive rate) as x- and y-axes respectively, considering a continuous range of decision thresholds. The top left corner is the optimal location of the ROC space since there both the specificity and the sensitivity are 1. The area under the ROC curve (AUC) measures the two-class prediction performance. An AUC of 0.5 indicates a classification performance no better than chance. The closer the AUC is to 1, the better is the classification potential of the maps (Cockx et al., 2007).

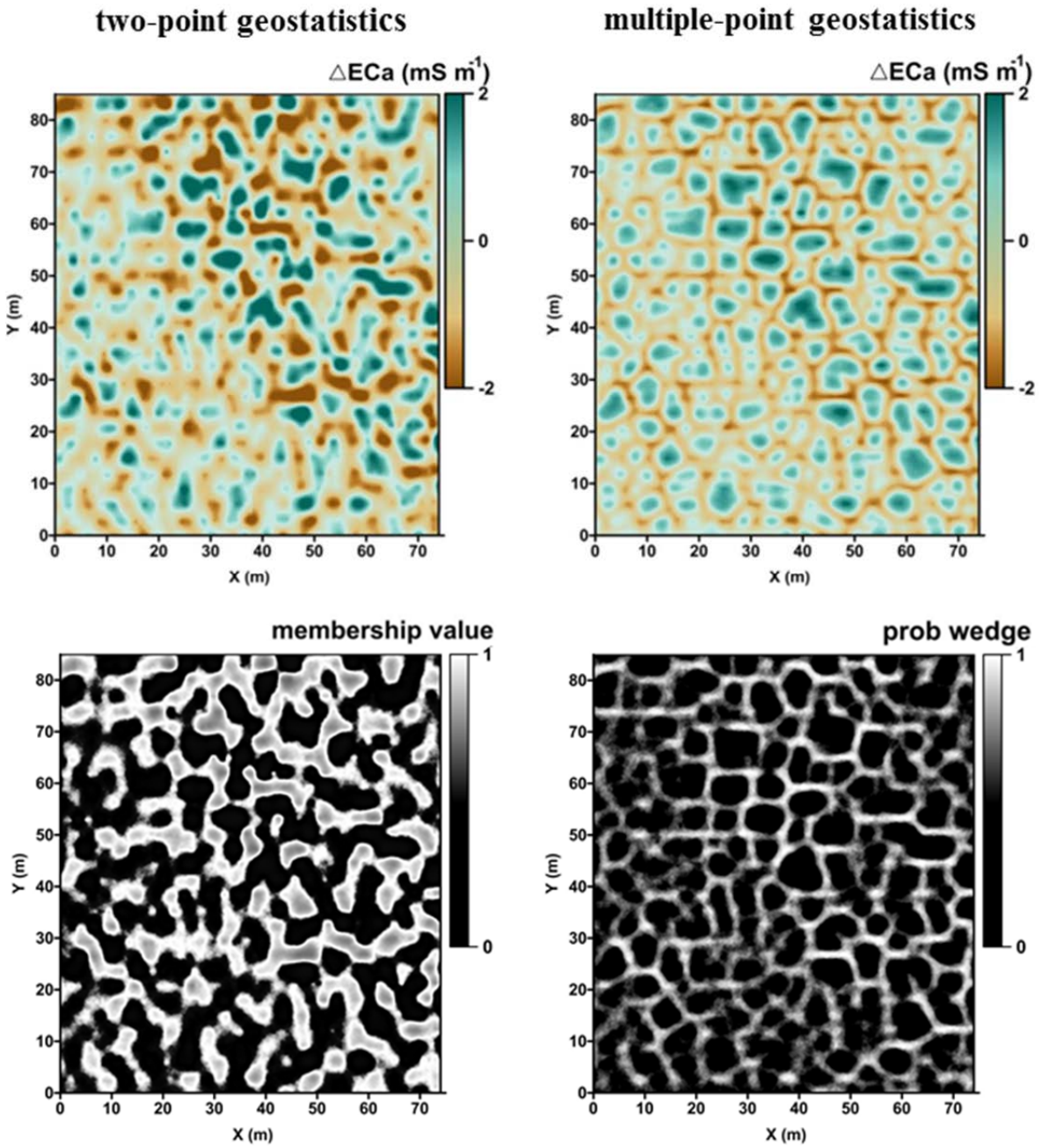


Figure 8.5 Map of the sensor data interpolated with ordinary kriging and derived fuzzy membership value map indicating the possibility to find wedge material in the subsoil (left) and map of the sensor data and probability map to find wedge material in the subsoil simultaneously generated with MPG (E-type of 50 simulations) (right).

8.3 Results and discussion

Figure 8.5 shows the maps generated with traditional two-point geostatistics (left) and multiple-point geostatistics (right). When we compare these maps with the georectified aerial photograph of the polygonal crop marks (Figure 8.6), it is clear that both maps delineate the major ice-wedges very well, especially considering the between-line distance of the input data which was large in relation to the scale of the soil features (Figure 8.2a).

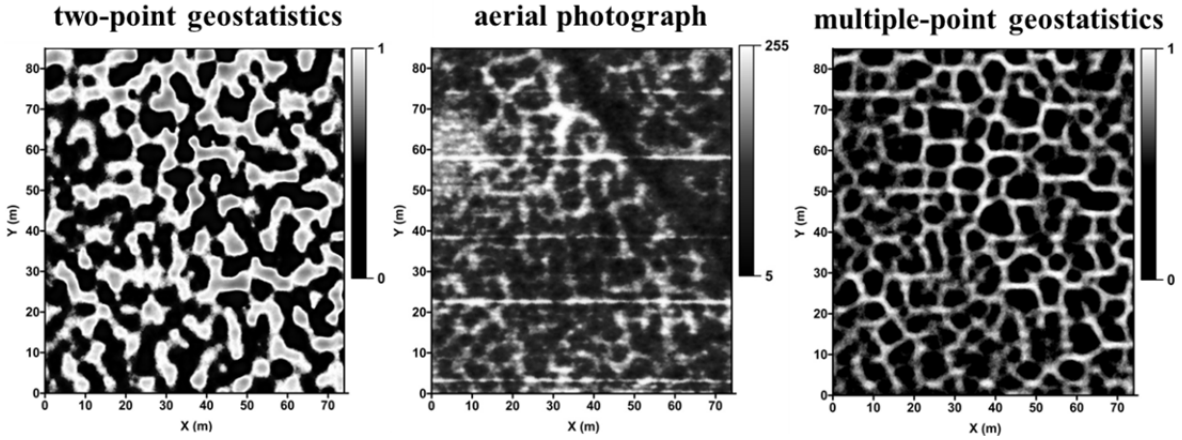


Figure 8.6 Comparison of the aerial photograph (centre) with the TPG (left) and MPG (right) prediction map for the presence of wedge material in the subsoil.

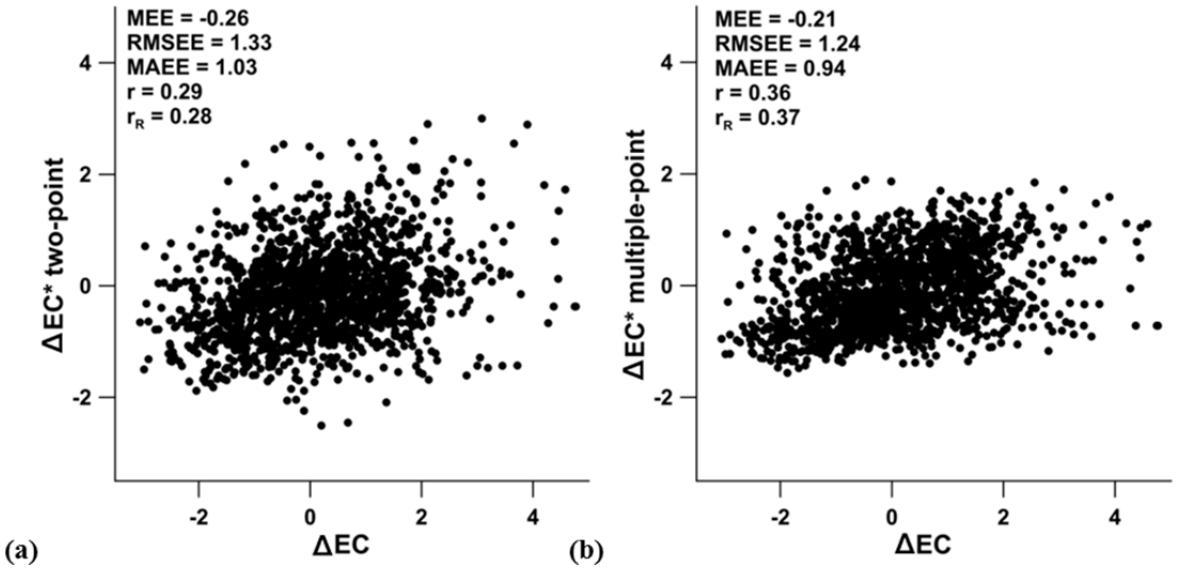


Figure 8.7 Validation results for both interpolated sensor data maps (Figure 8.5 - top) using the independent validation data of 9 measurement lines (Figure 8.2b): scatterplots and validation indices (MEE = mean estimation error, RMSEE, root mean square estimation error, MAEE = mean absolute estimation error, r = Pearson's correlation coefficient, r_R = Spearman's rank correlation coefficient) for the map interpolated with two-point geostatistics (left) and the map interpolated with multiple-point geostatistics (right).

However, the polygonal pattern was much better reconstructed in the MPG maps. The maps based on TPG showed more smoothed polygons and a lack of connectivity for the smaller polygons. This better pattern reconstruction of the MPG maps is due to the use of a TI as a structural model which explicitly implies a multiple-point pattern.

In addition to reconstructing the patterns correctly, the prediction maps also need to be locally accurate. To quantify this local accuracy, we validated the maps as described in section 8.2.4. Figure 8.7 shows the validation results for the interpolated sensor data maps (Figure 8.5 – top). Although the validation scatterplots show a smoothing effect for both maps (the slope shown by the data in both plots is less than one), they predicted the sensor data reasonably well. The scatterplot cloud was more elongated for the MPG E-type map, the correlation coefficients were closer to 1, and the validation indices closer to 0. This shows that the enhanced pattern reconstruction obtained with MPG does not come at the cost of local accuracy.

Figure 8.8 shows the ROC curves for the maps predicting the presence of wedge material in the subsoil (Figure 8.5 – bottom) using the 94 classified bore hole samples (Figure 8.2b). The fuzzy membership value map had an AUC of 0.73 and the probability map created with MPG an AUC of 0.84. This means that the probability of ranking a randomly chosen location with wedge material higher than a randomly chosen location with host material, is higher for the MPG map than for the fuzzy membership value map. Hence, the MPG map was better able to map the polygonal network in the subsoil.

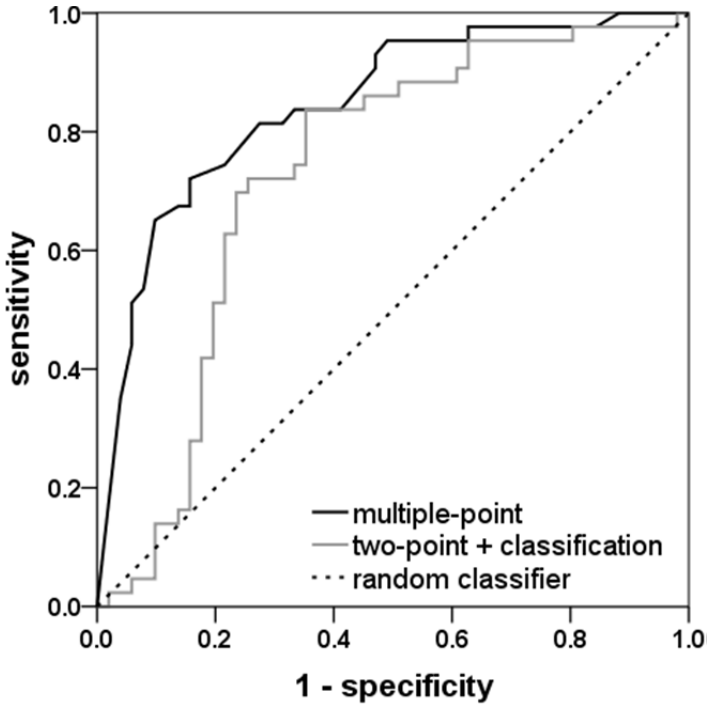


Figure 8.8 Validation results for both maps predicting the presence of wedge material in the subsoil (Figure 8.5 – bottom) using the 94 classified bore hole samples (Figure 8.2b): receiver-operating characteristic (ROC) curves.

8.4 Conclusions

This study shows that bivariate MPG can be used for the processing of proximal soil sensor data. Based on a bivariate TI, we interpolated the proximal soil sensor data between the measurement lines and simultaneously predicted a target variable, i.e. the location of fossil ice-wedge polygons in the subsoil. The use of the sensor data as ancillary variable guaranteed local accuracy, while the multiple-point structural model (TI) ensured pattern reconstruction.

This was the first application of multivariate MPG in soil science, but the flexibility of the method opens up a wide range of potential applications. The variables to be simulated can be categorical and/or continuous and for each variable conditioning data can be given as input data. Furthermore, the (multivariate) TI can be data driven, knowledge driven or a combination of both, like the TI used in this chapter.

Chapter 9

3D reconstruction of sedimentary layers: an industrial application

The aim of this chapter is to translate some of our research findings into an industrial application. We present a TPG and MPG solution to solve a problem that is daily faced by dredging firms, i.e. the reconstruction of the depositional pattern of sedimentary layers in a channel to be dredged. Predicting the thickness of sedimentary layers has direct economic consequences.

9.1 Introduction

An adequate planning of a dredging project requires an accurate prediction of the sediment volumes to dredge and the uncertainties attached to these volume predictions. Next to the total volume to dredge, one is also interested in the depth and thickness of the sedimentary layers for each location along the dredging route. This information is needed to select appropriate dredging equipment, to ensure sufficient disposal options and to draw up a project budget and schedule.

This chapter presents a real case study of a channel to be dredged by DEME N.V. (Dredging, Environmental & Marine Engineering), a Belgian dredging and hydraulic engineering group. The channel will be dredged over a length of ca. 35 km to a depth of 9.5 m allowing larger ships access. For confidential reasons, no geographical details about the study area can be provided. Bore hole samples that were collected in a pre-dredging survey provided information about the sedimentary layers: a soft sediment layer, consisting of sand and loose (soft) clay, covered a hard sediment layer, consisting of compacted (stiff) clay. The aim of this research was to reconstruct the depositional pattern of the sedimentary layers in order to optimize the dredging campaign.

DEME N.V. solved this case study by applying nearest neighbour interpolation whereby the predictions at unsampled locations are provided by the nearest observation. More advanced geostatistical tools that allow to simulate sedimentary layers are indicator kriging (IK) (e.g. Bastante et al., 2005) and sequential indicator simulation (SISIM) (e.g. Seifert and Jensen, 2000). These non-Gaussian approaches allow to simulate categorical variables (Goovaerts, 1997). An advantage of SISIM over its kriging counterpart is that each equiprobable SISIM realization represents an alternative scenario for which the total volume to dredge can be calculated, resulting in a probability distribution of total volume predictions, and corresponding project budget predictions. An alternative is using a MPG approach to map sedimentary layers (e.g. Jung and Aigner, 2012). Bastante et al. (2008) and dell' Arciprete et al. (2012) compared TPG indicator approaches with MPG approaches to model depositional patterns.

This chapter presents two alternative solutions to reconstruct the sedimentary layers in the channel to be dredged: SISIM, a two-point geostatistical algorithm, and IMPALA, a multiple-point geostatistical algorithm. The aim was to provide improved workflows, while ensuring practicality for the company involved. Before the presentation of the case study and the results, we summarize the theory behind both algorithms. All geostatistical analyses were performed with Isatis (Bleinès et al., 2011).

9.2 Theory

9.2.1 Two-point geostatistical algorithm: SISIM

Sequential Indicator Simulation (SISIM) is a widely used sequential simulation technique for categorical variables (Goovaerts, 1997). We used the SISIM algorithm as implemented in Isatis. Consider the simulation of the spatial pattern of a finite number *K* of mutually exclusive categories z_k conditional to the data set $\{z(\mathbf{x}_\alpha), \alpha = 1,2,...\}$. SISIM first transforms all categorical data $z(\mathbf{x}_\alpha)$ into a vector of *K* indicator data:

$$i(\mathbf{x}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{x}_\alpha) = z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1,..., K\,. \tag{9-1}$$

These indicator data are migrated to their closest grid node. Then, a random path is defined visiting each unsampled grid node **x**.

The *K* conditional probabilities of occurrence for each category z_k, i.e. $p^*(\mathbf{x}, z_k|(n))$, are predicted by simple or ordinary kriging using the neighbouring indicator data, consisting of both the conditioning data and the previously simulated grid nodes. One can use a separate indicator variogram model for each $i(\mathbf{x}_\alpha; z_k)$, accounting for class-specific patterns of spatial variation, or one representative variogram model. The latter is allowed when the shape and the anisotropy of the different indicator variograms are similar. It is

known as median indicator kriging because one generally uses the variogram model for the median indicator (in the case of continuous variables).

The K predicted probabilities are normalized ensuring their sum to be one. The vector of the K predicted probabilities is then considered as the conditional probability distribution function (cpdf). The simulated category for $\mathbf{x}$ is the category that has been randomly sampled from this cpdf. Then, the algorithm proceeds to the next node along the random path and the previous steps are repeated (Goovaerts, 1997).

Just as conventional 3D variogram models, 3D indicator variogram models are generally fitted to the experimental variograms calculated in the horizontal and vertical direction. The nugget effect C_0 must be chosen to be equal in all directions. Geometric anisotropy allows one to model different ranges a for different directions, whereas zonal anisotropy allows one to model different sills C_1 in different directions (Gringarten and Deutsch, 2001).

9.2.2 Multiple-point geostatistical algorithm: IMPALA

IMPALA stands for Improved Multiple-point Parallel Algorithm using a List Approach (Straubhaar et al., 2010). It is a recent MPG algorithm for categorical variables that is implemented in Isatis (Bleinès et al., 2011). The algorithm proceeds in three steps. First the TI is scanned, then the conditioning data are migrated to their closest grid node, followed by the sequential simulation. Similar to SNESIM (Strebelle, 2002), IMPALA uses a multi-grid approach to capture structures within the TI that are defined at different scales. The user defines the number of subgrids m. First, the grid nodes $\mathbf{x}$ with coordinates that are a multiple of 2^{m-1} are simulated (coarsest grid), followed by the grid nodes with coordinates that are a multiple of 2^{m-2}, etc. (Figure 9.1a).

IMPALA scans the TI for all possible TI patterns $\mathbf{d}_n(\mathbf{y})$, i.e. the pattern configurations for each location $\mathbf{y}$ in the TI grid (except for the grid borders). TI patterns have a rectangular shape in 2D, and a box shape in 3D. The dimensions of the search template are defined by the user. For each TI pattern found in the TI, its frequency of occurrence is saved together with its frequency of occurrence having a central node value equal to z_i (chapter 2, Figure 2.5). These TI scan results are stored in lists, instead of trees. The TI scan is repeated for each of the m subgrids. The lag vectors of the search template, are rescaled for each subgrid, as shown in Figure 9.1b.

The conditioning data are migrated to all subgrids. The observed facies are assigned to the node in the simulation grid whose corresponding region contains the observed location. If two or more data points are migrated to the same grid simulation node, only the data point that is the closest to the centre of the corresponding region of the grid node is kept.

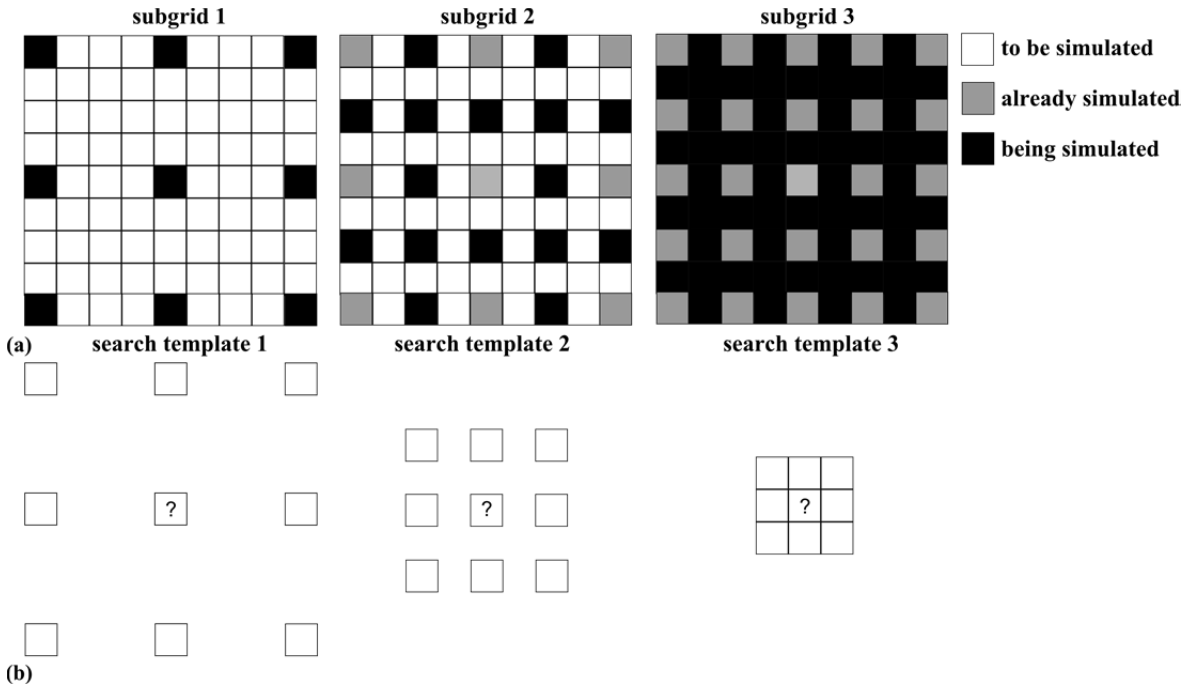


Figure 9.1 Principle of the multi-grid approach for $m = 3$.

For each subgrid, a random path is defined visiting all $\mathbf{x}$. Based on the data event $\mathbf{d}_n(\mathbf{x})$ of $\mathbf{x}$, $p^*(\mathbf{x}; z_k|(n))$ is retrieved from the corresponding list. A random value is drawn from this cpdf and the node is used as conditioning data point for the next $\mathbf{x}$. More details about IMPALA can be found in Straubhaar et al. (2010) and the software manual of Isatis (Bleinès et al., 2011).

Whereas 3D TPG requires a 3D variogram model, 3D MPG requires a 3D training image. Because finding a 3D TI is challenging, this problem is often overcome by constructing 3D TIs from 2D TIs (Comunian et al., 2012; dell'Arciprete et al., 2012; Okabe and Blunt, 2004).

9.3 Data set and initial data analysis

Because the geographical details of the study area are confidential, the coordinates were transformed and are relative to a local datum (Figure 9.2). The channel to be dredged is part of a big river that flows southeast and connects an economically important city to the sea. The channel is ca. 35 km long and its width ranges between ca. 280 and 430 m. It should be dredged to a depth of 9.5 m with side slopes of 1:5 or 1:10, aiming at a total water volume of 68 418 145 m³.

The dredging firm provided us with a data set of 155 bore holes (Figure 9.2). Note that many bore holes were located outside the boundaries of the channel to be dredged, especially in the central part of the channel where most of the bore holes were located to the west of the channel. The bore holes had been sampled in four different sampling stages,

partly by local authorities and partly by the dredging firm. In this research, we treated all bore holes equally. For each bore hole, the water depth and the lithology of the sediment layers had been reported. The dredging firm assigned a stratigraphy (12 categories) to the lithological borehole intervals.

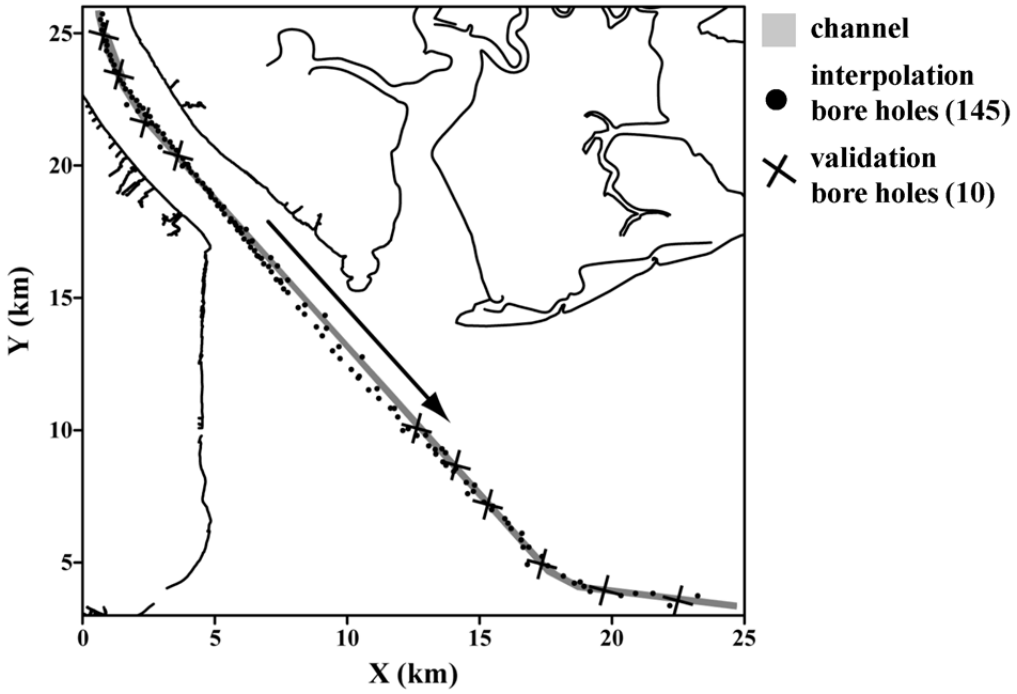


Figure 9.2 Channel to be dredged (X = Eastings; Y = Northings) with indication of the main flow direction and the 155 bore hole locations (145 interpolation bore holes and 10 validation bore holes).

We further classified the stratigraphy in soft (code 1) and hard (code 2) sediment material, each requiring a different dredging strategy. We assigned a code 0 to the water layer, i.e. between the channel bed and the water surface (assumed to be at the 0 m reference level). We added water as a separate class since the water depth is an important variable to be predicted for the entire channel. The average water depth of the 155 bore holes was 7.83 m with a standard deviation of 1.26 m and ranged between 1.49 m and $\geq$ 9.5 m.

Since the data were measured at different support sizes, we first converted the bore holes into composite samples of the same dimension. The bore holes were cut into intervals of the same length (0.1 m) honouring the boundaries between the three categories (water, soft material, hard material) using Isatis (Bleinès et al., 2011). This resulted in a total of 14 746 composite samples.

In contrast to the previous chapter, we did not have a true reference image in this study. Therefore, we randomly selected 10 validation bore holes, corresponding to 950 composite samples, from the 57 bore holes that were located within the channel

boundaries. We chose this rather limited number of validation bore holes to have left sufficient prediction bore holes within the channel boundaries.

Table 9.1 shows the proportions of composite samples for each category. Based on these proportions, we calculated an initial volume prediction: a total volume of 11 940 314 m³ sediment material needs to be removed, consisting of 4 931 986 m³ soft material and 7 008 328 m³ hard material.

Table 9.1 Number of composite samples (0.1 m) and proportions of water, soft sediment layer and hard sediment layer of all composite data, the interpolation composite data and the validation composite data.

	all data (155 bore holes)	interpolation data (145 bore holes)	validation data (10 bore holes)
number of composites	14746	13796	950
% water	82.33	82.29	83.05
% soft layer	7.30	7.19	8.84
% hard layer	10.37	10.53	8.11

9.4 Results and discussion

9.4.1 Nearest neighbour interpolation

Figure 9.3 shows the prediction map as obtained by nearest neighbour interpolation. We defined a simulation grid of 40x40x0.1-m. We chose a rather low resolution in the X- and Y-direction to constrain the simulation time together with a high resolution in the Z-direction because the spatial variation was expected to be larger in the Z-direction than in the X- and Y-direction.

The most northern part of the channel (near X = 2 km and Y = 25 km) has a water depth that is generally larger than 9.5 m. The centre of the channel (near X = 10 km and Y = 12 km) has the thickest hard sediment layer, whereas the bend of the channel (near X = 17 km and Y = 5 km) has the thickest soft sediment layer. The boundaries between the layers do not represent the natural boundaries: their shape solely depends on the data and their sampling configuration.

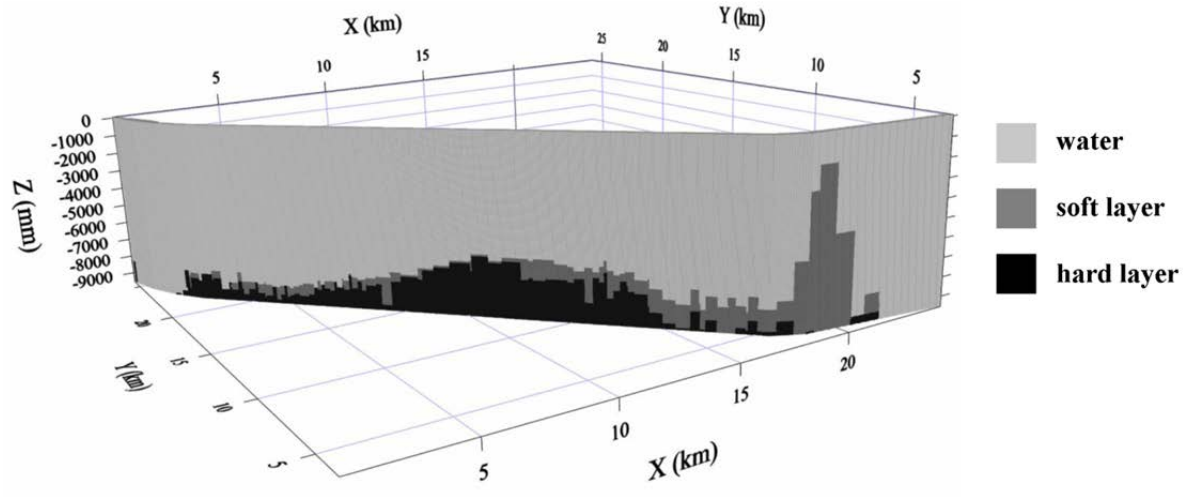


Figure 9.3 Nearest neighbour prediction map.

9.4.2 Two-point geostatistical reconstruction

Figure 9.4 shows the horizontal and vertical experimental indicator variograms for the three categories. The horizontal variogram was calculated from data pairs in the major flow direction of the channel (135° clockwise from the north axis). We applied median SISIM since the spatial patterns of the three categories were similar. The 3D variogram was modelled as a spherical variogram with $C_0 = 0.01$, $C_1 = 0.09$, $a_x = 29\,000$ m, $a_y = 29\,000$ m and $a_z = 5.5$ m (Figure 9.4). The sill was assumed to be constant for all directions and the range to be different for the vertical direction (i.e. geometric anisotropy).

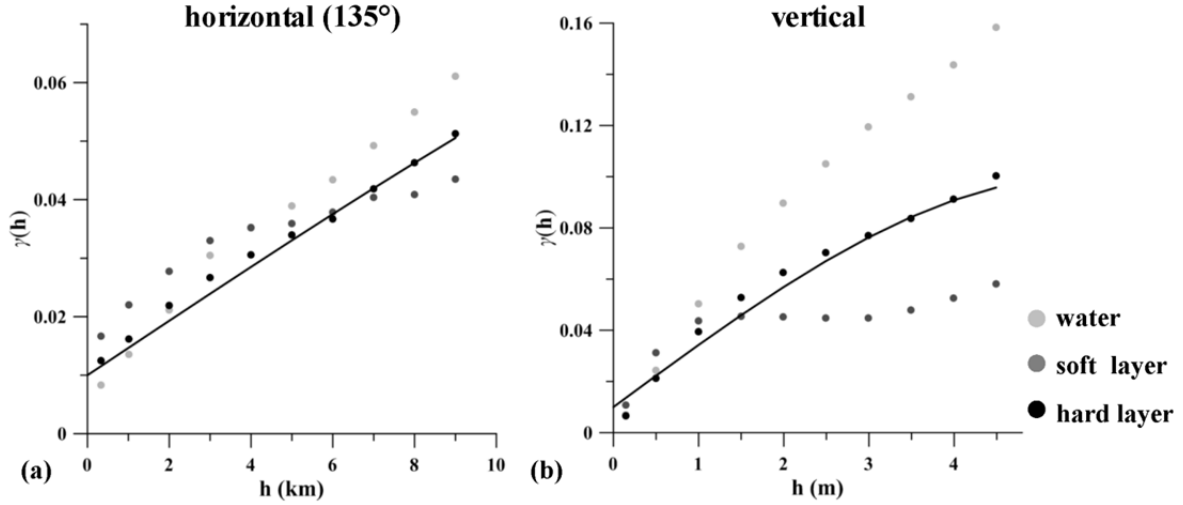


Figure 9.4 3D variogram model used for the TPG reconstruction with (a) the experimental variograms (dots) calculated in the horizontal plane (direction 135°) and the variogram model (line) in the horizontal plane and (b) the experimental variograms (dots) calculated in the vertical direction (tolerance 0°) and the variogram model (line) in the vertical direction.

The same simulation grid as for the nearest neighbour interpolation was used. The search window was defined as an ellipsoid with a radius of 1600 m in the X-direction, 1600 m in the Y-direction and 5 m in the Z-direction. The maximum number of conditioning data was set to 100 and the optimum number of already simulated grid nodes to 100.

Figure 9.5 shows the first SISIM realization together with the conditional mode (most simulated category) of 50 realizations. Both maps respect the sequence of water over soft material over hard material. The simulation map is rather noisy. It shows some isolated pixels and noisy edges between the different layers. The conditional mode map does not show noise and has sharper boundaries between the layers.

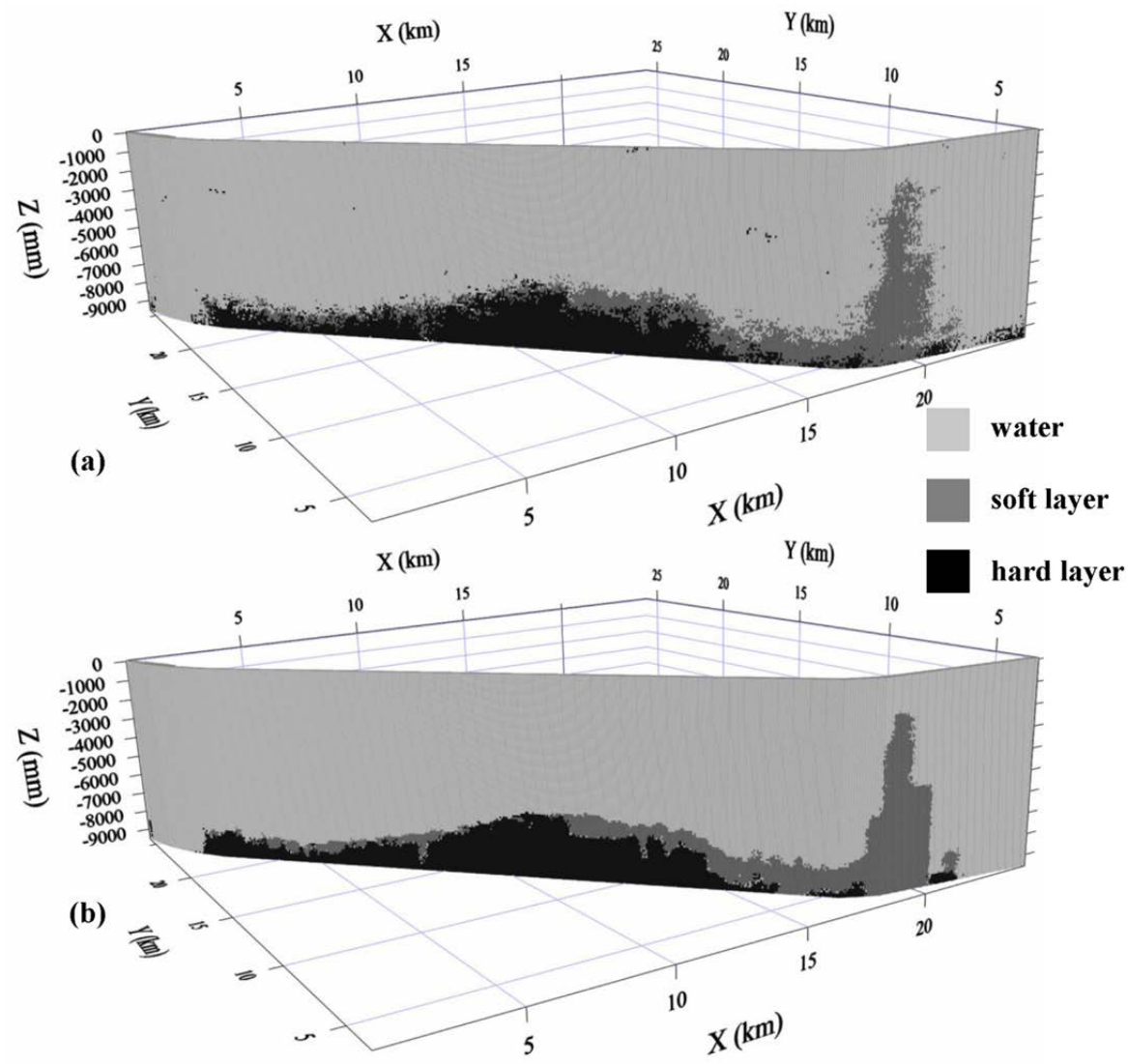


Figure 9.5 (a) First SISIM realization and (b) conditional mode of 50 SISIM realizations.

9.4.3 Multiple-point geostatistical reconstruction

We artificially built a simple 3D TI based on our prior knowledge of the depositional pattern and some summary statistics of the bore hole samples. It consist of three subsequent layers: a water layer of 7.9 m, a soft layer of 1.3 m and a hard layer of 0.3 m (Figure 9.6). We chose these depths to have similar relative proportions in the TI as in the composite data (Table 9.1). Just as for the variogram model, we assumed isotropy in the X- and Y-plane. The TI is a 3D grid of 5 km in the X-direction, 5 km in the Y-direction and 9.5 m in de the Z-direction with a resolution of 40x40x0.1-m. Note that the TI has the same resolution as the simulation grid but that it is much smaller. The simplicity of the TI pattern does not require a larger TI; the TI already provides sufficient pattern repetition.

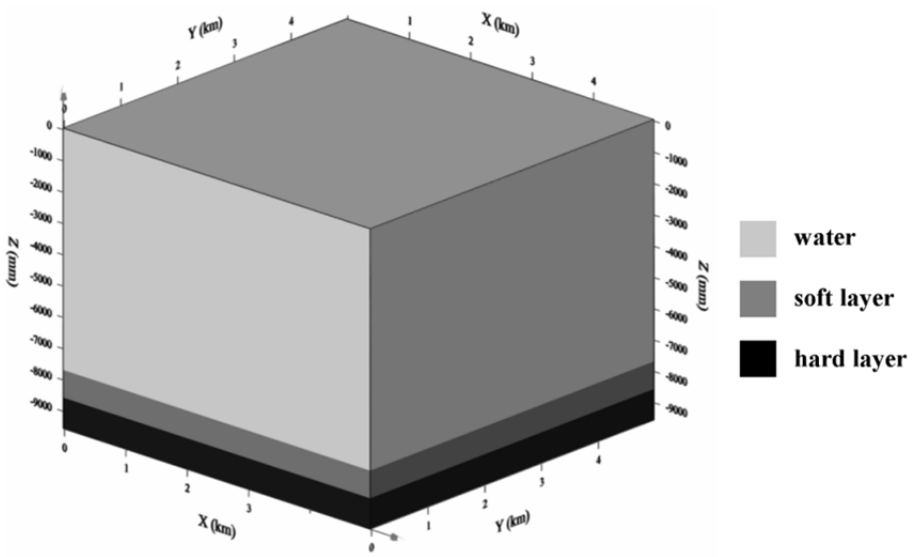


Figure 9.6 3D training image used for the MPG reconstruction of the water and sediment layers.

We used four IMPALA subgrids and defined a search box with a half-side-length of 400 m in the X-direction, 400 m in the Y-direction and 2 m in the Z-direction. The simulation grid was equal to the grid used for the nearest neighbour reconstruction.

Figure 9.7 shows the first MPG realization together with the conditional mode of 50 realizations. Overall, the depositional patterns predicted by the MPG approach were very similar to these predicted by the TPG approach (Figure 9.5). However, the MPG realization is less noisy and corresponds better to the conditional mode map.

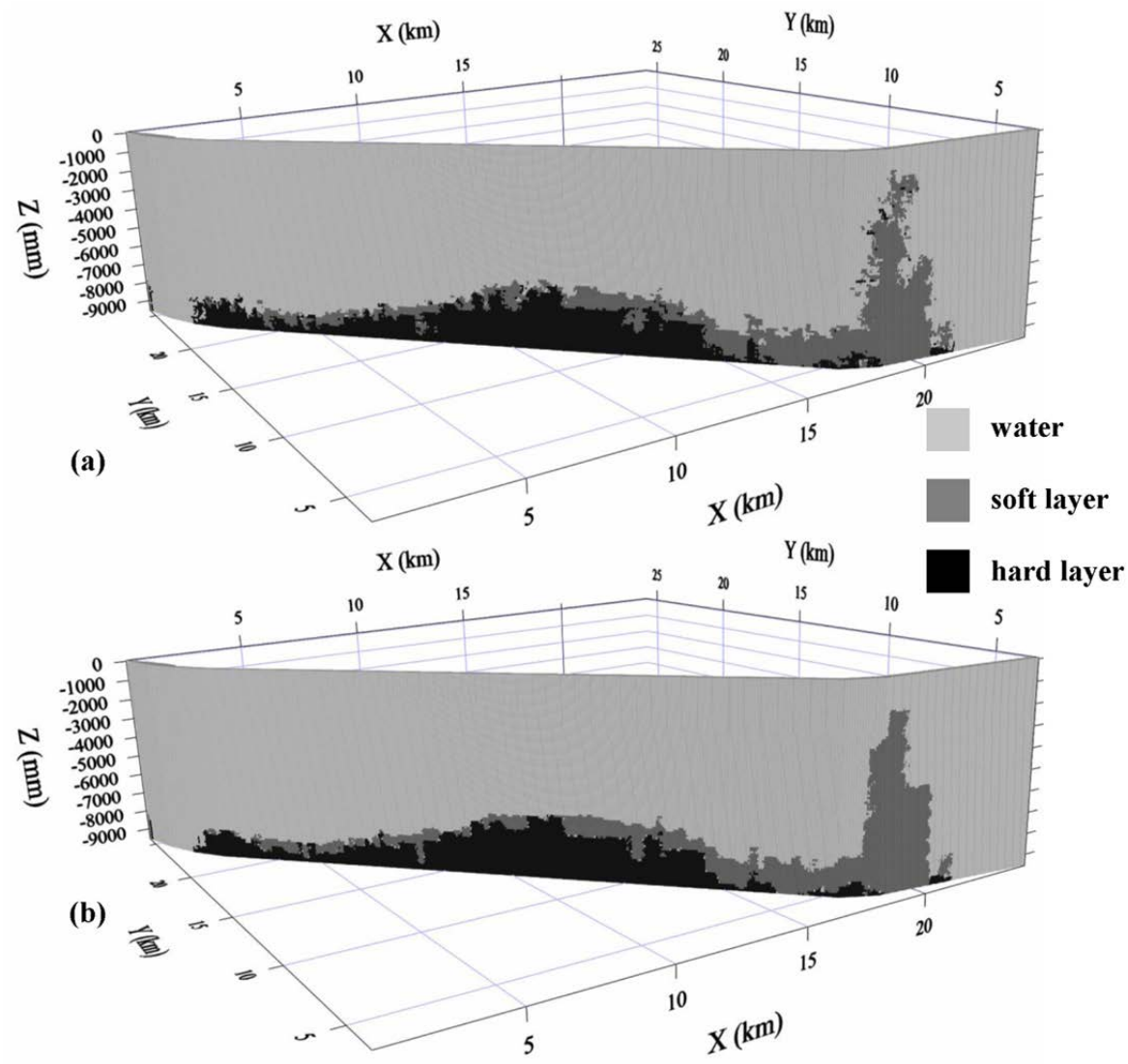


Figure 9.7 (a) First IMPALA realization and (b) conditional mode of 50 IMPALA realizations.

9.4.4 Evaluation and discussion

Validating the prediction maps (Figure 9.3, Figure 9.5b and Figure 9.7b) using the 950 validation composite samples showed that all approaches produced similar and accurate predictions. The overall accuracy was 0.93 for the nearest neighbour prediction, 0.94 for the TPG prediction and 0.96 for the MPG prediction. The measure of agreement kappa (Cohen, 1960) was 0.78 for the nearest neighbour prediction, 0.82 for the TPG prediction and 0.87 for the MPG prediction. Each method had a rather high validation performance, which is partly due to the proximity of the validation data to the interpolation data. However, the suggested TPG and MPG approaches scored consistently higher than the currently used nearest neighbour interpolation. The small discrepancies between the suggested methods are due to the better MPG prediction for the water layer (the most represented category). The soft and hard sediment layer were slightly better predicted by the TPG approach. dell'Arciprete et al. (2102) also concluded from their comparative study

that SISIM reproduced the distribution of the less abundant categories more efficiently than MPG.

Table 9.2 gives the summary statistics of the total volumes to be dredged calculated from the 50 TPG and MPG realizations. The predicted volumes are generally smaller for the MPG than for the TPG reconstruction.

Table 9.2 Summary statistics of the total volume to be dredged (soft and hard layer) calculated from the 50 TPG and MPG realizations.

	Nearest neighbour	TPG	MPG
nb of realizations	1	50	50
min volume (m^3)	-	10 239 769	10 023 483
mean volume (m^3)	10 246 528	10 426 351	10 286 270
max volume (m^3)	-	10 712 893	10 591 233
standard deviation (m^3)	-	113 625	124 146

The nearest neighbour prediction resulted in a total volume to dredge of 10 246 528 m^3. An advantage of using SISIM or IMPALA is thus that they provide a measure of the uncertainty about the volume to dredge. However, it is important to keep in mind that SISIM and IMPALA both target realizations with a predefined histogram: SISIM realizations target the histogram of the input data and IMPALA realizations the histogram of the TI (here similar to the histogram of the input data). This could explain the rather small standard deviations of the volume predictions.

An interesting additional way to evaluate the applied methodologies is generating unconditional realizations. Eliminating the influence of the conditioning data clearly reveals the consequences of selecting a certain algorithm (and its parameters) and a certain model of spatial variation. Figure 9.8 shows a SISIM and IMPALA unconditional realization generated with the 3D variogram model, the TI, and the SISIM and IMPALA parameter settings as described above. Both realizations reconstructed the layer geometry. The IMPALA realizations were less noisy and reconstructed the correct sequence of the layers.

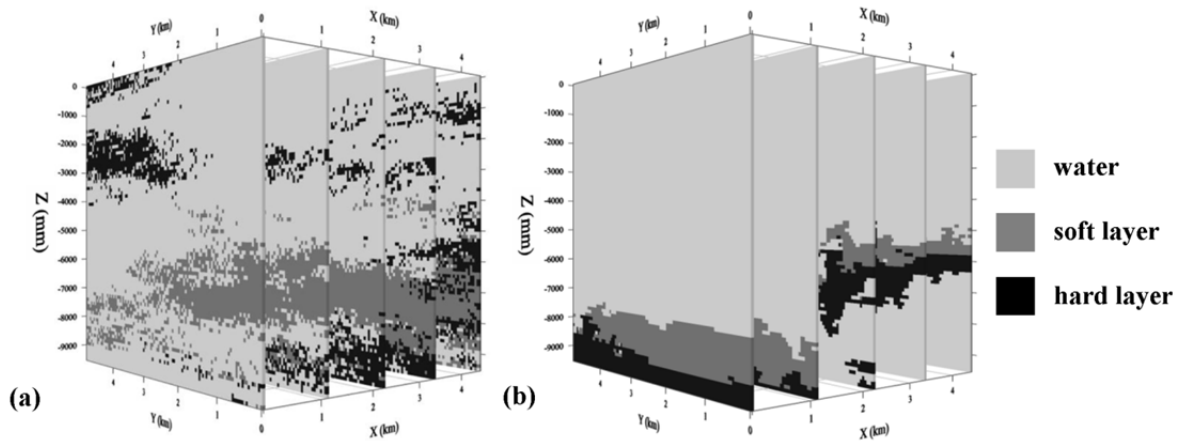


Figure 9.8 An unconditional simulation of (a) the applied TPG technique en (b) the applied MPG technique.

9.5 Conclusions

For this case study, both methods were equivalent. They both required a comprehensive model of spatial variation, whether this is a variogram model or a TI, and a thorough selection of their input parameters. We believe that the choice between a TPG or MGP method for other similar applications will strongly depend on the number of conditioning data and on the complexity of the depositional pattern. Using a TI will be advantageous when the number of conditioning data is small or when the depositional pattern is more complex (Jung and Aigner, 2012).

We found good results by using a very simple 3D TI, which is in line with the Occam's Razor principle. A TI is a conceptual model and can therefore be less complex than the actual depositional pattern. It is the task of the algorithm to anchor the TI patterns to the available conditioning data. Note that the unconditional simulations already showed some spatial variability and did not simply mirror the parallel sediment layers of the TI. We believe that the ease of building a TI is a crucial point to transfer the innovative MPG approach to industrial applications. On the other hand, when expert knowledge or indirect observations (e.g. seismic data) provide more information about the expected depositional pattern, it is no problem to use a more advanced TI.

This chapter showed a practical application of MPG to reconstruct sedimentary layers, suggesting a broader applicability and transferability of the MPG approach to industrial applications. Both TPG and MPG approaches performed well for this case study, demonstrating the complementarity of both approaches. For similar case studies, the user should select an appropriate technique considering his own experience, the number and quality of the available samples and the complexity of the depositional pattern.

Chapter 10

General conclusions and future research

This chapter reflects on the most important conclusions of this thesis and discusses the formulated research objectives (chapter 1). Finally, some ideas for future research are put forward.

10.1 Conclusions

This research contributed to –what is called in medicine– translational research. Translational research facilitates the translation of findings from fundamental science to more practical applications. The foundations of multiple-point geostatistics were already laid during the last decade. Its practical applications are however still scarce, especially in soil science. We indulged in the developed methodology and evaluated its application in soil science, a field where traditional variogram-based geostatistics is still the dominating spatial modelling technique. We hereby focused on theoretical aspects that are important for the implementation of the technique, but priority was given to practicality, *a trademark of geostatistics that explains its success and application in such diverse fields* as stated by Goovaerts (1997).

10.1.1 Collect a test data set of complex soil patterns.

A typical example of complex soil patterns are the ice-wedge pseudomorphs that show up in various parts of Flanders (Belgium) and in other parts of the mid-latitudes of the northern hemisphere. We selected an appropriate field based on aerial photography and collected both direct and indirect observations of the soil. This data set was not only the

basic for most of our MPG experiments, it was also the first time that a subsoil polygonal network was imaged so accurately with a proximal soil sensor.

10.1.2 Fit a non-Gaussian model of spatial variation to the test data set.

We concluded that the spatial pattern of the ice-wedge pseudomorphs could not be modelled with a Gaussian RF and a predefined variogram model. We therefore assessed a theoretical solution to model the polygonal pattern and adapted a geometric RF model to fit our test case. This alternative model was well evaluated because it reconstructed the connectivity of low values. The approach of using soil knowledge to select and fit a non-Gaussian RF model also contributes to the trend of applying more knowledge driven modelling approaches in pedometrics.

However, the theoretical correctness of the geometric RF model conflicts with its practicality. One cannot expect soil scientists to develop alternative –mathematically more complex– random functions for each spatial modelling questions. The difficulty of developing alternative random functions to model complex spatial patterns asks for a more practical approach, that could be provided by MPG.

10.1.3 Perform a sensitivity analysis on an appropriate MPG algorithm.

We found the Direct Sampling (DS) algorithm to be an appropriate MPG algorithm to reconstruct soil patterns due to its flexibility and its wide range of potential applications. Because a well thought setting of the input parameters of DS is crucial for its performance, we derived some general implementation guidelines from the results of a sensitivity analysis on DS in collaboration with the DS developers.

The main three parameters are the acceptance threshold *t*, the fraction of the TI to scan *f* and the number of neighbours *n*. Choosing a small *t* together with a large *f* and *n* generally gives good simulation results but requires a long simulation time and minimizes the variability between different simulations (especially for continuous simulations). Finding an optimal balance for these three parameters is thus important. When one wants to reduce CPU time or the risk of patching, it is advised to first adapt parameter *f*. Especially for categorical simulations, it is good practice to always add one post-processing step for noise removal. For bivariate simulations, we showed that the weights given to each variable clearly affected the simulation quality. Continuous variable simulations could be improved by adding an auxiliary categorical variable that is co-simulated with a relative small weight. To improve data conditioning, it is interesting to put the weights given to the conditioning data (parameter δ) higher than the weights given to the already simulated nodes.

10.1.4 Evaluate the potential of MPG to reconstruct complex soil patterns using the test data set.

We found that MPG, using an appropriate TI, was well suited to reconstruct the polygonal pattern of the ice-wedge pseudomorphs. Both the simulation and prediction maps honoured the conditioning data and reproduced the spatial pattern. The prediction maps, which are particularly important in soil science, were also locally accurate. It is interesting that MPG can be used for the reconstruction of continuous and categorical variables, since both frequently appear in soil science. Especially for mapping categorical variables, the traditional geostatistical toolbox is rather limited.

We used a continuous TI that was built using the geometric RF model and one that was built from more densely measured neighbouring areas. Despite the fact that the geometric RF model was explicitly fit to the test case (using soil knowledge), the data driven TI gave the best results. For the categorical case, we used a photograph showing a spatial pattern that was expected to be similar to the studied phenomenon and adapted it to the histogram of the observations, which gave good results.

We concluded that using an entirely data-driven TI is a straightforward and promising approach. When using a knowledge driven TI, such as a model outcome or a photograph that is deemed to show similar spatial patterns, we found good results when the image was first transformed to have a histogram similar to the conditioning data.

A comparison with a TPG approach for the continuous case showed the stronger pattern reproduction capacity of MPG. Whereas TPG simulation algorithms reproduce the variogram and let the multiple-point statistics depend on the chosen algorithm, MPG reproduces the multiple-point statistics of the TI. However, we want to remark that comparing the mapping performance of MPG and TPG is delicate. The reconstruction quality of both approaches strongly depends on the number and quality of the conditioning data, the appropriateness of the model of spatial variation (which is a variogram for TPG and a training image for MPG) and the implementation of the algorithm (its parameter settings).

10.1.5 Investigate whether MPG can be used for the processing of proximal soil sensor data.

We concluded that MPG is an appropriate toolbox for the processing of proximal soil sensor data. In a first case (chapter 7), we interpolated proximal soil sensor data in areas that were inaccessible for the mobile soil sensor. Characteristic for this approach is its very easy implementation. The only DS input file needed contained the neighbouring high density data interpolated to a regular grid between the measurement lines and showing gaps in inaccessible areas. When the input parameters of DS were thoroughly set, the sensor values in the inaccessible areas could be well reconstructed. When the gaps were

large relative to the size of the soil features, only simulation maps could be provided, whereas the gaps were small relative to the size of the soil features, both simulation and prediction maps could be provided.

In a second case (chapter 8), we used bivariate MPG to interpolate sensor data within measurement lines (assuming no inaccessible areas) and to predict a categorical target variable simultaneously. The suggested approach gave good results, especially in comparison with a more traditional approach of applying ordinary kriging followed by a fuzzy *k*-means classification. However, in contrast to the previous case, the approach is more difficult to implement because a bivariate TI needs to be constructed. Hence, we believe that this approach is promising but not immediately useful in practical applications. On the other hand, we demonstrated one of the first applications of the only recently developed bivariate MPG.

Consequently, to evaluate this research objective, we showed two processing problems that could be solved using MPG. However, we believe that there is a wide spectrum of other processing steps of proximal soil sensor data, or high density data in general, where MPG can be useful. Processing high density data with a variogram function as model of the spatial structure results in a loss of information. The data itself often reveal their multiple-point patterns (such as connectivity, curvilinearity, repetitivity), but this information is ignored in the variogram model. We therefore believe that the current trend of collecting more and more high density observations goes hand in hand with the application of MPG.

10.1.6 Evaluate the practical use of MPG in an industrial application.

To date, MPG is already being used in the petroleum industry. However, the technique could be of assistance in other industrial fields. We applied a MPG approach to reconstruct the sedimentary layers in a channel to be dredged by DEME N.V. The MPG approach consisted of using the IMPALA algorithm and a simple categorical 3D TI. We also suggested a TPG approach, i.e. sequential indicator simulation. Both stochastic geostatistical approaches showed a rather equivalent performance. This is because the sedimentary layers did not have a very complex spatial pattern (it could be represented with a variogram model) and because there were sufficient bore hole samples. Both approaches gave better reconstruction results than a nearest neighbour interpolation, a deterministic interpolation method that is currently used by the company due to budget and time constraints and a lack of training. A second advantage of the stochastic geostatistical approaches is that they allow to calculate the uncertainty related to the predicted sediment volumes to dredge, allowing some risk analysis.

10.1.7 General conclusions

Our general conclusion is that MPG is an innovative technique that can be a valuable part of the pedometrician's toolbox. This was proven by the different successful MPG applications in soil science throughout this thesis. We believe that TPG and MPG are complementary techniques and the user should select the technique that is best suited to solve the particular problem. We do not state that MPG is a better method than TPG, but we believe that it is more flexible.

MPG is flexible because it can be both knowledge and data driven, and therefore belongs to the *hybrid* approaches. Soil scientists often have some prior conceptual knowledge about the studied phenomenon. It is easier to translate these concepts into an image than into a variogram function. Note that the approach reconnects with the first soil mappers who based their maps especially on their knowledge of the soil. Lark (2012a) calls it *a successful sign of scientific progress when one makes space in quantitative models of soil variation for understanding of the soil.*

The second reason for the flexibility of MPG is its capacity to model a broad range of spatial patterns. Soil scientists often face periodic, connected or curvilinear patterns that are difficult to represent with a random function model. However, the use of MPG is not restricted to complex spatial patterns. A TI can also be simple as was shown in chapter 9.

There are some perceptions about constructing a TI that prohibit soil scientists to experiment with MPG. An often mentioned constraint is that TI construction requires a larger effort than variogram modelling. Whether this is true or not, strongly depends on the situation. For instance, the TI that was built from the geometric RF model (chapter 6) and the bivariate TI (chapter 8) were indeed rather difficult to construct, but the TIs constructed from high density measurements (chapter 6 and 7) and the simple 3D TI (chapter 9) were obtained in a straightforward way. Today, there is an increasing availability of soil covariates helping the soil scientist to accurately model spatial variation. This soil covariate information, like aerial photographs or proximal soil sensor images, might be used to construct TIs which can be more generally applied (since they do not need to contain any local information).

The second perception is that TI construction is more subjective than variogram modelling. Indeed, constructing a TI forces the user to make decisions about the multiple-point statistics instead of just accepting them implicitly. However, we follow the opinion of Journel and Zhang (2006) and consider the explicit visualisation of multiple-point statistics in a TI as an advantage because it allows a visual inspection of the full model of spatial variation applied. Multiple-point maps are very sensitive to the chosen TI, but one should realise that this sensitivity is not stronger than the sensitivity to the combination of a variogram model and implicit high-order assumptions.

Just as soil scientists implemented variogram-based geostatistics as developed in the mining industry, we believe that the time is ready to also implement multiple-point geostatistics as developed in the petroleum geology. We want to encourage soil scientists to rethink the –often blindly applied– variogram modelling procedure by evaluating first if the expected spatial pattern cannot be better or more easily modelled with a training image.

10.2 Future research

'In theory, theory and practice are the same. In practice, they are not.' Albert Einstein.

The MPG theory is rather recent and therefore exposed to diverse theoretical improvements that are suggested in scientific literature. Main MPG research lines are the assessment of the sensitivity to the input training image, the use of parameterized TIs, the development of appropriate (multiple-point) validation statistics to evaluate the pattern reconstruction of MPG simulations, and the search for improvements to existing MPG algorithms or the development of new MPG algorithms aiming to increase the simulation quality and the computational efficiency,

The search for theoretical improvements of MPG is essential, but it should not delay the implementation of MPG in practice because often a lot can be learned from practical applications. The main obstacle to a wide implementation of MPG in practice is the need for user friendly and easily attainable MPG software packages. The SNESIM algorithm that is implemented in SGeMS and the IMPALA algorithm in Isatis are pioneering software tools, but the other developed algorithms and the theoretical improvements should follow.

Concerning the application of MPG in soil science, we suggest further research to focus especially on the construction of TIs from high density soil sensor data. Solutions should be found to judge the suitability of these data to be used as the basic for TI construction and to develop criteria for soil science TIs. A related research question is for instance which data transformations, such as the histogram transformation that was often applied in this thesis, are required. Another idea for future research is the development of libraries of parameterized soil TIs for frequently appearing soil features, as has already been realised in hydrogeology. Geometric RF models, such as the RF model for the polygonal ice-wedges, could serve as the basic to construct such libraries.

Appendix

Supplementary figures chapter 5

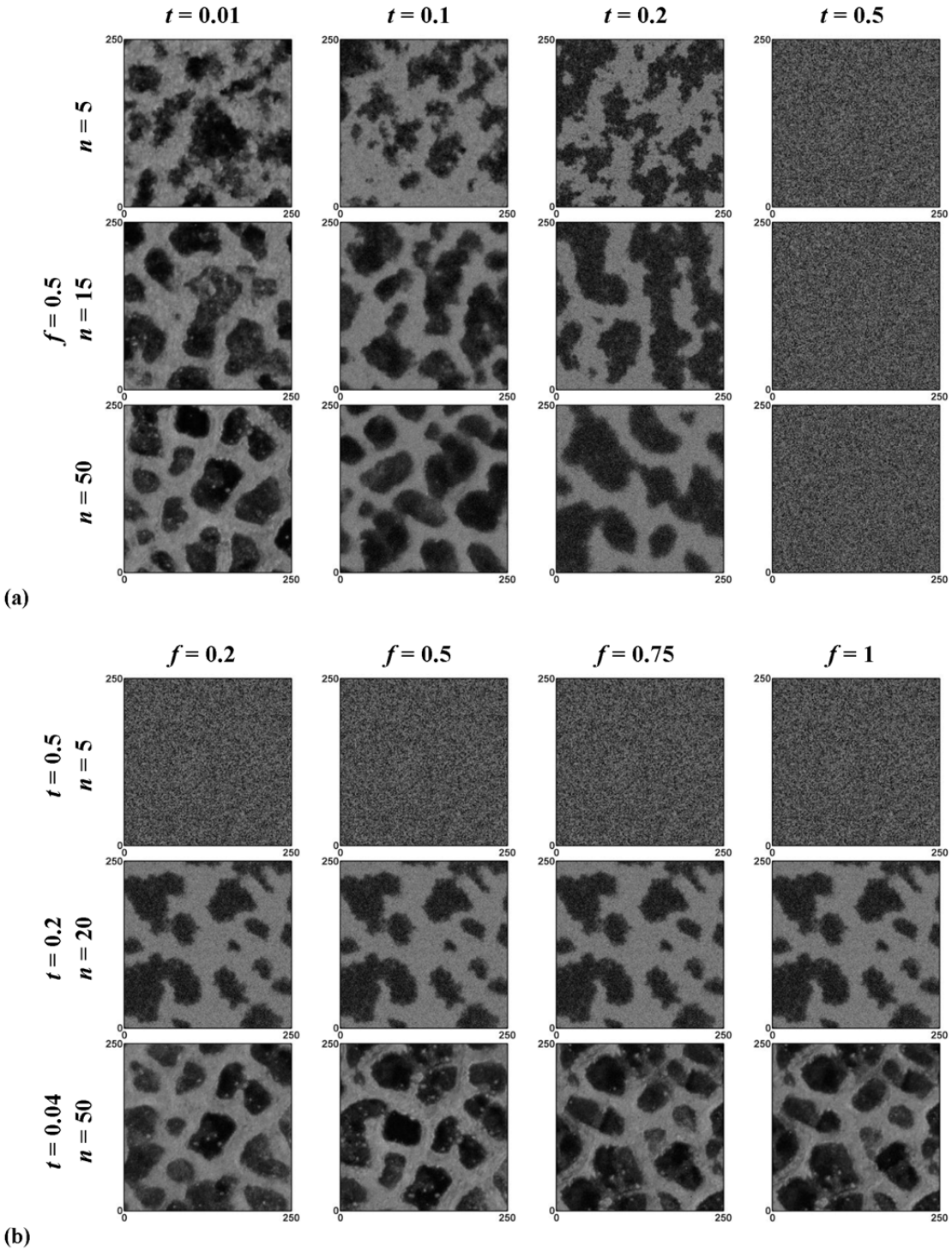


Figure A (a) First out of 10 unconditional simulations illustrating the effect of parameters *t* and *n* with *f* = 0.5 and (b) first out of 10 unconditional simulations illustrating the effect of *f* for constant *t* and *n* based on the continuous ice-wedge TI (Figure 5.3a).

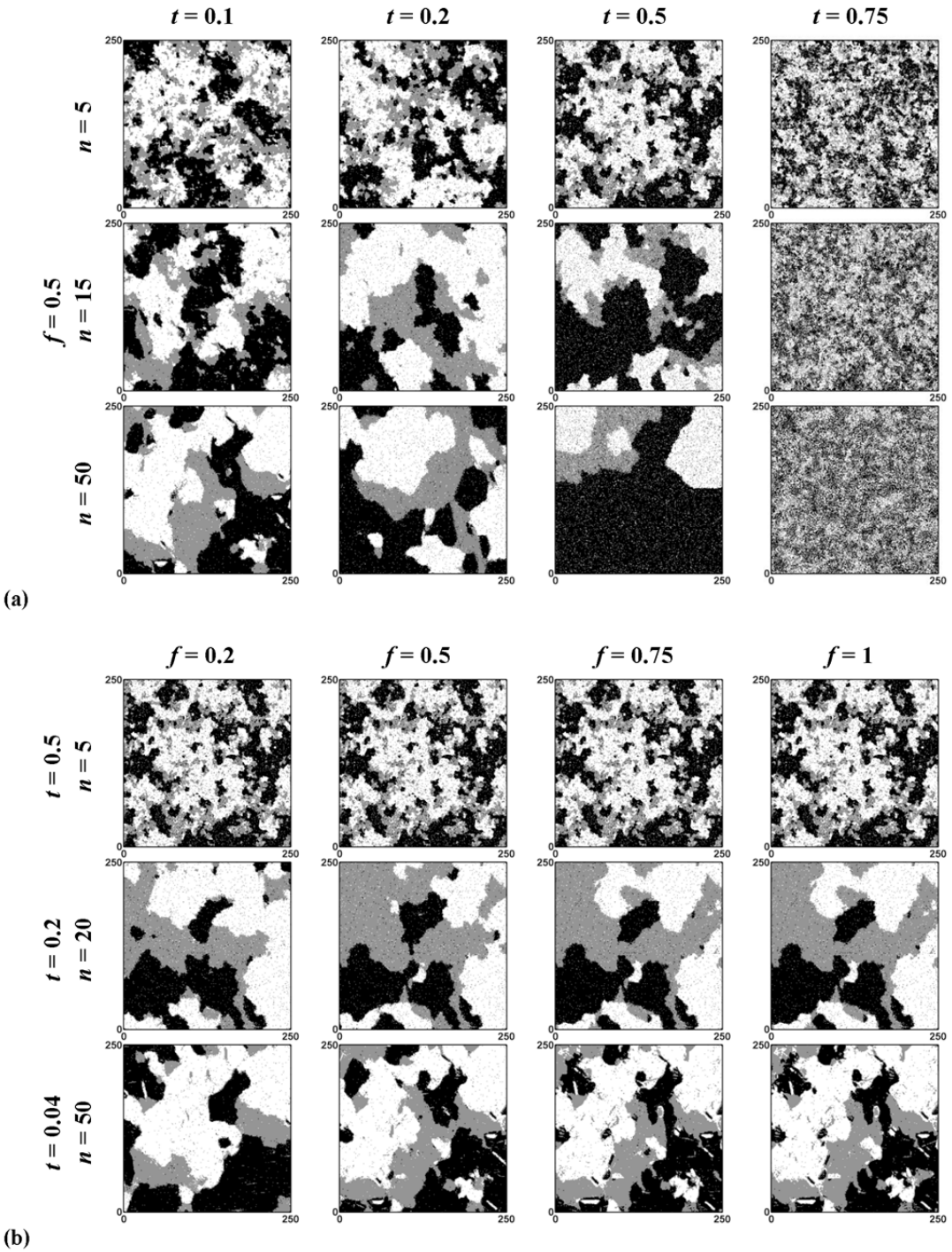


Figure B (a) First out of 10 unconditional simulations illustrating the effect of parameters *t* and *n* with *f* = 0.5 and (b) first out of 10 unconditional simulations illustrating the effect of *f* for constant *t* and *n* based on the categorical marble TI (Figure 5.3d).

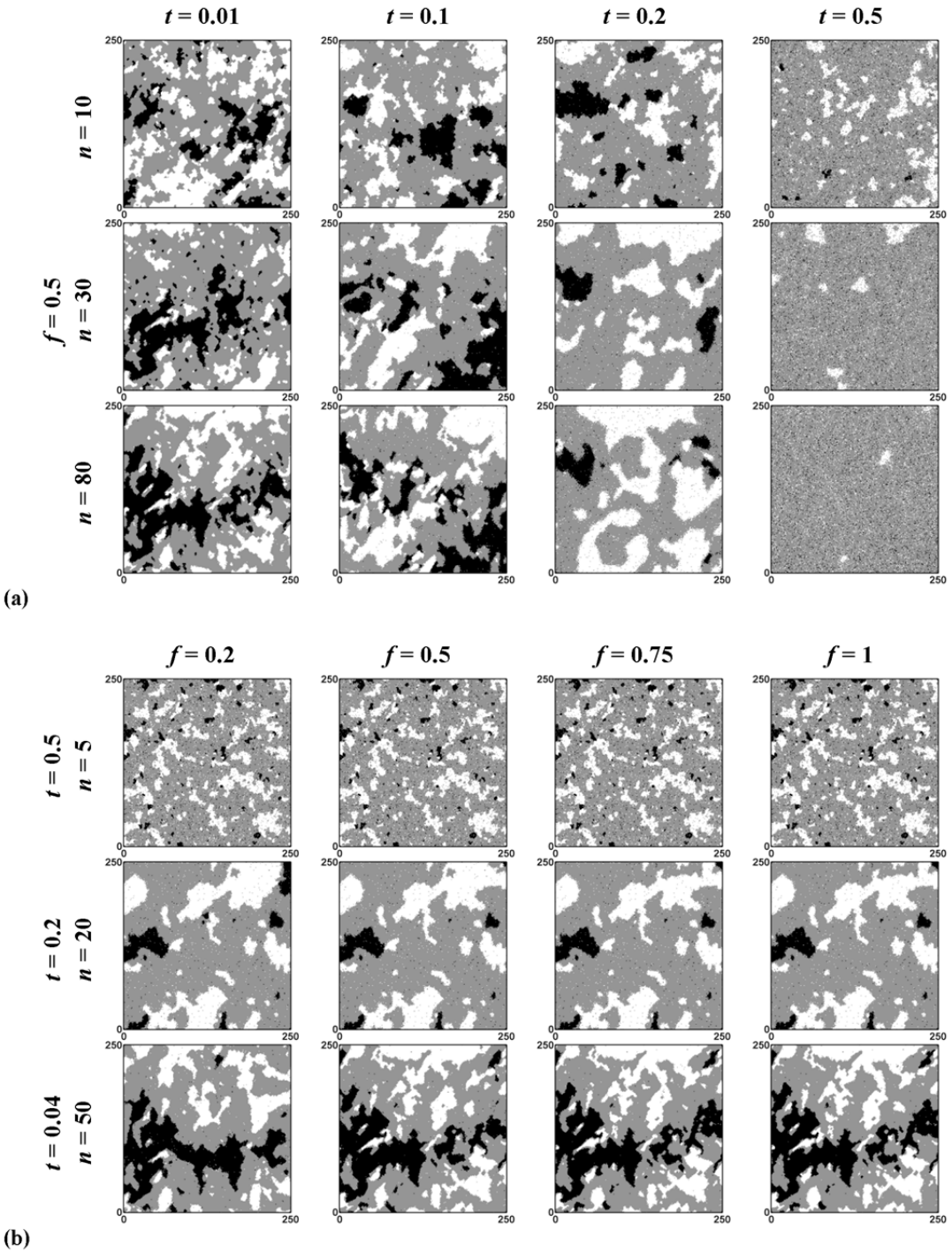


Figure C (a) First out of 10 unconditional simulations illustrating the effect of parameters *t* and *n* with *f* = 0.5 and (b) first out of 10 unconditional simulations illustrating the effect of *f* for constant *t* and *n* based on the categorical snow crystals TI (Figure 5.3f).

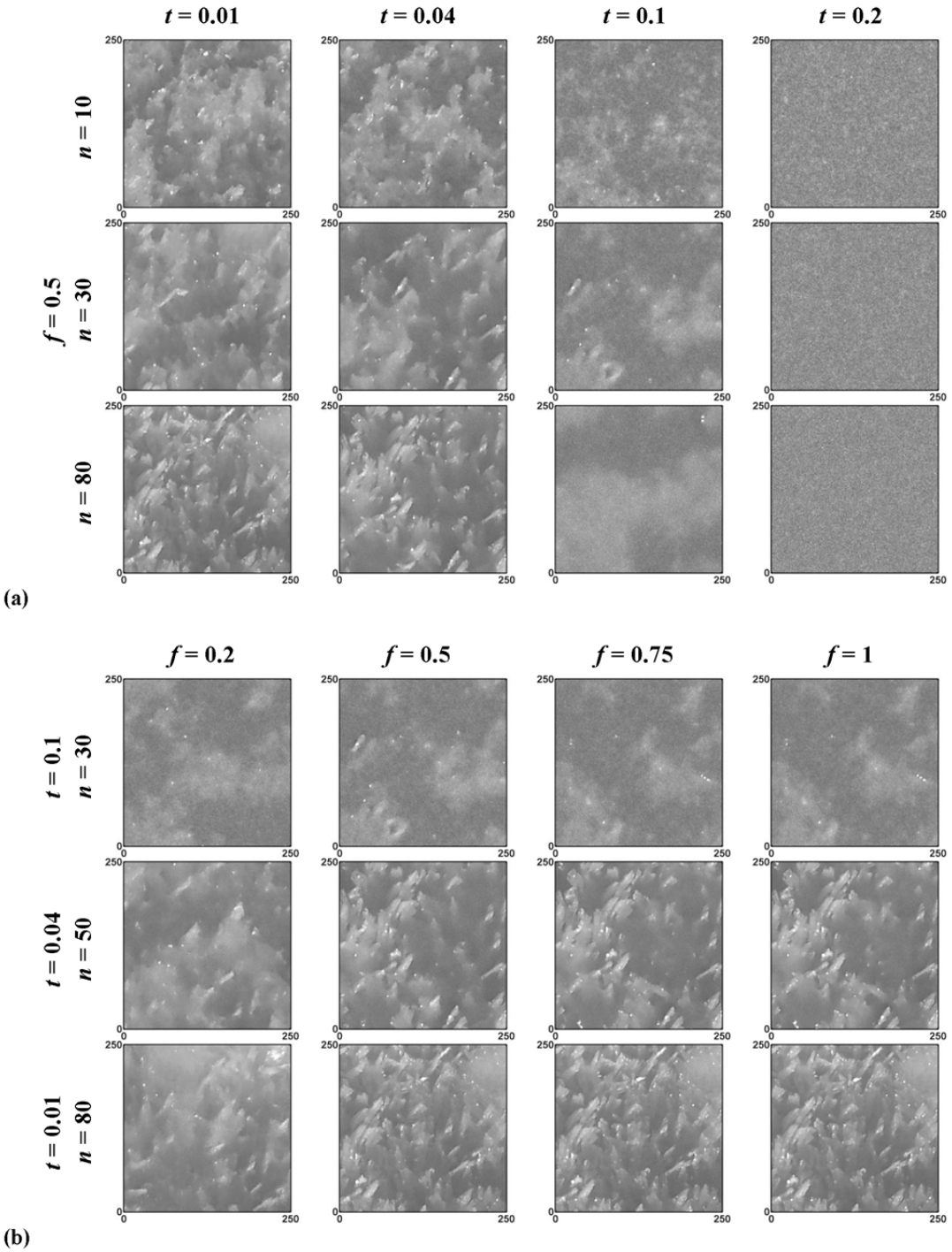


Figure D (a) First out of 10 unconditional simulations illustrating the effect of parameters *t* and *n* with *f* = 0.5 and (b) first out of 10 unconditional simulations illustrating the effect of *f* for constant *t* and *n* based on the continuous snow crystals TI (Figure 5.3e).

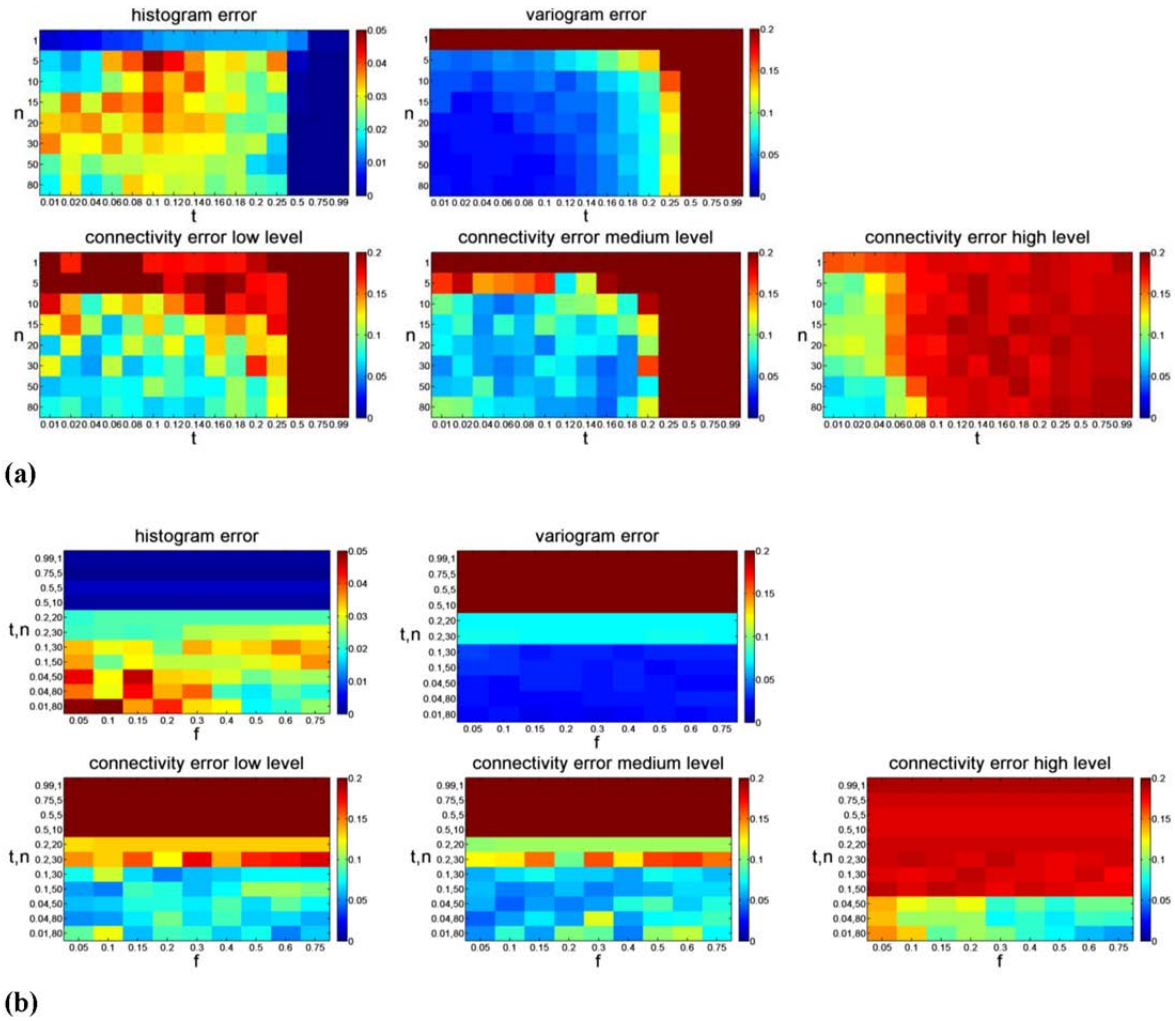
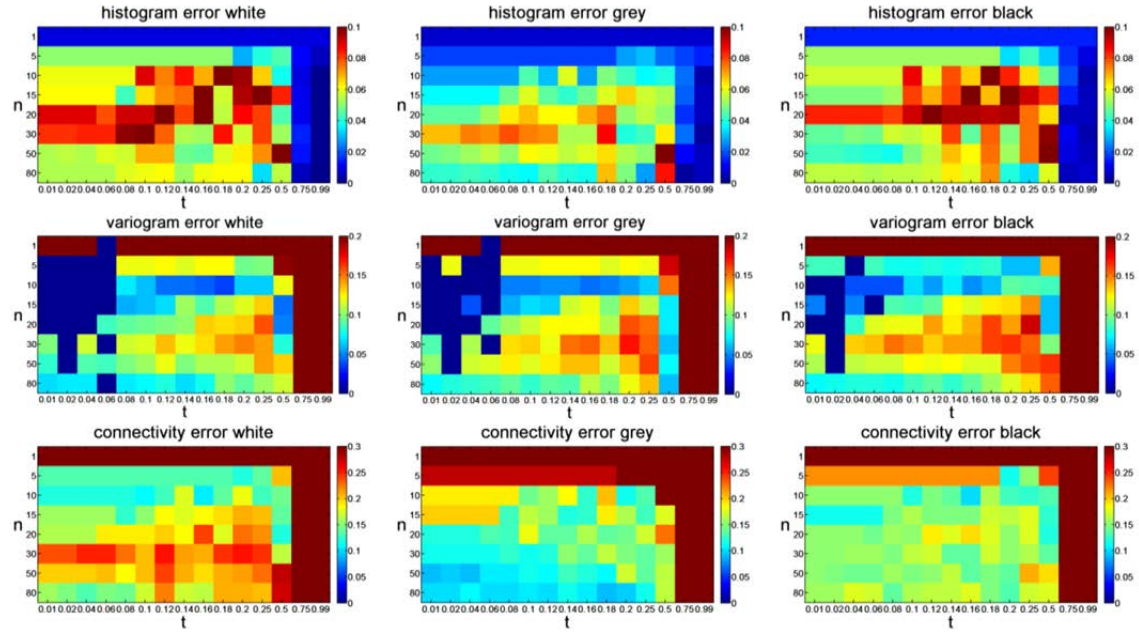
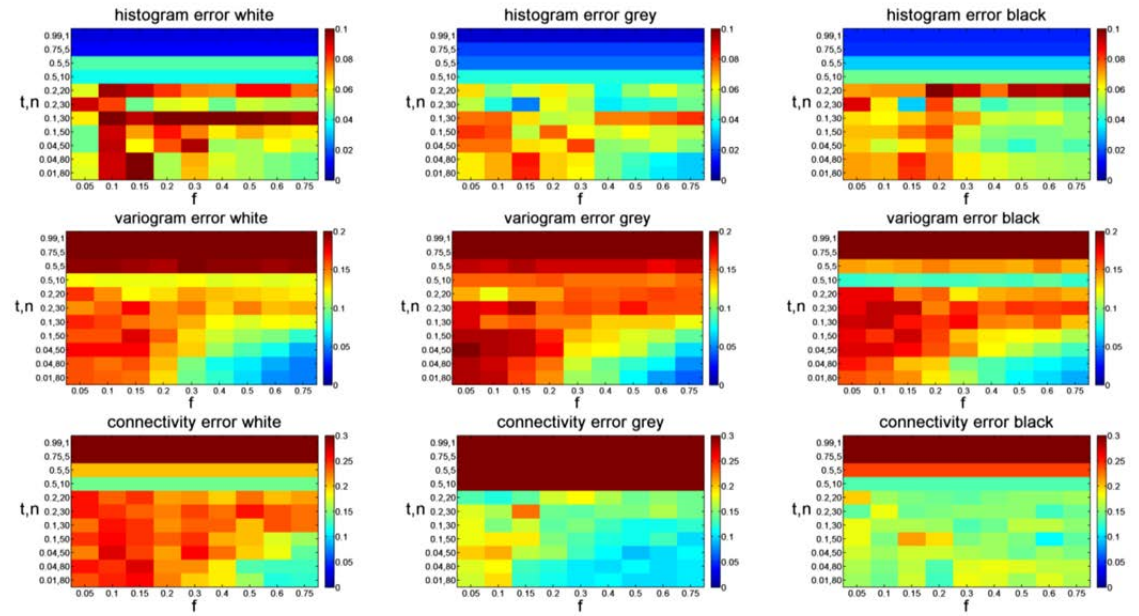


Figure E Influence of (a) *t* and *n* (for *f* = 0.5) and (b) *f* on the quality indicators based on the continuous ice-wedge TI (Figure 5.3a).

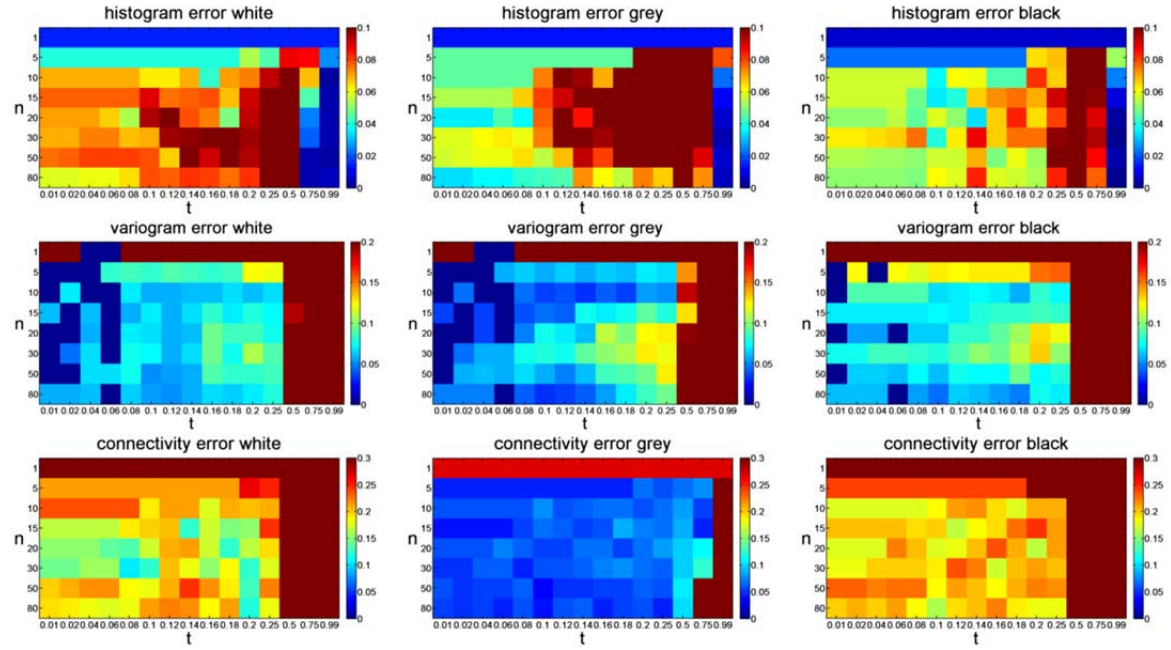


(a)

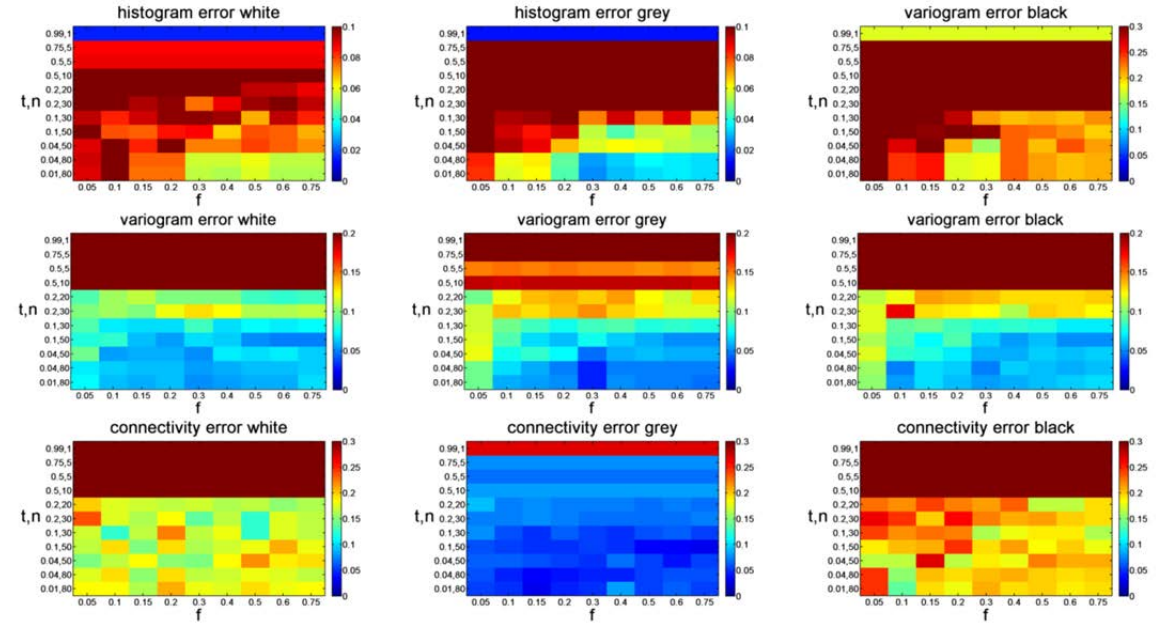


(b)

Figure F Influence of (a) *t* and *n* (for *f* = 0.5) and (b) *f* on the quality indicators based on the categorical marble TI (Figure 5.3d).



(a)



(b)

Figure G Influence of (a) *t* and *n* (for *f* = 0.5) and (b) *f* on the quality indicators based on the categorical snow crystals TI (Figure 5.3f).

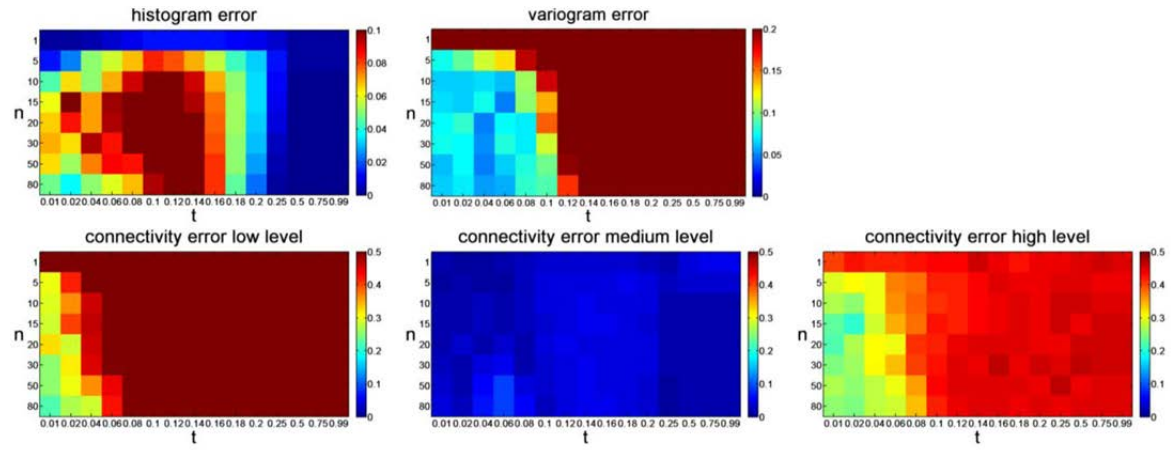


Figure H Influence of (a) *t* and *n* (for *f* = 0.5) and (b) *f* on the quality indicators based on the continuous snow crystals TI (Figure 5.3e).

References

Adamchuk, V.I., Hummel, J.W., Morgan, M.T. and Upadhyaya, S.K. 2004. On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture* 44, 71–91.

Arpat, G.B. and Caers, J. 2007. Conditional simulation with patterns. *Mathematical Geology* 39, 177–203.

Baer, J.U., Kent, T.F. and Anderson, S.H. 2009. Image analysis and fractal geometry to characterize soil desiccation cracks. *Geoderma* 154, 153–163.

Bastante, F.G., Ordóñez, C., Taboada, J. and Matias, J.M. 2008. Comparison of indicator kriging, conditional indicator simulation and multiple-point statistics used to model slate deposits. *Engineering Geology* 98, 50–59.

Bastante, F.G., Taboada, J., Alejano, L.R. and Ordóñez, C. 2005. Evaluation of the resources of a slate deposit using indicator kriging. *Engineering Geology* 81, 407–418.

Bleinès, C., Bourges, M., Deraisme, J., Geffroy, F., Jeannée, N., Lemarchand, O., Perseval, S., Poisson, J., Rambert, F., Renard, D., Touffait, Y. and Wagner, L. 2011. *ISATIS 2011 Case studies*. Geovariances, Avon Cedex.

Boisvert, J.B., Pyrcz M.J. and Deutsch C.V. 2007. Multiple-point statistics for training image selection. *Natural Resources Research* 16, 313–321.

Burgess, T.M. and Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. the semivariogram and punctual kriging. *Journal of Soil Science* 31, 315–331.

Buylaert, J.P., Ghysels, G. Murray A.S., Thomsen, K.J., Vandenberghe, D., De Corte, F., Heyse, I. and Van den Haute, P. 2009. Optical dating of relict sand wedges and composite-wedge pseudomorphs in Flanders, Belgium. *Boreas* 38, 160–175.

Caers, J. and Zhang, T. 2004. Multiple-point geostatistics: a quantitative vehicle for integrating geologic analogs into multiple reservoir models. Integration of outcrop and modern analog data in reservoir models, *AAPG memoir* 80, 383–394.

Catt, J.A. 1979. Soils and Quaternary Geology in Britain. *Journal of Soil Science* 30, 607–642.

Chilès, J.P. and Guillen, A. 1984. Variogrammes et krigeages pour la gravimétrie et le magnétisme. *Sciences de la Terre – Série Informatique Géologique* 20, 455–468.

Cockx, L., Ghysels, G., Van Meirvenne, M. and Heyse, I. 2006. Prospecting ice-wedge pseudomorphs and their polygonal network using the electromagnetic induction sensor EM38DD. *Permafrost and Periglacial Processes* 17, 163–168.

Cockx, L., Van Meirvenne, M. and De Vos, B. 2007. Using the EM38DD soil sensor to delineate clay lenses in a sandy forest soil. *Soil Science Society of America Journal* 71, 1314–1322.

Cockx, L., Van Meirvenne, M., Vitharana, U.W.A., Verbeke, L.P.C., Simpson, D., Saey, T. and Van Coillie, F.A.B. 2009. Extracting Topsoil Information from EM38DD

Sensor Data using a Neural Network Approach. *Soil Science Society of America Journal* 73, 2051–2058.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements* 20, 37–46.

Comunian, A., Renard, P. and Straubhaar, J. 2012. 3D multiple-point statistics simulation using 2D training images. *Computers & Geosciences* 40, 49–65.

Comunian, A., Renard, P., Straubhaar, J. and Bayer, P. 2011. Three-dimensional high resolution fluvio-glacial aquifer analog - Part 2: Geostatistical modeling. *Journal of Hydrology* 405, 10–23.

Corwin, D.L. and Lesch, S.M. 2005. Characterizing soil spatial variability with apparent electrical conductivity: I. Survey protocols. *Computers and Electronics in Agriculture* 46, 103–133.

Cressie N. 1985. Fitting variogram models by weighted least squares. *Mathematical Geology* 17, 563–586.

Cresto Aleina, F., Brovkin, V., Muster, S., Boike, J., Kutzbach, L. and Zuyev, S. 2012. Poisson-Voronoi diagrams and the polygonal tundra. *Geophysical Research Abstracts* 14, EGU2012-1963-1.

Dansart, A.M., Bahr, J.M. and Atig, J.W. 1999. Using ground-penetrating radar to map fossil permafrost wedges that are preferential flow paths for leaching to groundwater. Geological Society of America, *Abstracts with Program* 31, A–76.

Dathe, A., Eins, S., Niemeyer, J. and Gerold, G. 2001. The surface fractal dimension of the soil-pore interface as measured by image analysis. *Geoderma* 103, 203–229.

de Gruijter, J.J. and McBratney, A.B. 1988. A modified fuzzy k-means method for predictive classification, In: Bock, H.H. (Ed.), *Classification and Related Methods of Data Analysis*, Elsevier, North Holland, pp. 97–104.

de Gruijter, J.J., McBratney, A.B. and Taylor, J. 2010. Sampling for High-Resolution Soil Mapping. Chapter 1, In: Rossel, V.R.A., McBratney, A.B., Minasny, B. (Eds.), *Proximal Soil Sensing. Progress in Soil Science 1*. Springer Science+Business Media B.V., pp. 3–14.

De Iaco, S. and Maggio, S. 2011. Validation techniques for geological patterns simulations based on variogram and multiple point statistics. *Mathematical Geoscience* 43, 483–500.

dell'Arciprete, D., Bersezio, R., Felleti, F., Giudici, M., Comunian, A. and Renard, P. 2012. Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments. *Hydrogeology Journal* 20, 299–311.

De Smedt, P., Van Meirvenne, M., Meerschman, E., Saey, T., Bats, M., Court-Picon, M., De Reu, J., Zwertvaegher, A., Antrop, M., Bourgeois, J., De Maeyer, P., Finke, P.A., Verniers, J. and Crombe, P. 2011. Reconstructing palaeochannel morphology with a mobile multicoil electromagnetic induction sensor. *Geomorphology* 130, 136–141.

Deutsch, C.V. and Journel, A.G. 1997. *GSLIB: Geostatistical Software Library and User's Guide*. 2nd Edition. Oxford University Press, New York.

Deutsch, C.V. and Wang, L. 1996. Hierarchical object based stochastic modeling of fluvial reservoirs. *Mathematical Geology* 28, 857–880.

Doolittle, J. and Nelson F. 2009. Characterising Relict Cryogenic Macrostructures in Mid-Latitude Areas of the USA with Three-Dimensional Ground-Penetrating Radar. *Permafrost and Periglacical Processes* 20, 257–268.

Dudewicz, E.J. and Mishra, S.N. 1988. *Modern mathematical statistics*. John Wiley & Sons, New York.

Dutilleul, P., Haltigin, T.W and Pollard, W.H. 2009. Analysis of polygonal terrain landforms on Earth and Mars through spatial point patterns. *Environmetrics* 20, 206–220.

Emery, X. 2004. Properties and limitations of sequential indicator simulation. *Stochastic Environmental Reseach and Risk Assessment* 18, 414–424.

Emery, X. and Ortiz, J.M. 2011. A comparison of random field models beyond bivariate distributions. *Mathematical Geosciences* 43, 183–202.

French, H.M. 2007. *The Periglacial Environment, 3rd Edition*. John Wiley and Sons, Chichester.

French, H.M., Demitroff, M. and Forman, S.L. 2003. Evidence for Late-Pleistocene Permafrost in the New Jersey Pine Barrens (Latitude 391N), Eastern USA. *Permafrost and Periglacial Processes* 14, 259–274.

Ghysels, G. 2008. Bijdrage tot de studie van de kenmerken, de genese en de datering van periglaciale polygonale wigstructuren in België. PhD dissertation, Ghent University, Ghent, Belgium (in Dutch).

Ghysels, G. and Heyse, I. 2006. Composite-wedge pseudomorphs in Flanders, Belgium. *Permafrost and Periglacial Processes* 17, 145–161.

Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Pres, New York.

Gringarten, E. and Deutsch, C.V. 2001. Variogram interpretation and modeling. *Mathematical Geology* 33, 507–534.

Guardiano, F.B. and Srivastava, R.M. 1993. Multivariate geostatistics: beyond bivariate moments, In: Soares, A. (Ed.), *Geostatistics-Troia, Vol. 1*. Kluwer Academic Publishers, Dordrecht, pp. 133–144.

Harding, S.A., Murray, D.A. and Webster, R. 2010. MVARIOGRAM procedure, In: Payne, R.W. (Ed.), *GenStat Release 13 Reference Manual, Part 3 Procedure Library PL21*. VSN International, Hemel Hempstead.

Harry, D.G. and Gozdzik, J.S. 1988. Ice wedges: Growth, thaw transformation, and palaeoenvironmental significance. *Journal of Quaternary Science* 3, 39–55.

Heuvelink, G.B.M. and Webster, R. 2001. Modelling soil variation: past, present, and future. *Geoderma* 100, 269–301.

Hu, L.Y. and Chugunova, T. 2008. Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review. *Water Resources Research* 44, W11413.

Huysmans, M. and Dassargues, A. 2009. Application of multiple-point geostatistics on modelling groundwater flow and transport in a cross-bedded aquifer (Belgium). *Hydrogeology Journal* 17, 1901–1911.

Huysmans, M. and Dassargues, A. 2011. Direct multiple-point geostatistical simulation of edge properties for modelling thin irregularly shaped surfaces. *Mathematical Geosciences* 43, 521–536.

Isaaks, E.H. and Srivastava, R.M. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, New York.

Islam, M.M., Saey, T., Meerschman, E., De Smedt, P., Meeuws, F., Van De Vijver, E. and Van Meirvenne, M. 2011. Delineating water management zones in a paddy rice field using a Floating Soil Sensing System. *Agricultural Water Management* 102, 8–12.

Jones, A., Stolbovoy, V., Tarnocai, C., Broll, G., Spaargaren, O. and Montanarella, L. 2010. *Soil Atlas of the Northern Circumpolar Region*. Publication Office of the European Union, Luxembourgh.

Journel, A.G. and Huijbregts, C.J. 1978. *Mining Geostatistics*. Academic Press, New York.

Journel, A. and Zhang, T. 2006. The necessity of a multiple-point prior model. *Mathematical Geology* 38, 591–610.

Jung, A. and Aigner, T. 2012. Carbonate geobodies: hierarchical classification and database – a new workflow for 3D reservoir modelling. *Journal of Petroleum Geology* 35, 49–65.

Kachanoski, R.G., Hendrickx, J.M.H. and de Jong, E. 2002. Electromagnetic induction. In: Dane, J.H., Topp, G.C (Eds.), *Methods of Soil Analysis, Part 1, Physical Methods, Third Edition*. Soil Science Society of America, pp.497–501.

Kolstrup, E. 1986. Reappraisal of the upper Weichselian periglacial environment form Danish frost wedge casts. *Palaeogeography, Palaeoclimatolgoy, Palaeoecology* 56, 237–249.

Kolvos, A., Christakos, G., Hristopulos, D.T. and Serre, M.L. 2004. Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. *Advances in Water Resources* 27, 815–830.

Krige, D.G. 1951. A statistical approach to some mine valuations and allied problems at the Witwatersrand. Master Thesis, University of Witwatersrand, Johannesburg, South Africa.

Kullback, S. and Leibler, R.A. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.

Lark, R.M. 2009. A stochastic-geometric model of soil variation. *European Journal of Soil Science* 60, 706–719.

Lark, R.M. 2012a. Towards soil geostatistics. *Spatial Statistics* 1, 92–99.

Lark, R.M. 2012b. A stochastic geometric model for continuous local trends in soil variation. *Geoderma* 189–190, 661–670.

Le Coz, M., Genthon, P. and Adler, P.M. 2011. Multiple-point statistics for modeling facies heterogeneities in a porous medium: the Komadugu-Yobe alluvium, Lake Chad Basin. *Mathematical Geosciences* 43, 861–878.

Lefebvre, S. and Hoppe, H. 2006. Appearance-space texture synthesis. *ACM Transactions on Graphics 25*, 541-548.

Liu, Y.H. 2006. Using the Snesim program for multiple-point statistical simulation. *Computers & Geosciences* 32, 1544-1563.

Lusch, D.P., Stanley, K.E., Schaetzl, R.J., Kendall, A.D., Van Dam, R.L., Nielsen, A., Blumer, B.E. Hobbs, T.C., Archer, J.K., Holmstadt J.L.F. and May, C.L. 2009. Characterization and Mapping of Patterned Ground in the Saginaw Lowlands, Michigan: Possible Evidence for Late-Wisconsin Permafrost. *Annals of the Association of American Geographers* 99, 445–466.

Mackay, J.R. and Burn, C.R. 2002. The first 20 years (1978-1979 to 1998-1999) of ice-wedge growth at Illisarvik experimental drained lake site, western Arctic coast, Canada. *Canadian Journal of Earth Sciences* 39, 95–11.

Mariethoz, G. 2009. Geological stochastic imaging for aquifer characterization. PhD Dissertation, University of Neuchâtel, Neuchâtel, Switzerland.

Mariethoz, G. 2010. A general parallelization strategy for random path based geostatistical simulation methods. *Computers & Geosciences* 36, 953–958.

Mariethoz, G. and Kelly, B.F.J. 2011. Modeling complex geological structures with elementary training images and transform-invariant distances. *Water Resources Research* 47, W07527.

Mariethoz, G., McCabe, M. and Renard, P. 2012. Spatiotemporal reconstruction of gaps in multivariate fields using the Direct Sampling approach. *Water Resources Research* 48, W10507.

Mariethoz, G. and Renard, P. 2010. Reconstruction of incomplete data sets or images using Direct Sampling. *Mathematical Geosciences* 42, 245–268.

Mariethoz, G., Renard, P. and Straubhaar, J. 2010. The Direct Sampling method to perform multiple-point geostatistical simulations. *Water Resources Research* 46, W11536.

Matheron, G. 1962. *Traité de Géostatistique Appliqué*, Tome 1. Memoires du Bureau de Recherches Géologiques et Minières, Paris.

Matheron, G. 1965. *Les variables régionalisées et leur estimation*. Masson, Paris.

McBratney, A.B. and Moore, A.W. 1985. Application of fuzzy sets to climatic classification. *Agricultural and Forest Meteorology* 35, 165–185.

McBratney, A.B. and Odeh, I.O.A. 1997. Applications of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma* 77, 85–113.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S. and Shatar, T.M. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma* 97, 293–327.

McBratney, A.B. and Webster, R. 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science* 37, 617–639.

McNeill, J.D. 1980. *Electromagnetic terrain conductivity measurement at low induction numbers*. Geonics Ltd, Ontario.

Meerschman, E., Cockx, L. and Van Meirvenne, M. 2011. A geostatistical two-phase sampling strategy to map soil heavy metal concentrations in a former war zone. *European Journal of Soil Science* 62, 408–416.

Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M. and Renard, P. 2013. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. *Computers & Geosciences* 52, 307–324.

Minasny, B. and McBratney, A.B. 2002. *FuzMe version 3*. Australian Centre for Precision Agriculture, The University of Sidney, New South Wales.

Mirowski, P.W., Tetzlaff, D.M., Davies, R.C., McCormick, D.S., Williams, N. and Signer, C. 2009. Stationarity scores on training images for multipoint geostatistics. *Mathematical Geosciences* 41, 447–474.

Morgan, A.V.M. 1971. Engineering problems caused by fossil permafrost features in the English Midlands. *Quaterly Journal of Engineering Geology & Hydrogeology* 4, 111–114.

Murton, J.B. and French, H.M. 1993. Thaw modification of frost-fissure wedges, Richards Island, Pleistocene Mackenzie Delta, western Arctic Canada. *Journal of Quaternary Science* 8, 185–196.

Nyquist, H. 1928. Certain topics in telegraph transmission theory. *Trans. AIEE* 47, 617-611.

Okabe, H. and Blunt, M.J. 2004. Prediction and permeability for porous media reconstructed using multiple-point statistics. *Physical Review* E, 70.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S.K. 2000. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. 2nd Edition*. John Wiley & Sons, Chichester.

Payne, R.W., Harding, S.A., Murray, D.A., Soutar, D.M., Baird, D.B., Glaser, A.I., Channing, I.C., Welham, S.J., Gilmour, A.R., Thompson, R. and Webster, R. 2009. *Genstat Release 12 Reference Manual, Part 2 Directives*. VSN International, Hemel Hempstead.

Plug, L.J. and Werner, B.T. 2002. Nonlinear dynamics of ice-wedge networks and resulting sensitivity to severe cooling events. *Nature* 417, 929–933.

Plug, L.J. and Werner, B.T. 2008. Modelling of ice-wedge networks. *Permafrost and Periglacial Processes* 19, 63–69.

Pontius, R.G.Jr. and Schneider, L.C. 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment* 85, 239–248.

Pyrcz, M.J., Boisvert, J.B. and Deutsch, C.V. 2008. A library of training images for fluvial and deepwater reservoirs and associated code. *Computers & Geosciences* 43, 542–560.

Remy, N., Boucher, A. and Wu, J. 2009. *Applied Geostatistics with SGeMS: A User's Guide*. Cambridge University Press, New York.

Renard, P. and Allard, D. 2012. Connectivity metrics for subsurface flow and transport. *Advances in Water Resources* 51, 168–196.

Renard, P., Straubhaar, J., Caers, J. and Mariethoz, G. 2011. Conditioning facies simulations with connectivity data. *Mathematical Geosciences* 43, 879–903.

Reynolds, J.M. 1997. An Introduction to Applied and Environmental Geophysics. Wiley & Sons, New York.

Romanovskij, N.N. 1973. Regularities in formation of frost-fissure polygons and development of frost-fissure polygons. *Biuletyn Peryglacjalny* 23, 237–277.

Ronayne, M.J., Gorelick, S.M. and Caers, J. 2008. Identifying discrete geologic structures that produce anomalous hydraulic response: An inverse modeling approach. *Water Resources Research* 44, W08426.

Saey, T., Islam, M.M., De Smedt, P., Meerschman, E., Van De Vijver, E., Lehouck, A. and Van Meirvenne, M. 2012. Using a multi-receiver survey of apparent electrical conductivity to reconstruct a Holocene tidal channel in a polder area. *Catena* 95, 104–111.

Saey, T., Simpson, D., Vermeersch, H., Cockx, L. and Van Meirvenne, M. 2009. Comparing the EM38DD and DUALEM-21S sensors for depth-to-clay mapping. *Soil Science Society of America Journal* 73, 7–12.

Saey, T., Simpson, D., Vitharana, U.W.A., Vermeersch, H., Vermang, J. and Van Meirvenne, M. 2008. Reconstructing the paleotopography beneath the loess cover with the aid of an electromagnetic induction sensor. *Catena* 74, 58–64.

Saey, T., Van Meirvenne, M., Vermeersch, H., Ameloot, N. and Cockx, L. 2009. A pedotransfer function to evaluate the soil profile textural heterogeneity using proximally sensed apparent electrical conductivity. *Geoderma* 150, 389–395.

Seifert, D. and Jensen, J.L. 2000. Object and pixel-based reservoir modeling of a braided fluvial reservoir. *Mathematical Geology* 32, 581–603.

Simpson, D., Van Meirvenne, M., Saey, T., Vermeersch, H., Bourgeois, J., Lehouck, A., Cockx, L. and Vitharana, U.W.A. 2009. Evaluating the Multiple Coil Configurations of the EM38DD and DUALEM-21S Sensors to Detect Archaeological Anomalies. *Archaeological Prospection* 16, 91–102.

Slavich, P.G. and Petterson, G.H. 1990. Estimating average rootzone salinity for electromagnetic (EM-38) measurements. *Australian Journal of Soil Research* 31, 2401–2409.

Straubhaar, J., Renard, P., Mariethoz, G., Froidevaux, R. and Besson, O. 2011. An improved parallel multiple-point algorithm using a list approach. *Mathematical Geosciences* 43, 305-328.

Strebelle, S. 2002. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology* 34, 1–21.

Strebelle, S., Payrazyan, K. and Caers, J. 2003. Modeling of a deepwater turbidite reservoir conditional to seismic data using principal component analysis and multiple-point geostatistics. *SPE Journal* 8, 227–235.

Tobbler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46, 234–240.

Triantafilis, J., Odeh, I.O.A. and McBratney, A.B. 2001. Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton. *Soil Science Society of America Journal* 65, 869–878.

Truong, P.N., Heuvelink, G.B.M. and Gosling, J.P. 2012. Web-based tool for expert elicitation of the variogram. *Computers & Geosciences* 51, 390-399.

Vandenberghe, J. and Pissart, A. 1993. Permafrost changes in Europe during the Last Glacial. *Permafrost and Periglacial Processes* 4, 121–135.

Van Meirvenne, M. and Goovaerts, P. 2001. Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. *Geoderma* 102, 75–100.

Venables, W.N. and Ripley, B.D. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, New York.

Vitharana, U.W.A., Saey, T., Cockx, L., Simpson, D., Vermeersch, H. and Van Meirvenne, M. 2008a. Upgrading a 1/20,000 soil map with an apparent electrical conductivity survey. *Geoderma* 148, 107–112.

Vitharana, U.W.A., Van Meirvenne, M., Simpson, D., Cockx, L. and De Baerdemaeker, J. 2008b. Key soil and topographic properties to delineate potential management classes for precision agriculture in the European loess area. *Geoderma* 143, 206–215.

Walters, J.C. 1994. Ice-wedge casts and relict polygonal patterned ground in North-East Iowa, USA. *Permafrost and Periglacial Processes* 5, 269–281.

Webster, R. and Oliver, M.A.. 2007. *Geostatistics for environmental scientists. 2nd edition.* John Wiley & Sons, Chichester.

Whittle, P. 1954. On stationary processes in the plane. *Biometrika* 41, 434–449.

Whittle, P. 1962. Topographic correlations, power-law covariance functions and diffusion. *Biometrika* 49, 305–314.

Zhang, T., Bombarde, S., Strebelle, S. and Oatney, E. 2006a. 3D porosity modeling of a carbonate reservoir using continuous multiple-point statistics simulation. *SPE Journal* 11, 375–379.

Zhang, T, Switzer, P. and Journel, A. 2006b. Filter-based classification of training image patterns for spatial simulation. *Mathematical Geology* 38, 63–80.

Zhang, J., Zhou, K., Velho, L., Guo, B. and Shum, H.Y. 2003. Synthesis of progressively-variant textures on arbitrary surfaces. *ACM Transactions on Graphics* 22, 295–302.

Curriculum vitae

Personal data

Name: Eef Meerschman
Address: Hippodroomstraat 41, 8790 Waregem
E-mail: eefmeerschman@hotmail.com
Phone: +32 499 294647

Date of birth: 22 December 1986
Place of birth: Kortrijk, Belgium
Nationality: Belgian

Education

2007-2009: Master of Bioscience Engineering: Environmental Technology
Ghent University
Thesis: Geostatistical inventory of heavy metal concentrations in the soil around Ypres as a consequence of World War I
Promoter: Prof. dr. ir. Marc Van Meirvenne

2004-2007: Bachelor of Bioscience Engineering
Ghent University

Professional experience

2009-2013: Doctoral researcher
PhD fellowship of the Fund for Scientific Research-Flanders (FWO-Vlaanderen).
Research Group Soil Spatial Inventory Techniques
Department of Soil Management
Ghent University
Promoter: Prof. dr. ir. Marc Van Meirvenne

Scientific publications

International publications with peer review and in the Science Citation Index (A1)

22. De Smedt, P., Van Meirvenne, M., Herremans, D., De Reu, J., Saey, T., **Meerschman, E.,** Crombé, P. and De Clercq, W. 2013. The 3-D reconstruction of medieval wetland reclamation through electromagnetic induction survey. *Scientific Reports* 3, 1517.

21. Lark, R.M., **Meerschman, E.** and Van Meirvenne, M. 2013. A stochastic geometric model of the variability of soil formed in Pleistocene patterned ground. Submitted for publication in *Geoderma*.

20. Islam, M.M., **Meerschman, E.**, Saey, T., De Smedt, P., Van De Vijver, E. and Van Meirvenne, M. 2013. Delineating and evaluating variably puddled zones in a paddy rice field using electromagnetic induction based soil sensing. Submitted for publication in *Computers and Electronics in Agriculture*.

19. **Meerschman, E.,** Van Meirvenne, M., Mariethoz, G., Islam, M.M., De Smedt, P., Van De Vijver, E. and Saey, T. 2013. Using bivariate multiple-point statistics and proximal soil sensor data to map fossil ice-wedge polygons. *Geoderma*, *in press* (DOI: 10/1016/j.geoderma.2013.01.016).

18. **Meerschman, E.,** Van Meirvenne, M., Van De Vijver, E., De Smedt, P., Islam, M.M. and Saey, T. 2013. Mapping complex soil patterns with multiple-point geostatistics. *European Journal of Soil Science* 64, 183–191.

17. Saey, T., De Smedt, P., De Clercq, W., **Meerschman, E.,** Islam, M.M. and Van Meirvenne, M. 2013. Identifying soil patterns at different spatial scales with a multi-receiver EMI sensor. *Soil Science Society of America Journal* 77, 382–390.

16. Van Meirvenne, M., Islam, M.M., De Smedt, P., **Meerschman, E.,** Van De Vijver E. and Saey, T. 2013. Key variables for the identification of soil management classes in the Aeolian landscapes of north-west Europe. *Geoderma* 199, 99–105.

15. De Smedt, P., Saey, T., Lehouck, A., Stichelbaut, B., **Meerschman, E.,** Islam, M.M., Van De Vijver, E. and Van Meirvenne, M. 2013. Exploring the potential of multi-receiver EMI survey for geoarchaeological prospection: a 90 ha dataset. *Geoderma* 199, 30–36.

14. **Meerschman, E.,** Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M. and Renard, P. 2013. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. *Computers & Geosciences* 52, 307–324.

13. De Smedt, P., Van Meirvenne, M., Davies, N., Bats, M., Saey, T., De Reu, J., **Meerschman, E.,** Gelorini, V., Zwertvaegher, A., Antrop, M., Bourgeois, J., De Maeyer, P., Finke, P.A., Verniers, J. and Crombé, P. 2012. A multidisciplinary approach to reconstructing Late Glacial and Early Holocene landscapes. *Journal of Archaeological Science* 40, 1260–1267.

12. Saey, T., De Smedt, P., Islam, M.M., **Meerschman, E.,** Van De Vijver, E., Lehouck, A. and Van Meirvenne, M. 2012. Depth slicing of multi-receiver EMI measurements to enhance the delineation of contrasting subsoil features. *Geoderma* 189–190, 514–521.

11. Saey, T., Islam, M.M., De Smedt, P., **Meerschman, E.,** Van De Vijver, E., Lehouck, A. and Van Meirvenne, M. 2012. Using a multi-receiver survey of apparent electrical conductivity to reconstruct a Holocene tidal channel in a polder area. *Catena*, 95, 104–111.

10. Islam, M.M., **Meerschman, E.,** Saey, T., De Smedt, P., Van De Vijver, E. and Van Meirvenne, M. 2011. Comparing apparent electrical conductivity measurements on a paddy field under flooded and drained conditions. 2011. *Precision Agriculture* 13, 384–392.

9. Saey, T., De Smedt, P., **Meerschman, E.,** Islam, M.M., Meeuws, F., Van de Vijver, E., Lehouck, A. and Van Meirvenne, M. 2012. Electrical conductivity depth modelling

with a multireceiver EMI sensor for prospecting archaeological features. *Archaeological Prospection* 19, 21–30.

8. Saey, T., Van Meirvenne, M., Dewilde, M., Wyffels, F., De Smedt, P., **Meerschman, E.,** Islam, M.M, Meeuws, F. and Cockx L. 2011. Combining multiple signals of an electromagnetic induction sensor to prospect land for metal objects. *Near Surface Geophysics* 9, 309–317.

7. Islam, M.M., Saey, T., **Meerschman, E.**, De Smedt, P., Meeuws, F., Van De Vijver, E. and Van Meirvenne, M. 2011. Delineating water management zones in a paddy rice field using a floating soil sensing system. *Agricultural Water Management* 102, 8–12.

6. Islam, M.M., Cockx, L., **Meerschman, E.,** De Smedt, P., Meeuws, F. and Van Meirvenne, M. 2010. A floating sensing system to evaluate soil and crop variability within flooded paddy rice fields. *Precision Agriculture* 12, 850–859.

5. **Meerschman, E.,** Van Meirvenne, M., De Smedt, P., Saey, T., Islam, M.M., Meeuws, F., Van De Vijver, E. and Ghysels, G. 2011. Imaging a polygonal network of ice-wedge casts with an electromagnetic induction sensor. *Soil Science Society of America Journal* 75, 2095–2100.

4. De Smedt, P., Van Meirvenne, M., **Meerschman, E.,** Saey, T., Bats, M., Court-Picon, M., De Reu, J., Werbrouck, I., Zwertvaegher, A., Antrop, M., Bourgeois, J., De Maeyer, P., Finke, P.A., Verniers, J. and Crombé, P. 2011. Reconstructing palaeochannel morphology with a mobile multi-coil electromagnetic induction sensor. *Geomorphology* 130, 136–141.

3. **Meerschman, E.,** Cockx, L. and Van Meirvenne, M. 2011. A geostatistical two-phase sampling strategy to map soil heavy metal concentrations in a former war zone. *European Journal of Soil Science* 62, 408–416.

2. **Meerschman, E.,** Cockx, L., Islam, M.M., Meeuws, F. and Van Meirvenne, M. 2011. Geostatistical assessment of the impact of World War I on the spatial occurrence of soil heavy metals. *AMBIO: A Journal of the Human Environment* 40, 417–424.

1. Saey, T., Van Meirvenne, M., De Smedt, P., Cockx, L., **Meerschman, E.,** Islam, M.M. and Meeuws, F. 2011. Mapping depth-to-clay using fitted multiple depth response curves of a proximal EMI sensor. *Geoderma* 162, 151–158.

Book chapters

1. Islam, M.M., **Meerschman, E.,** Cockx, L., De Smedt, P., Meeuws, F. and Van Meirvenne, M. 2011. Comparing apparent electrical conductivity measurements on a paddy field under flooded and drained conditions. In: Stafford, J.V. (Ed.), *Precision Agriculture 2011*. Czech Centre for Science and Society, Prague, Czech Republic. ISBN: 978-80-904830-5-7, p43–50.

Conference and workshop proceedings

8. Lark, R.M., **Meerschman, E.** and Van Meirvenne, M. 2013. A stochastic-geometric model of soil variation in Pleistocene patterned ground. In: *Geophysical Research Abstracts, Vol. 15, EGU2013-2234, EGU General Assembly 2013*, Vienna, Austria.

7. **Meerschman, E.,** Van De Vijver, E., Mariethoz, G. and Van Meirvenne, M. 2012. Using bivariate multiple-point statistics for the processing of proximal soil sensor data. In: Goméz-Hernandez (Ed.), *Proceedings of geoENV 2012: IX conference on Geostatistics for Environmental Applications*, Polytechnic University of Valencia, Valencia, Spain, pp. 219–220

6. **Meerschman, E.,** Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M., Renard, P. 2012. Guidelines to perform multiple-point statistical simulations with the Direct Sampling algorithm. In: *Expanded abstract collection from Ninth International Geostatistics Congress*, Oslo, Norway, abstract number P–029.

5. Pirot, G., **Meerschman, E.,** Mariethoz, G., Straubhaar, J., Van Meirvenne, M. and Renard, P. 2011. Optimizing Direct Sampling algorithm's parameters to performing multiple-points geostatistical simulations. In: *Proceedings of AGU Fall Meeting 2011*, San Francisco, California, USA, abstract number H53F–1475.

4. **Meerschman, E**. and Van Meirvenne, M. 2011. Using bivariate multiple-point geostatistics and proximal soil sensor data to map fossil ice-wedge polygons. In: Jakšík, O., Klement, A. and Borůvka, L. (Eds.), *Pedometrics 2011 – Innovations in Pedometrics*, the Czech University of Life Sciences, Prague, Czech Republic, pp. 51.

3. Islam, M. M., Van Meirvenne, M., Loonstra, E., **Meerschman, E.,** De Smedt, P., Meeuws, F., Van De Vijver, E. and Saey, T. 2011. Key properties for delineating soil management zones. In: Adamchuk, V.I. and Viscarra Rossel, R.A. (Eds.), *Proceeding of the second Global Workshop on Proximal Soil Sensing*, McGill University, Montreal, Canada, pp. 52–55.

2. **Meerschman, E**. and Van Meirvenne, M. 2010. Regional characterization of soil heavy metals in a former World War I battle area. In: Cockx, L., Van Meirvenne, M., Bogaert, P. and D'Or, D. (Eds.), *Book of abstracts of geoENV 2010: 8th International Conference on Geostatistics for Environmental Applications*, Ghent University, Ghent, Belgium, pp. 195–197.

1. **Meerschman, E**. and Van Meirvenne, M. 2010. Application of multiple-point geostatistics in soil science: the reconstruction of polygonal networks of ice-wedge pseudomorphs. In: Cockx, L., Van Meirvenne, M., Bogaert, P. and D'Or, D. (Eds.), *Book of abstracts of geoENV 2010: 8th International Conference on Geostatistics for Environmental Applications*, Ghent University, Ghent, Belgium, pp. 28–31.