

Updating soil information with digital soil mapping

Bas Kempen

Thesis committee**Thesis supervisor**

Prof. dr. ir. A. Veldkamp
Professor of Land Dynamics
Wageningen University

Thesis co-supervisors

Dr. ir. G.B.M. Heuvelink
Associate professor, Land Dynamics group
Wageningen University

Dr. D.J. Brus
Senior researcher, Team Soil Geography
Alterra, WUR

Other members

Prof. dr. ir. A.K. Bregt, Wageningen University
Prof. dr. R.M. Lark, British Geological Survey, Nottingham, United Kingdom
Dr. P.A. Finke, Ghent University, Belgium
Dr. D.G. Rossiter, University of Twente

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC).

Updating soil information with digital soil mapping

Bas Kempen

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 December 2011
at 4 p.m. in the Aula.

B. Kempen

Updating soil information with digital soil mapping, 218 pages.

Thesis Wageningen University, The Netherlands (2011)

With references, with summaries in Dutch and English

ISBN 978-94-6173-090-9

Table of Contents

1	Introduction	11
2	Updating the 1:50 000 Dutch soil map using legacy soil data: a multinomial logistic regression approach	25
3	Pedometric mapping of soil organic matter using a soil map with quantified uncertainty	57
4	Three-dimensional mapping of soil organic matter content using soil type-specific depth functions	81
5	Digital soil type mapping with the generalized linear geostatistical model	113
6	Efficiency comparison of conventional and digital soil mapping for updating soil maps of a cultivated peatland	139
7	Discussion and conclusions	175
	References	195
	Summary	205
	Samenvatting	209
	About the author	215



Dankwoord

Op het moment dat ik deze woorden schrijf is het exact vijf jaar geleden dat ik als promovendus begon bij de toenmalige leerstoel Bodeminventarisatie en Landevaluatie en bij Team Bodemgeografie van Alterra. Nu, 33.000 regels computerscript, 600 bodemmonsters, ruim 1300 meter grondboren (de eerlijkheid gebiedt mij er direct bij te vermelden dat ik hiervan slechts een kwart zelf heb gedaan) en acht werkplekken later, ligt er dit proefschrift. Promoveren gaat uiteraard niet vanzelf en veel werk gebeurt eenzaam achter de computer. Toch zijn er talloze personen geweest die mij de afgelopen vijf jaar op allerlei manieren hebben bijgestaan en hiermee een belangrijke rol hebben gespeeld bij de totstandkoming van dit proefschrift. Vanzelfsprekend wil ik hen hiervoor bedanken. En dat ga ik nu doen!

Allereerst mijn drie begeleiders: Gerard, Dick en Jetse. Mijn dank aan jullie is moeilijk in woorden uit te drukken. Wat ik wel kan zeggen is dat een promovendus zich geen beter begeleiding kan wensen zoals ik deze van jullie al die jaren heb gehad. Het feit dat mijn promotieonderzoek zo soepel is verlopen is voor een groot deel aan jullie te danken. Ik heb ontzettend veel van jullie geleerd op het gebied van geostatistiek, steekproeftheorie, het doen van wetenschappelijk onderzoek en het schrijven van artikelen. Het was me een groot genoegen om met jullie te mogen samenwerken! Tom, als voormalig leerstoelhouder, en Mirjam, als voormalig teamleidster van Bodemgeografie, jullie hebben deze AIO-positie mogelijk gemaakt met als doel de samenwerking tussen universiteit en Alterra te bevorderen en jullie hebben mij daarop aangesteld. Ik hoop dat ik daarin geslaagd ben. Hartelijk dank voor het gestelde vertrouwen.

Ik heb vijf jaar deel uit mogen maken van Alterra's team Bodemgeografie. Een clubje mensen met een voorliefde voor de bodem: voor de één in de vorm van zand, klei, veen en voor de ander in de vorm van bits en bytes. Bedankt voor alle gezelligheid tijdens lunches, uitjes en etentjes. Voor mij waren het vijf plezierige jaren en ik ben blij dat ik sowieso nog een aantal jaren in jullie midden mag vertoeven. Er zijn een aantal Bodemgeografen die ik hier in het bijzonder wil noemen. Folkert, bij jou kon ik altijd terecht met verzoeken voor en vragen over bodemgegevens uit het bodemkundig informatiesysteem. Als projectleider van het grote veenactuali-

satie project heb jij mij de ruimte gegeven om binnen dit project mijn eigen weg te zoeken wat uiteindelijk heeft geleid tot hoofdstukken 5 en 6 van dit proefschrift. Mijn dank daarvoor is groot. Ik ben verheugd dat we de kennis opgedaan in dit proefschrift komende jaren kunnen gaan inzetten in de praktijk. Fokke wil ik bedanken voor het beantwoorden van vele bodemkundige vragen door de jaren heen en voor de week die we samen in het Drentse zand en veen hebben doorgebracht om mijn veldbodemkundige kennis wat op te krikken. Ik heb veel van je geleerd! Dennis, R-goeroe van team Bodemgeografie, bedankt voor al je waardevolle hulp met R. Martin, bedankt voor het redigeren van de Nederlandse samenvatting en voor dat L^AT_EX bestandje van je (proefschrift was het toch?). Dat laatste heeft me enorm op weg geholpen met de opmaak van dit proefschrift. Gert, Willy, Ebbing en Matheijs wil ik hartelijk danken voor hun enorme inzet tijdens het verzamelen van de validatiedata voor de hoofdstukken 5 en 6. Joop, als teamleider van Bodemgeografie heb je me tijdens het laatste jaar veel vrijheden in mijn werkzaamheden gegeven en heb je het voor elkaar gekregen dat ik, ondanks economisch zwaar weer, mijn carrière komende jaren kan voortzetten bij het team. Ik ben je daar erg erkentelijk voor.

Bij de leerstoel Landdynamiek dank ik alle AIOs en oud-AIOs voor alle gezelligheid tijdens de koffiepauzes, lunches en allerhande sociale activiteiten. Ook hier wil ik een aantal mensen persoonlijk noemen. Jantiene, we zijn ongeveer tegelijkertijd als AIO begonnen en zijn nu ook bijna tegelijkertijd klaar. Bijna vijf jaar hebben we de hoogtepunten en dieptepunten (gelukkig waren dat er niet zoveel) in het leven van een promovendus gedeeld (met bijbehorende stemmingen) tijdens onze gezamenlijke kantoordag op vrijdag in Gaia en op onze 'koffiehangplek' in Atlas. Bedankt voor je gezelschap. Ik beschouw het als een eer dat ik al die jaren je koffiemesje, secretaresse en ArcGIS helpdesk heb mogen zijn ;). Verder wil ik Arnaud bedanken voor het altijd netjes terugbetalen van z'n snoepautomatleningen ;), Wieteke voor shaRing, Catherine voor het (niet) in leven houden van onze kantoorplanten tijdens mijn afwezigheid ;), en Dirk voor de vele goed gesprekken en het zo nu en dan delen van de PhD-blues; nu nog die Zimbabwe paper gepubliceerd krijgen. Nynke, bedankt voor van alles en nog wat en bij deze ook alvast voor de foto's! Marthijn en Gert, het was me een groot genoegen om jullie vier jaar te mogen assisteren tijdens het bodemkundig veldpracticum voor eerstejaars: van de Drentse keileem naar de Limburgse löss en van de fruitboomgaarden in de Betuwe naar de schorren van Zeeland. Ik vond het erg leuk om te doen en voor mij was het denk ik net zo leerzaam als voor de studenten zelf. Henny, bedankt voor alle onmisbare administratieve hulp.

En dan mijn paranimfen: Maarten en Bouke. Bouke, ondanks de ontelbare oorlogen die we met dobbelstenen hebben uitgevochten, bedankt voor 21 jaar hechte vriendschap en je niet aflatende interesse in mijn bezigheden. Maarten: oud-studiegenoot, squashmaatje, ex-huisgenoot, mede-concertganger en worthy opponent aan de speltafel; bedankt dat je ook tijdens mijn promotie aan mijn zijde wil staan.

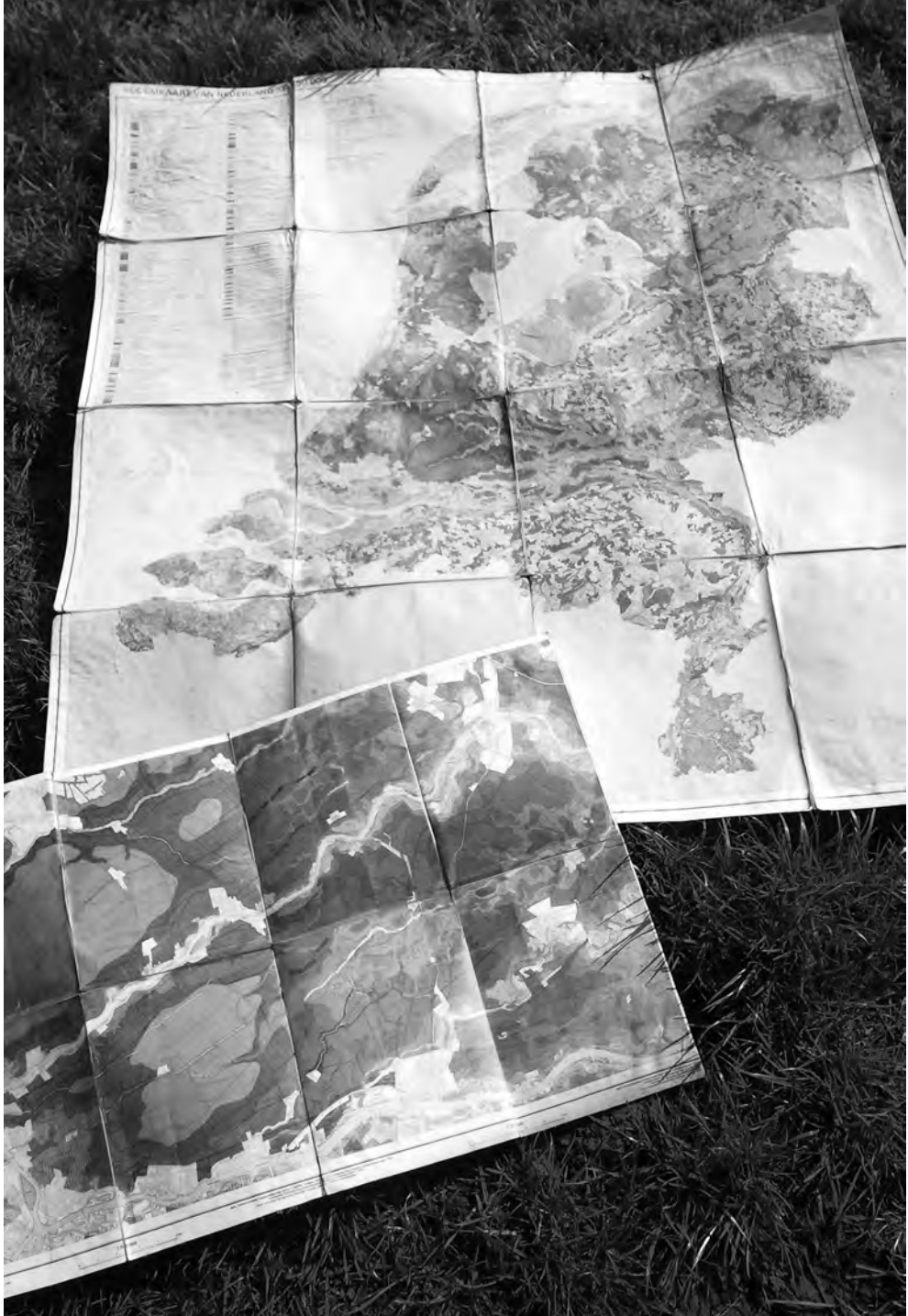
Voor de broodnodige ontspanning in de vorm van lange spelavonden en -middagen,

filmmarathons, BBQs, zuurkoolgies, pannenkoekenfeestjes, bioscoopbezoekjes, zeiltochtjes en uitstapjes naar Jena zorgden Sipke en Marlide, Hans en Priscilla, Bram en Irka, Johan en Ebertine, en sinds kort ook Wessel en Imie. Maarten, jij hoort hier uiteraard ook bij maar jou had ik al genoemd ;). Ook aan jullie mijn hartelijke dank voor jullie jarenlange vriendschap. De vele partijtjes squash op maandag en woensdag waren niet alleen erg leuk maar hielpen ook om mijn hoofd na een lange werkdag te legen. Diederik, Noortje, Matthijs en Mike, bedankt daarvoor!

During the final months of my PhD I had the opportunity to visit the Department of Crop & Soil Sciences at Cornell University. Harold and Bianca, thank you so much for your hospitality and for making my stay, albeit short, a very pleasant one. It has been a great experience.

Pap en mam, bedankt voor de onvoorwaardelijke steun die ik altijd van jullie heb gehad. Het is goed te beseffen dat er ergens in Brabant altijd een deur voor me open staat.

En tenslotte: Lieve Els, al bijna 10 jaar mijn lief en mijn maatje. Hoeveel je voor me betekent hoef ik hier niet uit te leggen. Bedankt dat je er voor me bent.



Chapter 1

Introduction

1.1 Background

Soils are back on the global agenda (Hartemink, 2008). An increasing world population, rapid economic growth in large parts of the world and the energy crisis are putting pressure on the soil as a natural resource. The soil must feed an estimated nine billion people around 2050, but food production is threatened by soil fertility decline as a result of poor soil management and by competition for fertile lands from biofuels and animal feed. Besides food production, the soil plays an important role in climate change and the carbon cycle (Gorham, 1991; Lal, 2004). Global concerns about food security and climate change and increasing awareness about the importance of sustainable soil management to protect and preserve soil resources for society and environment, have instigated that soils are increasingly mentioned in global studies conducted by the United Nations, World Bank and FAO (Hartemink and McBratney, 2008). Soil legislation is developed on national and continental level. Examples are the adoption of a Resolution on soil in the US Senate (Hartemink and McBratney, 2008) and the Thematic Strategy for Soil Protection of the European Union (EU) (Commission of the European Communities, 2006; Bouma and Droogers, 2007; Bouma, 2010). The objective of the EU Thematic Soil Strategy is the protection and sustainable use of soils by preventing further degradation, preserving its functions and restoring degraded soils (European Commission, 2006). These functions are i) production of food and biomass, ii) storing, filtering and transforming compounds, iii) providing habitat and gene pool, iv) providing the physical and cultural environment for human activities, v) source of raw materials, vi) acting as carbon pool, and vii) archive of geological and archeological heritage (Commission of the European Communities, 2006). Along with the soil functions, the EU defined eight main soil threats. These are erosion, organic matter decline, contamination, salinisation, compaction, soil biodiversity loss, sealing, floods and landslides.

Renewed global interest in soil has fuelled demand for accurate, up-to-date, high-resolution geographical soil information to assess effects of soil threats on soil functions, to study the role of soil as a source or sink of greenhouse gasses and to support policy-making in soil resource management to secure food production. To meet this demand, several soil mapping projects were recently initiated across the globe, of which the *GlobalSoilMap.net* project¹ (Sanchez et al., 2009) is perhaps the most ambitious. This project aims to make a new digital soil map of the world using digital soil mapping (McBratney et al., 2003), which comprises quantitative, state-of-the art technologies for soil mapping.

Also in the Netherlands there is growing demand for soil information for a great variety of purposes. The national soil map at scale 1:50 000 is the main source of soil information. This map was initially created for soil suitability analysis for various land-use systems (van Lynden et al., 1985; Sonneveld et al., 2010), but since the

¹<http://www.globalsoilmap.net/>

1990s it is increasingly being used for environmental and agro-economic analyses in support of policy-making. Examples include modelling of nutrient and pollutant fluxes in the soil (van der Salm et al., 1996; Hack-ten Broeke et al., 1999; Kros et al., 2011), inventories and monitoring of carbon stocks (de Groot et al., 2005; Schulp and Veldkamp, 2008; Reijneveld et al., 2009), modelling soil subsidence (Hoogland et al., 2011), implementation of the EU Thematic Soil Strategy (Bouma and Droogers, 2007) and simulation studies on greenhouse gas emissions from peat soils (Nol et al., 2010; van Beek et al., 2011). These environmental applications of soil information greatly outnumber the land suitability studies (de Vries et al., 2008). A consequence of the shift soil data applications is that the national soil map often fails to meet the requirements of the current generation soil data users. User inventories on data requirements held between 2004 and 2008 revealed several shortcomings in the current status of soil spatial information in the Netherlands (de Vries et al., 2008). Digital soil mapping can play a key role in tackling these deficiencies and in supplying soil data users with the information necessary to address current issues in soil protection, conservation and agro-environmental policy-making.

The remainder of this chapter introduces conventional and digital soil mapping in section 1.2. Section 1.3 focusses on soil information in the Netherlands and describes trends in soil mapping, the development of soil information system *BIS* and the current status of soil information. Section 1.4 defines the objectives and research questions of this thesis, provides a description of the study area and presents an outline of the thesis.

1.2 Soil mapping

1.2.1 Conventional soil mapping

Conventional soil mapping (CSM) matured in the second half of the 20th century (Schelling, 1970; Dijkerman, 1974). The original methodologies were refined over time, although important methodological differences between schools and countries remain until today. The soil survey manual (Soil Survey Division Staff, 1993)—considered a standard work by many—provides the major principles and practices for making and using soil surveys at the end of the 20th century. Soil maps are typically created with the free survey method. Free survey starts with physiographic aerial photo interpretation in which the soil surveyor employs a mental soil-landscape model to delineate soil boundaries based on landscape features. Aerial photo interpretation is followed by field survey. Based on the mental model, the surveyor selects sample locations from which the most useful information is likely to be obtained (Bregt, 1992b). The field observations are used to check and further develop the mental model of soil formation, to confirm or modify soil boundaries and to de-

termine map unit composition. Sampling effort concentrates on 'problem areas', i.e. areas where the soil pattern does not fit the mental model (Rossiter, 2000) and on areas with more variation. Free survey results in a general-purpose soil class map and a set of soil profile descriptions. Each map unit is characterized by one or more representative soil profiles, which are used for the interpretation of the soil map (Bregt, 1992b).

In the Netherlands, conventional soil maps are created with free survey but without the aerial photo interpretation stage. Soil boundaries are located and drawn based on field observations. Nowadays a high-resolution digital elevation model (DEM) assists the surveyor in the delineation process. Sampling locations are chosen purposively and are considered 'representative' for the field that is surveyed. These follow a fairly regular pattern across the survey area. At the selected locations soil profiles are described and classified from auger borings. A smaller set of detailed descriptions is obtained from soil pits. In addition, extra (often non-recorded) observations are made to determine the boundary of the map delineations.

The conventional approach to soil mapping can produce accurate maps but is labour intensive and often impractical in large, inaccessible areas such as for example in Australia (Bui and Moran, 2003). Because of the qualitative nature of conventional methods, i.e. the use of soil-landscape models is not formalized in the methodology, these are often considered as much an art as a science (Hewitt, 1993). Main criticisms of CSM include irreproducibility because the soil surveyor's mental soil-landscape model employed in deriving soil mapping units is often not recorded; soil bodies that are strictly represented as discrete, homogeneous entities (information is lost because the reporting format requires the data to be classified), which in many cases is unrealistic; lack of quantified measures of accuracy; and 'static' representation of soil variation (Hewitt, 1993; Goovaerts and Journel, 1995; Hartemink et al., 2010). Moreover, CSM is based on soil classification, which is a hierarchical, rigid system. The legend associated to a classification system is central to mapping, whereas ideally the soil data and their applications should be. This is especially true when considering today's environmental-centered research approach in which soils are an integral part of an ecosystem that interact with other environmental factors (Grunwald, 2010; Bouma, 2010).

1.2.2 Digital soil mapping

Quantitative methods for the study of the spatial distribution of soils based on regionalised variable theory were introduced in the early 1980s (Burgess and Webster, 1980; McBratney et al., 1982). These methods quickly gained popularity, assisted by the rise of the personal computer, and led to the establishment of the Pedometrics Working Group (later Commission) of the International Union of Soil Sciences in the late 1980s and the first conference on Pedometrics in 1992. Pedometrics is defined as

the '*application of mathematical and statistical methods for the study of the distribution and genesis of soils*' (Webster, 1994; Burrough et al., 1994).

Digital soil mapping (DSM) applies pedometric methods to map (predict) the spatial (and temporal) distribution of soils (McBratney et al., 2000, 2003; Grunwald, 2006). It is defined as the '*creation and population of spatial soil information systems by use of field and laboratory observational methods coupled with spatial and non-spatial soil inference systems*' (Lagacherie and McBratney, 2007). It aims to extend the functionality of soil information systems from storage of (digitised) conventional soil maps to production of soil maps to meet current and future demand for accurate, up-to-date soil information. Developments in DSM gained momentum in the 1990s (e.g. McKenzie and Austin, 1993; Odeh et al., 1994; Knotters et al., 1995), triggered by the great explosion in computation power and information technology. With that came vast amounts of high-resolution environmental data such as DEMs and satellite imagery. During the first decade of the 21st century DSM moved forward from research to operational phase and became a true global venture (Lagacherie and McBratney, 2007; Hartemink et al., 2008; Sanchez et al., 2009; Boettinger et al., 2010). DSM is also referred to as '*pedometric mapping*' (Hengl, 2003), which is occasionally used in this thesis as well. Furthermore, in this thesis '*digital soil map*' refers to a map generated by pedometric methods and not to a digitised, conventional soil map.

The conceptual framework of DSM, like that of CSM, is based on the original model of Jenny (1941) that describes soil formation in terms of the main soil forming factors: climate, organisms, relief, parent material and time. DSM formalizes the relationships between soil and soil forming factors by deriving empirical relationships between observations on soil and explanatory variables that represent the soil forming factors (McBratney et al., 2003). These explanatory variables are derived from spatially referenced environmental data layers such as DEMs, satellite images, geology maps and soil maps. Once quantified in form of a statistical model, the relationships between soil and soil forming factors can be used to predict soil at locations where field observations are lacking but where environmental data are available. There is a great variety of models in the DSM toolbox such as classification and regression trees (Bui et al., 2006; Grimm et al., 2008), neural networks (Behrens et al., 2005), various forms of kriging (Lark et al., 2006; Hengl et al., 2007a), expert systems (Skidmore et al., 1996), Bayesian maximum entropy (D'Or and Bogaert, 2004; Brus et al., 2008) and Kalman filtering (Webster and Heuvelink, 2006). Empirical models are the most widely applied but process-based or mechanistic models exist as well (e.g. Stacey et al., 2006; Minasny and McBratney, 2006b; Finke, 2011).

Digital soil mapping is, in contrast to CSM, data-centered. It takes the soil data as starting point, instead of a legend, and transforms these data into soil information required for a specific application, i.e. DSM is environmental-centered soil mapping which better fits the current view of soils as being part of an ecosystem (Grunwald, 2010) than does CSM. Various models can be used for this purpose, depend-

ing on the application. DSM is not limited to use of one specific model such as in CSM and is thus less rigid. Moreover, the main drawbacks of CSM do not apply to DSM. DSM methods are reproducible: a prediction model is documented in computer script which can be stored and rerun. Soil spatial variation can be represented with different models of spatial variation. It is generally modelled as a continuous phenomenon with a continuous model of spatial variation (e.g. kriging) but discrete or mixed models of spatial variation can be applied as well (Heuvelink, 1996; Heuvelink and Huisman, 2000). Use of (geo)statistical models does not only produce a soil map but also quantifies the associated prediction uncertainty. In addition, it is assumed that DSM is much more efficient than CSM. Despite these advantages compared to CSM, DSM methods have drawbacks of their own. Conventional soil maps are general-purpose maps that provide information on the three-dimensional variation of a wide range of soil properties (including hard-to-measure properties) via representative profiles, whereas digital soil maps are generally specific-purpose maps: a map is created of a specific soil property at a specific depth interval. Multivariate methods can be used for DSM but these become prohibitive when a large number of properties is considered. Furthermore, complex soil forming processes might be difficult to quantify and represent by environmental explanatory variables, while these can be more easily taken into account in free soil survey. Also, DSM methods for soil classes are limited by the number of classes that can be handled (Brus et al., 2008). While conventional methods are standardized (e.g. Soil Survey Division Staff, 1993), DSM methods generally lack standardization. This can complicate transferability of soil prediction models developed in one region to similar and/or other regions (Grunwald, 2009). Finally, the success of DSM depends on the availability of (up-to-date) soil data and environmental data layers. In areas with limited data, producing accurate maps with pedometric techniques is challenging (Stoorvogel et al., 2009).

1.3 Soil information in the Netherlands

1.3.1 Trends in soil inventories

The Netherlands has a long-standing tradition in land and soil inventories. In this section I give a brief overview, which is largely based on Buurman and Sevink (1995) and van der Pouw and Finke (1999). Already in the 1860s the first geological map with nationwide coverage was produced at scale 1:200 000, with a legend that distinguished different types of unconsolidated deposits (the Netherlands being a sedimentary basin). During the first decades of the 20th century it became apparent that the geological approach to soil survey could not sufficiently answer agricultural questions regarding soil management and soil fertility. This led to the development of soil science and soil survey as an independent discipline. An important contribu-

tion was made in the 1930s by W.A.J. Oosting, who is considered one of the founding fathers of soil science in the Netherlands. Oosting developed a new approach to soil survey in which soils and their spatial distribution were studied in coherence with geology, geomorphology, vegetation and soil forming processes. His approach was adopted and further developed by C.H. Edelman and became the cornerstone of soil survey in the Netherlands. Until the 1960s soil survey had a physiographic nature. In 1950 the first provisional soil map of the Netherlands at scale 1:400 000 was published, followed by publication of the first soil map with nationwide coverage at scale 1:200 000 in 1965. During fieldwork for this map the need for a standardized approach to soil survey and for a soil classification system based on measurable soil properties became apparent. This resulted in the publication of a morphometric classification system (de Bakker and Schelling, 1966), a legend for a new nationwide soil survey at scale 1:50 000 and a standardized methodology (Steur, 1961; Buringh et al., 1962). The first sheet of the 1:50 000 soil map was published in 1964, the last in 1995. In 1984 a soil information system became operational (see section 1.3.2). One year later, a nationwide soil map at scale 1:250 000 soil map was published (Steur et al., 1985).

Quantitative methods for soil inventories were introduced in the mid-1980s and 1990s to meet changing soil information needs that resulted from new societal environmental concerns (see section 1.1). In this period research focused on a great variety of topics such as the development of pedotransfer functions for difficult-to-measure soil properties (Wösten and van Genuchten, 1988), methods for soil mapping (Bregt et al., 1992; Heuvelink and Bierkens, 1992; Knotters et al., 1995), strategies for updating soil survey information (Brus et al., 1992), comparison of (simple) quantitative methods for soil mapping with conventional methods (Bregt et al., 1987; Brus et al., 1996), numerical methods for soil classification (McBratney and Gruijter, 1992), the effect of soil spatial variation on soil processes (Finke et al., 1992; Finke, 1993), development of design-based sampling methods for soil map accuracy assessment (de Gruijter and Marsman, 1985; Marsman and de Gruijter, 1986; Brus, 1994; Domburg et al., 1994; Brus et al., 1999; Brus, 2000) and the fundamental difference between design-based and model-based sampling (de Gruijter and ter Braak, 1990; Brus and de Gruijter, 1993, 1997). After 2000, quantitative research became more applied and was often carried out in support of research in other soil science disciplines such as soil quality (e.g. Brus et al., 2002; Brus and Jansen, 2004).

1.3.2 BIS: the Dutch soil information system

Digital archiving of soil information started in the late 1960s with the storage of chemical data from soil horizons of well-described reference soil profiles (Leeters et al., 1990). In the second half of the 1970s map sheets of the 1:50 000 map as well as standard forms with soil profile descriptions were digitized and stored. This created the first digital soil database that contained both soil point and areal data. In 1983 the

project 'BIS' (Dutch acronym for 'soil information system') was initiated with the aim to develop a relational database (ORACLE) for storage of soil point data and maps. *BIS* became operational in 1984. Initially it only stored point data from reference soil profiles. In 1988 digitization of the 1:50 000 soil map was completed and the map was linked to the database. Between 1984 and 1990 *BIS* was gradually extended with point data from various research and rural land restructuring projects as well as large-scale (1:10 000-1:25 000) soil maps created for the latter. In 1990 *BIS* contained almost 4 400 soil profile descriptions with measurement data on a great variety of soil properties and over 80 000 basic profile descriptions with hand-estimates of basic soil properties such as texture fractions and soil organic matter content. Storage of soil data and maintenance of *BIS* was funded from governmental programme funds.

The mid-1980s saw a growing demand for quantitative information on the quality of the 1:50 000 soil map and on the variation of soil properties within the map units. Focus was not only on the traditional soil properties but also soil properties that had become relevant for environmental research such as the Phosphate Sorption Capacity and the Phosphate Saturation Degree. In 1988, a project was initiated to obtain statistically reliable information about soil spatial variation within the map units of the 1:50 000 soil map. This project developed into the Netherlands Soil Sampling Programme in the 1990s (Visschers et al., 2007). At the same time there was a growing awareness that data in *BIS* became outdated, particularly data on the presence and thickness of peat. This led to the revision and updating of several map sheets of the 1:50 000 map (which were the last update activities until to date) (Makken and de Vries, 1989; Finke et al., 1996). In 2004, government funding for maintenance and extension of *BIS* halted. Since that time no new (up-to-date) point data and maps were added to *BIS*. Also the funding of strategic research on methods for collecting and processing (mapping) soil data came to a standstill, which hampered methodological developments.

In the context of updating and upgrading soil information, it is worth noting that in the late 1990s a large project was initiated, funded by the national government and regional water authorities, to update the groundwater table maps (Finke, 2000). In addition, a more dynamic (high-resolution) description of the groundwater table was required. For this purpose, a method was developed to obtain a large set of parameters describing groundwater table dynamics (Finke et al., 2004). This method used geostatistical techniques that are very similar to techniques that are nowadays used in DSM.

Since 2004 *BIS* can be accessed through the internet². Presently, *BIS* stores the 1:50 000 and 1:250 000 soil maps, regional soil maps at scales 1:10 000 and 1:25 000, a soil sample archive with 15 000 samples from 8 000 locations with measurements on texture and chemical and physical properties, soil physical characteristics of the main texture classes, and spatially-referenced soil profile descriptions from over 307 000 loca-

²www.bodemdata.nl

tions (de Vries et al., 2008).

1.3.3 Current status of soil information in BIS

The 1:50 000 national soil map still is the main source of soil information in the Netherlands. Almost a hundred ministries, research and education institutes, consultancy and utility companies, provinces, and over 150 municipalities (40% of the total) obtained a license for use of the digitized map (de Vries et al., 2008). However, soil data user inventories held between 2004 and 2008 revealed four shortcomings of the current soil spatial information in *BIS* that hamper efficient and reliable use in today's agro-environmental research:

1. *Outdated soil information.* The map sheets of the 1:50 000 soil map are 20–50 years old. Outdated soil information particularly concerns the areas with peat soils (ca. 527 000 ha). Intensive agricultural use and deep drainage in these areas causes mineralization of peat layers. Recent inventories on the status of peat soils showed great changes in soil conditions in the north-eastern agricultural peatlands. An estimated 47% of the deep peat soils have changed to shallow peat soils or mineral soils and approximately 50–60% of the shallow peat soils have changed to mineral soils (van Kekem et al., 2005; de Vries et al., 2009). It is estimated that 365 000 ha peat soils require updating.
2. *Lack of thematic soil maps.* Thematic maps of basic soil property such as organic matter, texture fractions, pH or phosphate sorption capacity are essential input for various environmental models that predict and evaluate the effect of policy measures on for example crop growth, soil acidification, pesticide and nutrient leaching, carbon sequestration and land-surface greenhouse gas emission. At the present, *BIS* does not allow storage of soil property maps. These maps are currently made on ad-hoc basis. There is no standardized methodology to transform soil data in *BIS* to soil information that is needed as input for environmental research. Typically the description of a representative profile is used to determine the soil property of the map unit that is linked to that profile. The soil property value of interest is then assigned to the map delineations that comprise the map unit. Often a generalized soil map with 21 map units is used for this purpose given the very large number of map units of the 1:50 000 map. This approach results in maps with crisp boundaries (lateral and vertical). In many cases a continuous or mixed discrete-continuous representation of spatial variation is more truthful. Also, no use is made of environmental ancillary data and soil point data stored in *BIS*. Sometimes DSM methods are used to create thematic soil maps but the resulting maps are not stored in *BIS* and the used methods are not always documented.

3. *Lack of quantified measures of accuracy.* Soil maps and point data stored in *BIS* are not error-free. Errors in soil maps propagate to output of environmental models that use these maps as input and may for instance affect the accuracy of estimates of nutrient or pesticide leaching, soil carbon stock or greenhouse gas emissions. Therefore there is a growing demand for quantitative information on the accuracy of soil data stored in *BIS* and maps derived from these data.
4. *Lack of detailed soil information.* Many environmental applications require soil information on a local scale for which the 1:50 000 map does not provide sufficient detail. Large-scale (1:10 000–1:25 000), digitally available soil maps cover roughly 13% of the country (400 000 ha).

1.4 Content of the thesis

1.4.1 Objectives

There is a need to update the national soil map and to extend *BIS* with full-coverage thematic maps of all major soil properties with quantified accuracy at multiple scales. In 2009 the Dutch government commissioned a six-year research programme named *BIS2014* with the aim to update and upgrade soil information in *BIS*. This programme gave a new impulse to research on soil mapping methods and aims to update the 1:50 000 soil map for the areas with peat soils. It has been recognized that DSM can play a key role herein. However, despite a long history in soil survey, extensive research on quantitative methods for soil inventory between the late 1980s and mid 1990s and rapid global advancements in DSM methods during the past decade, DSM has not yet been applied in an operational way in the Netherlands. Soil mapping kept its qualitative nature until to date. The main objective of this thesis is therefore to investigate and evaluate the merits of digital soil mapping for updating soil information in the Netherlands.

The main objective can be divided into the following sub-objectives with their respective research questions:

1. Updating the 1:50 000 soil map.
 - (a) Can the accuracy of the current 1:50 000 soil map be improved by updating using a simple digital soil mapping model and legacy soil data from *BIS*?
 - (b) Does incorporation of spatial dependence in a model for digital soil type mapping result in a more accurate soil map than use of a non-spatial model?

2. Development of digital soil mapping models for updating soil property maps.
 - (a) How can a soil map with quantified uncertainty be used for mapping the soil organic matter content (SOM) and does this approach improve prediction accuracy compared to use of a conventional geostatistical approach?
 - (b) How can pedological knowledge be integrated with geostatistics for modelling the three-dimensional spatial variation of soil properties?
 - (c) Do depth functions of SOM constructed by pedometric methods give better predictions of SOM stock than conventional depth functions?
3. Accuracy and efficiency assessment of digital and conventional soil mapping for updating soil information.
 - (a) What is the effect of mapping intensity (in a monetary unit ha^{-1}) on accuracy of digital soil type and property maps?
 - (b) Are digital soil mapping methods more efficient than conventional methods for updating soil type and soil property maps?
 - (c) How can soil maps best be validated?

1.4.2 Study area

The case studies presented in this thesis are all located in the province of Drenthe, the Netherlands. In Chapters 2 and 3 the province itself is the study area. The methodology developed in Chapter 4 is illustrated with a case study for a small area in south-central Drenthe. Chapters 5 and 6 focus on a study area in the southeast of Drenthe: the cultivated peatlands.

The province of Drenthe is situated in the northeastern part of the Netherlands between $52^{\circ}12'$ and $53^{\circ}12'$ northern latitude and $6^{\circ}7'$ and $7^{\circ}5'$ eastern longitude (Fig. 1.1). The size of Drenthe is about $2\,680\text{ km}^2$. Agriculture (dairy farming, potatoes, wheat, maize) covers 67% of the area; natural areas, 21%; built-up areas, 9%; the remainder of the area is composed of water and infrastructure. Altitude ranges between -1 and 30 m above sea level.

Formation of the present-day landscape of Drenthe started at the end of the Mid-Pleistocene during the Saalian ice age (240-128 ka BP). A continental ice sheet covered the northern part of the Netherlands 160 000 years ago. Under the ice sheet a thick layer of glacial till was deposited. A deep glacial meltwater valley formed east of the till plateau, the Hunze valley, which was partly filled with fluvio-glacial sands when the ice sheet melted. The upper part of the till weathered during the Eemien interglacial (128-116 ka BP). The first part of the Weichselian ice age (116-73 ka BP) was cold and wet. Meltwater incised and eroded the till plateau, forming an extensive system of brook valleys. In the course of the Weichselien the brook valleys were

partly filled with eroded material of the till and fluvial sands. Aeolian cover sands were deposited on the glacial till and in the brook valleys during the coldest and driest part of the Weichselian (40-11.5 ka BP). This further levelled the relief and formed a slightly undulating landscape with ridges and flats. The climate became more gentle in the Holocene (since 11.5 ka BP). Groundwater levels rose and fen peat started forming in the brook valleys. At the same time, oligotrophic peat formed in depressions on the till plateau. Around 5000 BP oligotrophic peat formation started in the Hunze valley as a result of a changed drainage situation. The peat on the plateau and in the Hunze valley quickly expanded over the landscape forming vast, raised highmoor bogs. Eventually these bogs covered one third of Drenthe (Spek, 2004). During the Middle Ages, drift-sands and open-field agricultural complexes formed on the till plateau as a result of the plaggen-based, open-field farming system. Large-scale, systematic peat cutting in the highmoor swamps and subsequent reclamation for agriculture began in the 17th century and continued until the mid-20th century. This transformed the highmoor swamps to a man-made, agricultural landscape: the cultivated peatlands (or, translated from Dutch, the *colonized peat landscape*).

The soils in Drenthe formed during the Holocene. Podzols formed in infertile coversand deposits on the till plateau, while brown forest soils formed in richer, more loamy parent material. Plaggen soils were formed by plaggen-based agriculture in the open fields surrounding medieval settlements on the plateau (Blume and Leinweber, 2004; Spek, 2004). These soils have a thick (>30 cm) humic topsoil. The plaggen-horizon typically covers a podzol or a brown forest soil. Peat soils dominate the centres of the brook valleys. Hydromorphic earth soils, which are soils with a humic topsoil that overlies a C-horizon that may have gley features, are found in the brook valley-plateau transition zone. Raw sand soils, which are soils without pronounced signs of soil formation, are found in the drift-sand complexes bordering the open fields. Soils of the cultivated peatlands are distinguished from the peat soils of the brook valleys by their strong human disturbances to a considerable depth as a result of deep cultivation and their anthropogenic topsoil, which is a mixture of sand and peat fragments.

1.4.3 Outline

This thesis comprises seven chapters, including this introductory chapter. Chapters 2 to 6 form the core of this thesis. These chapters address the objectives and research questions presented in section 1.4.1. In *Chapter 2* the national soil map for the province of Drenthe is updated using a simple DSM model and legacy soil data. The accuracy of the updated map is compared with that of the current map (Question 1a). *Chapter 3* presents a model for spatial prediction of quantitative soil properties that uses a soil type map with quantified uncertainty (Question 2a). In *Chapter 4* a method is proposed for three-dimensional DSM of SOM based on soil type-specific depth functions. Predictions with these depth functions are compared with those

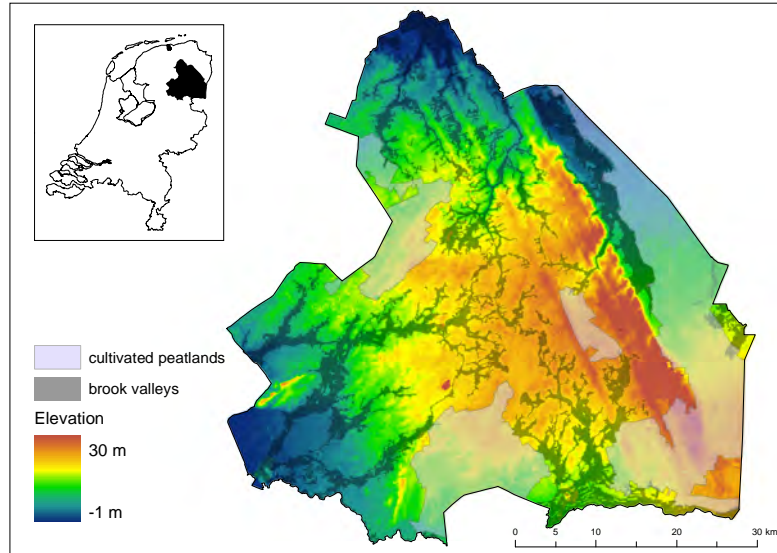


Figure 1.1: The province of Drenthe. The light grey-shaded area indicates the extent of the cultivated peatlands while the dark grey-shaded are shows the system of brook valleys. The inset shows the location of Drenthe in the Netherlands.

obtained by conventional depth functions (Question 2b, 2c). *Chapter 5* returns to the topic of updating the national soil map and introduces the generalized linear geostatistical model for digital soil type mapping (Question 1b). In *Chapter 6* the efficiency of DSM methods is compared with that of conventional methods. In addition the effect of mapping intensity on map accuracy is assessed for digital soil type and property maps (Questions 3a-b). Validation of the soil maps is an important topic in each of the Chapters 2 to 6 (Question 3c). *Chapter 7* concludes this thesis and summarizes and discusses the main research findings of this thesis.

Chapters 2, 3 and 4 are based on articles that have been published in refereed journals, whereas Chapters 5 and 6 are submitted to refereed journals. Chapters 2 to 6 can be read separately—although a reference will be made occasionally to a table or figure in an earlier chapter—as a consequence there might be some overlap between chapters. Literature references have been combined in the *References* section at the end of this thesis.



Chapter 2

Updating the 1:50 000 Dutch soil map using legacy soil data: a multinomial logistic regression approach

The 1:50 000 national soil map of the Netherlands is gradually becoming outdated. Intensive land and water management have had great impact on particularly peat soils. The area of these soils has substantially declined through oxidation of the peat layer soil since the survey was completed. This chapter assesses the possibility of updating the national soil map for the province of Drenthe by multinomial logistic regression modelling of ten major soil types using legacy soil data from BIS and high-resolution environmental covariates. Special attention is given to model selection. A framework for selecting logistic regression models was taken from the literature and adapted for the purpose of soil mapping. The model selection process was guided by pedological expert knowledge to ensure that the final models are not only statistically sound but also pedologically plausible. A prediction model was calibrated for each of the ten major soil types depicted on the soil map. The models were used to estimate the probability of occurrence of the soil types at the nodes of a raster with 25-m resolution. Shannon entropy was used to quantify the uncertainty associated to the predictions. The updated soil map was validated with independent probability sample data. The estimated overall purity of the updated map was 58%, which is 6% larger than the overall purity of the existing soil map.

Based on: B. Kempen, D.J. Brus, G.B.M. Heuvelink and J.J. Stoorvogel
Geoderma 151 (2009): 311–326

2.1 Introduction

The 1:50 000 soil map is the major source of soil information in the Netherlands. This map was completed in the early 1990s after more than three decades of field surveys. It is nowadays used for a wide variety of purposes such as agricultural and environmental policy making, nature and soil conservation and archeological prospection. The soil map, however, is gradually becoming outdated (de Vries and Brouwer, 2006; Rosing et al., 2006). Intensive land and water management has had great impact on soil conditions. Large areas of peat soils have disappeared through oxidation rates as a result of tillage and deep drainage. Recent inventories on the status of peat soils revealed that almost 50% of area originally mapped as deep peat soils has changed to shallow peat soils or mineral soils and approximately 50–60% of the mapped shallow peat soils are now mineral soils (van Kekem et al., 2005; de Vries et al., 2009). For two 1:50 000 map sheets in Drenthe, Finke et al. (1996) found that 82% of the deep peat soils had changed into shallow peat soils and that 63% of the shallow peat soils to mineral soils. Use of outdated soil information for environmental research or policy making may lead to erroneous conclusions or decisions.

Although the need for updating the soil map has been recognized for a long time, the last update activities took place in the early nineties. Four map sheets of the national soil map were updated between 1988 and 1993 (Finke et al., 1996). Brus et al. (1992) evaluated the merits of four update strategies for soil maps. Use of conventional soil mapping methods for updating are hampered by their costs. As fieldwork is a major cost component in a project on map updating (Finke, 2000), methods that reduce the amount of fieldwork, such as used in digital soil mapping (DSM), can be attractive for future update activities.

In DSM soil observations are related to readily available, spatially exhaustive environmental data using a (geo)statistical model. The relationships are then extended across a survey area to predict soil at unvisited locations (Bui and Moran, 2003; McBratney et al., 2003). Such methods also may have great potential in the Netherlands for updating soil maps, considering its data rich environment. The Dutch soil information system *BIS* contains soil profile descriptions and classifications at over 300,000 locations. Besides, an extensive suite of high-resolution environmental data is available. These data were combined with point observations to create maps of groundwater status for the sandy soils of the Netherlands using time series analysis and geostatistical techniques (Finke et al., 2004), whereas Brus et al. (2008) used over 8 000 legacy point observations to estimate the probabilities of occurrence of seven soil categories in the Netherlands. Here it is hypothesized that existing, recent soil profile observations can be used to update the peat and other map units of the national soil map. Furthermore, it is expected that the purity of the map units can be increased by using high-resolution ancillary data to delineate inclusions of soil classes other than the dominant soil class.

The objective of this chapter is to update the national soil map for the province of Drenthe without additional fieldwork by using legacy soil data from *BIS*. A soil map update is urgent for Drenthe considering the large area of peat soils in that province. Use of multinomial logistic regression (MLR) for digital soil type mapping is explored. MLR is widely used for spatial modeling in land use and ecology studies (e.g. May et al., 2008; Müller and Zeller, 2002; Rhemtulla et al., 2007; Suring et al., 2008). Application of MLR for DSM, however, is limited to only a few studies, see for instance Bailey et al. (2003), Campling et al. (2002), Debella-Gilo and Etzelmüller (2009), and Hengl et al. (2007b). For the study in this chapter legacy point data obtained from *BIS* are used to calibrate an MLR-model for each of the ten major map units depicted on the soil map of Drenthe. With these models soil type is re-mapped within each map unit. Careful attention is given to the process of model selection since this is perhaps the most crucial step in DSM. A framework for MLR-modeling was taken from the literature and adapted for soil mapping. The model selection framework is described and its application illustrated for one map unit. The accuracy of the updated soil map is assessed by independent probability sample data and compared to that of the existing soil map.

2.2 Methods

2.2.1 Study area

The province of Drenthe (2 680 km²) is situated in the northeastern part of the Netherlands (Fig. 1.1). The landscape in Drenthe is characterized by a glacial till plateau that is dissected by an extensive brook valley system. The till is covered with poor aeolian sand deposits that can be up to two metres thick. Cultivated peatlands border the plateau to the east and south. Podzols formed in the aeolian sand deposits. In loamy parent material, brown forest soils formed. Plaggen soils surround medieval settlements on the plateau. Peat soils dominate the centres of the brook valleys. Hydromorphic earth soils are found in the brook valley-plateau transition zone. Raw sand soils are found in drift-sand complexes. Peat soils of the cultivated peatlands are characterized by strong disturbances to a considerable depth due to deep cultivation. Their topsoil consists of sand mixed with small peat fragments.

2.2.2 Data sources

Soil data

The national soil map for Drenthe distinguishes 96 map units describing 61 soil types at the subgroup level of the Dutch soil classification system (de Bakker and Schelling,

1989), and 35 soil associations. Because it is practically unfeasible to calibrate an MLR-model for a variable with 96 possible outcomes, the legend of the map was generalized to ten map units, representing the major soil types (Fig. 2.1):

1. *Deep peat soils, (P)* (25 000 ha): soils with an organic (organic matter content >15%) surface horizon; at least 40 cm of peat within 80 cm from the surface;
2. *Deep peat soils with a mineral surface horizon, (mP)* (24 800 ha): soils with at least 40 cm of peat within 80 cm from the surface, and a sandy, clayey, or peat-colonial surface horizon less than 40 cm thick;
3. *Shallow peat soils, (PY)* (13 400 ha): soils with an organic surface horizon; at most 40 cm of peat within 80 cm from the surface;
4. *Shallow peat soils with a mineral surface horizon, (mPY)* (36 000 ha): soils with at most 40 cm of peat within 80 cm from the surface, and a sandy, clayey or peat-colonial surface horizon less than 40 cm thick;
5. *Brown forest soils, (BF)* (900 ha): soils with a B-horizon formed by weathering of minerals and incorporation of moder humus;
6. *Podzol soils (PZ)* (93 000 ha): xeromorphic and hydromorphic podzols;
7. *Earth soils (ES)* (13 000 ha): hydromorphic soils with a 15–50 cm thick humic A-horizon, overlying a sandy or loamy C-horizon that might have gley mottling;
8. *Plaggen soils (PS)* (17 000 ha): soils with an anthropogenic, humic A-horizon thicker than 30 cm overlying a podzol or brown forest soil; typical for the open fields on the Drenthe plateau;
9. *Till soils (TS)* (3 500 ha): soils with glacial till within 40 cm from the surface;
10. *Raw sand soils (RS)* (5 800 ha): These are sandy soils with a humus-poor topsoil less than 30 cm thick; subsoil only shows initial or no signs of soil formation.

Although some detail is lost by aggregating map units, the ten soil types still describe the major soil variation in Drenthe. *BIS* contains 16 282 soil profile descriptions that are located in Drenthe (Fig 2.1). Roughly 96% (15 556) of the soil profile observations are located in four areas where 1:10 000 soil surveys were carried out between 1996 and 2005. These areas cover 10% of the total area. The remaining 726 profile observations, collected during various research projects, are scattered across Drenthe. Because of the variety of data sources, the profile observations were collected with different sampling designs. The sampling locations in the 1:10 000 survey areas were selected by purposive sampling. The other locations were selected by both purposive ($n = 469$) and probability sampling ($n = 257$). The recorded soil types were reclassified to the ten major soil types. Cross-tabulation of observed versus mapped soil type showed that at 55% of the observation locations the recorded soil type corresponded to the soil type as depicted on the map. The podzol map unit is the most pure map unit (78%), while the shallow peat soils are the least pure (*PY*: 19% and *mPY*: 30%). The latter shows the effect of oxidation of the peat.

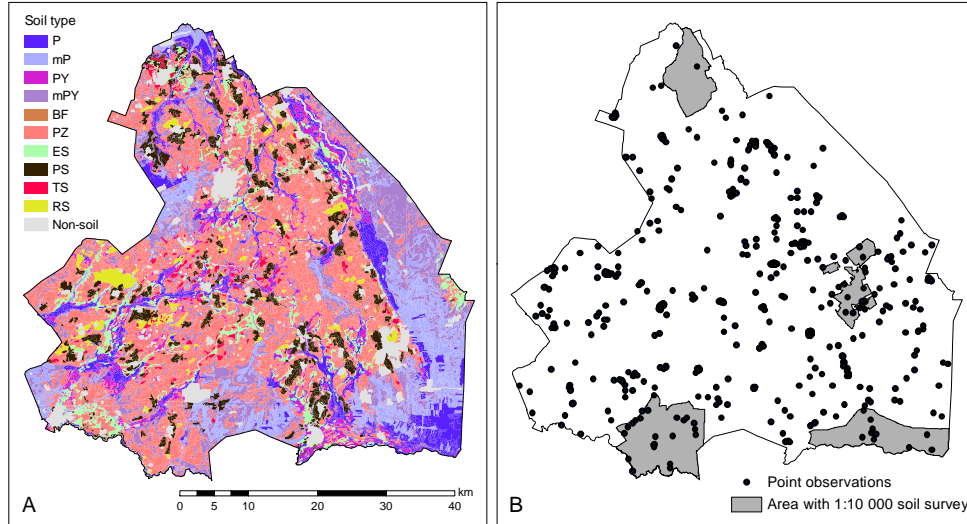


Figure 2.1: Soil map of Drenthe (A) and the locations of the profile observations (B).

Environmental ancillary data

Twelve primary environmental datasets were available (Table 2.1). Polygon maps were converted to raster maps with 25-m resolution. The DEM was used to derive four relative elevation layers using the local mean elevation within search radii of 250, 500, 750 and 1 000 m. Groundwater table classes were regrouped into three classes. Historic land cover (HLC) was regrouped into five classes and resampled to 25-m resolution. Three recent land cover layers were combined into a land cover map for the period 1997-2003 with five classes. The geomorphological units were grouped into 16 classes. The paleogeography grid contained 12 classes. After preprocessing the primary datasets, the environmental data layers were grouped into eight groups: (1) elevation, (2) relative elevation, (3) groundwater, (4) recent land cover, (5) historic land cover, (6) paleo-geography, (7) geomorphology, and (8) soil.

2.2.3 Multinomial logistic regression

The logistic model

The logistic model belongs to the family of generalized linear models and is used when the response variable is categorical. Suppose that variable Y_i represents the observed soil type at a sampling location, with $i = 1, \dots, n$ and n is the number of soil types in a survey area. In case n equals 2 and Y has outcomes y_1 and y_2 . Both the counts of Y_1 and Y_2 follow a binomial distribution. The probability of occurrence

Table 2.1: Available primary environmental ancillary data.

Dataset	Description	Res/Scale	Reference
Digital elevation model	Absolute elevation	25 m	www.ahn.nl
Groundwater			
Table class (GT)	Seasonal fluctuation of phreatic water levels	1:50 000	Finke (2000)
Dynamics class (GD)	Updated GT map: quantitative set of parameters describing groundwater dynamics	25 m	Finke et al. (2004)
GD—MHW	GD—mean highest water table	25 m	
GD—MLW	GD—mean highest water table	25 m	
Land cover			
HLC	Land cover in 1900	50 m	Knol et al. (2004)
LC1997	Land cover in 1997	25 m	Hazeu (2005)
LC2000	Land cover in 2000	25 m	
LC2003	Land cover in 2003	25 m	
Paleogeography	Reconstruction of the landscape of Drenthe by the end of the early Middle Ages (ca. 1000 AD)	1:50 000	Spek (2004)
Geomorphology	Geomorphological units	1:50 000	Koomen and Maas (2004)
Soil map (scale 1:50 000)	Spatial distribution of soil classes	1:50 000	Steur and Heijink (1991)

of Y_1 is π_1 and that of Y_2 is π_2 . Logistic regression relates probability π_1 to the linear predictor using the logit-link function:

$$g(\pi_1) = \ln\left(\frac{\pi_1}{\pi_2}\right) = \ln\left(\frac{\pi_1}{1 - \pi_1}\right) = \mathbf{x}^T \boldsymbol{\beta}, \quad (2.1)$$

where $g(\cdot)$ is the link function, \mathbf{x} is a vector of predictors, and $\boldsymbol{\beta}$ is a vector of model coefficients that are typically estimated by maximum likelihood. Eq. (2.1) can be rewritten as:

$$\frac{\pi_1}{1 - \pi_1} = \exp(\mathbf{x}^T \boldsymbol{\beta}) = \exp(\eta). \quad (2.2)$$

The quotient in Eq. (2.2) is referred to as the odds. From Eq. (2.2) follows that:

$$\pi_1 = \frac{\exp(\eta)}{1 + \exp(\eta)}. \quad (2.3)$$

The binomial logistic regression model is easily generalized to the multinomial case. In case of n soil types there are n variables Y_1, \dots, Y_n with associated probabilities π_1, \dots, π_n . Analogous to binomial logistic regression the odds $\frac{\pi_1}{\pi_n}, \dots, \frac{\pi_{n-1}}{\pi_n}$ are modelled by means of $\exp(\eta_1), \dots, \exp(\eta_n - 1)$. From $\sum_{i=1}^n \pi_i = 1$ it follows that:

$$\pi_i = \frac{\exp(\eta_i)}{\exp(\eta_1) + \exp(\eta_2) + \dots + \exp(\eta_n)}, \quad (2.4)$$

where $\eta_n = 0$. This model ensures that all probabilities are in the interval $[0, 1]$ and that the probabilities sum to 1.

Assessing model significance and contribution of predictors

The significance of the logistic regression model is assessed with the likelihood ratio test (Hosmer and Lemeshow, 2000). Central to this test is the deviance statistic, which is defined as :

$$D = -2 \ln \left(\frac{\text{likelihood fitted model}}{\text{likelihood saturated model}} \right), \quad (2.5)$$

where the quotient is the likelihood ratio. The larger the deviance D , the poorer the fit of the fitted model compared to the saturated model. The likelihood ratio test compares two logistic models by assessing the change in deviance due to inclusion of predictors:

$$G = D(\text{model without the predictor}) - D(\text{model with the predictor}). \quad (2.6)$$

G is the log-likelihood ratio statistic, which is χ^2 -distributed under the null hypothesis that the model coefficients are zero, assuming independent and normally distributed residuals. The likelihood ratio test is used to assess the significance of the overall model by comparing the deviance of the intercept-only model with the full model, and that of the individual predictors. This test can only be used to compare nested models. The significance of an individual model coefficient is assessed with the Wald statistic, which is obtained by comparing the estimated coefficient to an estimate of its standard error (Hosmer and Lemeshow, 2000):

$$W = \frac{\hat{\beta}}{\text{se}(\hat{\beta})}. \quad (2.7)$$

The Wald statistic follows the standard normal distribution under the null hypothesis that a model coefficient is zero. Important for the interpretation of the logistic regression is the value of $\exp(\beta)$, the odds ratio, which indicates the change in odds of an event resulting from a one-unit change in the predictor.

2.2.4 Model building

Pedological knowledge for regression modelling

Regression modeling is a popular method for quantifying the relationship between soil and ancillary data (e.g. Thompson and Kolka, 2005; Meersmans et al., 2008; Schulp and Veldkamp, 2008). Usually a set of predictors is derived from ancillary data, coefficients are estimated for these predictors, followed by an evaluation of the selected model on basis of some statistical performance criterion such as R^2 or Mallows' Cp statistics. The resulting regression model might be statistically sound but can be pedologically questionable if the selected predictors do not have a plausible relationship with the soil variable based on knowledge of the soil-landscape system.

An alternative to data-driven approaches to DSM are knowledge-driven approaches. It has become widely recognized that tacit knowledge of the soil-landscape system provides valuable information that should be integrated into the DSM process (Heuvelink and Webster, 2001; McKenzie and Gallant, 2007; Walter et al., 2007). Such knowledge can be used to build expert systems for soil mapping (Cook et al., 1996; Zhu et al., 2001) or to define a conceptual model of pedogenesis that forms the foundation of a statistical DSM model (Müller and Zeller, 2002; McKenzie and Gallant, 2007; McKenzie and Ryan, 1999). Use of knowledge of the soil-landscape system should be fully integrated throughout the model selection process for regression modelling. Each step of the process should be critically reviewed from a statistical as well as a pedological perspective to ensure a plausible prediction model.

Model building strategy

Hosmer and Lemeshow (2000) provide a methodological framework for building a binomial or multinomial logistic regression model. In this chapter their approach is adopted and extended into a framework for digital mapping of multinomial soil variables. This framework comprises eight steps, which are described hereafter.

(1) Definition of a conceptual model of pedogenesis

To ensure a sound pedological basis of the regression model, a conceptual model of pedogenesis is defined. This is an explicit, structured representation of knowledge of the soil-landscape system of the survey area. The conceptual model identifies the driving factors and processes controlling pedogenesis and soil spatial distribution.

(2) Collection of predictors from available environmental ancillary data

In DSM models, the drivers of pedogenesis are represented or proxied by predictors. The predictors are identified and collected from available environmental ancillary data. The result is a set of predictors, all of pedological importance, that are candidates for the MLR-model.

(3) Univariate analysis and selection of candidate predictors

Selection of predictors for an MLR-model from the set of candidates starts with a univariate analysis of each predictor. For categorical predictors this involves cross-tabulation of the response variable versus each predictor followed by the chi-square test of independence. Attention must be paid to contingency tables with zero frequency cells as these may cause numerical instability during parameter estimation, which is marked by extreme model coefficients and associated standard errors (Hosmer and Lemeshow, 2000). The analysis of the contingency tables is followed by the fit of a univariate MLR-model for each predictor that showed at least a moderate level of association with the response variable. Univariate MLR-models are also fit for continuous predictors. The estimated coefficient and odds ratio of each logit function of the univariate MLR-models should be checked for pedological consistency. Predictors that are significant in the univariate analysis are selected for the next step. Hosmer and Lemeshow (2000) suggest to retain predictors with p -value < 0.25 . The large p -value used is based on the work of Bendel and Afifi (1977) and Mickey and Greenland (1989) who showed that the 0.05 level often fails to identify predictors known to be important. Predictors that are only weakly correlated with the response variable may become strong predictors when taken together in the multivariate model. When univariate analysis resulted in a very large set of candidate predictors only the predictors with the strongest association to the response variable were selected from each variable group (Table 2.1), as predictors within each group are expected to be strongly associated.

(4) Multivariate analysis of selected candidate predictors

A multicollinearity assessment is carried out to identify associated predictors. Next, multivariate MLR-models are fitted, with the aim of selecting one or more preliminary models. The stepwise-forward method was used for model selection with entry probability 0.20 and removal probability 0.25, as recommended by Lee and Koval (1997). Selected MLR-models must be checked for numerical stability and multicollinearity. Numerical problems can be solved by replacing the predictor with another (associated) predictor that describes the same soil forming process, by grouping the levels of the predictor, by omitting the predictor from the model or by omitting the outcome class of the response variable that shows numerical instability (this will induce bias in the predictions). Multivariate analysis of candidate predictors might result in several competing models.

(5) Evaluation of adequacy of the multivariate model(s)

The fit of the MLR-model(s) is followed by verification of the importance of each included predictor using the Wald statistic (Eq. 2.7). When there are competing MLR-models, then verification is done with the best model. Competing models that are not nested cannot be compared with the likelihood ratio test but are compared with goodness-of-fit measures. Assessing goodness-of-fit of logistic regression models is not as straightforward as for linear regression models, and the appropriateness of the various goodness-of-fit measures for logistic regression models is a subject of debate

in the literature (e.g. Mittlböck and Schemper, 1996; Hosmer et al., 1997; Menard, 2000). Three goodness-of-fit measures were used: Pearson- χ^2 statistic, classification tables and the McFadden- R^2 . The Pearson- χ^2 statistic indicates how well the model fits the data. Hosmer and Lemeshow (2000) advise caution when using this statistic for models containing continuous predictors. The χ^2 -distribution then becomes an inadequate approximation of the true distribution of the statistic. Therefore the p -value for this statistic becomes meaningless, although the statistic itself is a good measure of model adequacy: the lower the statistic, the better the model fit. Classification tables were used to derive the calibration purity, which is the proportion of observation locations at which the soil map predicts the correct soil type. The McFadden- R^2 (Menard, 2000) measures the reduction in maximized log-likelihood. It is conceptually and mathematically close to the ordinary least squares R^2 .

Once an MLR-model is selected from the alternatives, the included predictors can be verified. Predictors that are not significant should be deleted from the model one by one, starting with the least significant. A new model is fitted each time a predictor is deleted and compared to the old model with the log-likelihood ratio test. Careful attention should be paid to predictors whose coefficient have changed markedly after another predictor is removed, indicating that the deleted predictor is a confounder of other predictors (Hosmer and Lemeshow, 2000). A strong confounder should be kept in the model, even when the predictor is not significant. Next the odds ratios of the predictors are checked for pedological consistency.

(6) Checking the assumption of linearity in the logit

Logistic regression assumes a linear relationship between continuous predictors and the logit. The Box-Tidwell approach and logit graphs were used to test this assumption. Box-Tidwell adds the transformed predictor $x \ln(x)$ to the model, where x is the value of the predictor (Hosmer and Lemeshow, 2000). Statistical significance of this predictor suggests non-linearity in the logit. The logit graph approach replaces the continuous predictor with a categorical predictor with four levels using the quartiles as cut-points. Estimated coefficients of this predictor are plotted against the midpoints of the quartiles. Non-linear plots indicate non-linearity in the logit. The relationship shown by the graph should be pedologically plausible, as before.

(7) Checking for interactions between predictors

To check whether interactions between predictors should be included in the MLR-model, pairwise interactions are created for each possible combination of predictors or only for those predictors the modeller expects to interact. The stepwise forward method was used to select interactions from all possible combinations of predictors. Interactions are tested for significance with the likelihood ratio test. Significant interactions are included unless these are not pedologically plausible or cause numerical instability. Goodness-of-fit statistics are used to check if the model fit improved.

(8) Statistical and visual assessment of the final model

Statistical assessment of the final MLR-model is based on the of goodness-of-fit mea-

asures as described in step 5. If the model is judged statistically acceptable then the model is applied to create a preliminary soil map. If unrealistic soil patterns are found the model should be adjusted. This means a return to step 4.

2.2.5 Model Application

The ten calibrated MLR-models were used to estimate the probabilities of occurrence of the ten soil types at the nodes of a 25-m raster covering Drenthe. The soil type with the largest probability at each grid node was used to construct a prediction map. The theoretical purity was computed as the mean of the maximum probability at each grid cell of the prediction grid (Brus et al., 2008). Prediction uncertainty was quantified by Shannon entropy:

$$H_z = - \sum_{i=1}^{n_z} \hat{\pi}(z_i, \mathbf{s}) \log_{n_z} \hat{\pi}(z_i, \mathbf{s}) \quad (2.8)$$

where $\hat{\pi}(z_i, \mathbf{s})$ is the estimated probability that random variable Z at location \mathbf{s} takes the value z_i , and n_z is the number of outcomes (Brus et al., 2008). By using the logarithm with base n_z the maximum entropy is 1, which occurs when all outcomes have equal probability. The minimum value for the entropy is zero, which occurs when there is no uncertainty and one of the outcomes has probability one. It should be noted that the entropy indicates whether the predicted soil type has a large probability, it does not indicate that the prediction itself is correct. The accuracy of the predicted soil types was validated with independent data.

2.2.6 Model Validation

Sampling strategy

The predictive soil map was validated with independent data collected by stratified simple random sampling (de Gruijter et al., 2006). Strata were obtained by overlaying the generalized national soil map, henceforth referred to as the *reference map*, with a map depicting three regions that roughly coincide with the major drainage basins and the areas with 1:10 000 soil maps. The latter map improves the spreading of the sample locations over the study area and facilitates separate estimation of purity for the subareas with high and low density of calibration data. This resulted in 34 strata. A total of 150 locations were allocated to the strata in proportion to their area, with a minimum of two per stratum to allow estimation of the sampling variance for each stratum. Locations where permission was denied or proved otherwise impossible to sample were replaced with locations from a reserve list.

Statistical inference

An indicator variable was derived that takes value 1 if mapped soil type equals observed soil type and 0 else. Several accuracy measures were estimated from the validation sample. The first is the overall purity (or global purity) of the reference map and updated map. The overall purity is defined as the proportion of the mapped area in which the predicted soil type, which is the soil type as depicted on the map, equals the true soil type. The purity was also estimated for each of the soil strata and 'map scale' strata separately. In addition, the map unit purities and class representations of the updated map were estimated. The map unit purity is the proportion of the map unit correctly classified. Class representation of soil type k is the proportion of the area where the actual soil type k occurs that is also mapped as type k .

The overall purity is estimated by (de Gruijter et al., 2006):

$$\hat{f} = \sum_{h=1}^l w_h \hat{f}_h \quad (2.9)$$

where w_h is the weight (relative area) of stratum h , \hat{f}_h is the estimated areal fraction of stratum h correctly classified, and l is the number of strata. The stratum fractions were estimated by the fraction correctly predicted locations in each stratum since the validation locations in each stratum were selected by simple random sampling:

$$\hat{f}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i \quad (2.10)$$

where n_h is the number of sampling locations in stratum h , and y_i is the indicator variable at sampling location i . The map unit purities and class representations of the updated map were estimated by the ratio-estimator (de Gruijter et al., 2006). This estimator must be used to obtain accuracy estimates for so-called *domains* (sub-areas of interest) that do not coincide with the sampling strata. The purity of map unit k of the updated soil map is estimated by:

$$\hat{p}^k = \frac{\sum_{h=1}^l A_h \bar{y}_h^k}{\sum_{h=1}^l A_h \bar{x}_h^k} \quad (2.11)$$

where A_h is the area of stratum h , \bar{y}_h^k is the sample mean of indicator $y_{i,h}^k$ taking value 1 if mapped and observed soil type at sampling location i equal soil type k and 0 else, and \bar{x}_h^k is the sample mean of indicator $x_{i,h}^k$ taking value 1 if the mapped soil type equals soil type k and 0 else. Class representation of soil type k is estimated by:

$$\hat{c}^k = \frac{\sum_{h=1}^l A_h \bar{y}_h^k}{\sum_{h=1}^l A_h \bar{z}_h^k} \quad (2.12)$$

where \bar{z}_h^k is the sample average of the indicator $z_{i,h}^k$ taking value 1 if the observed soil type equals soil type k and 0 else.

Eqs. 2.9 and 2.10 were also used to compare the predictive capabilities of the updated soil map and reference map by substituting y_i for $d_i = y_i^u - y_i^r$, the difference between the indicators for the updated (y_i^u) and for the reference map (y_i^r). This variable can have values -1, 0, and 1 and is used to estimate \hat{d} , which is the mean difference in actual purity of the updated and reference maps. Under the null hypothesis that the expected value of the estimated mean difference is zero, it was assumed that \hat{d} follows a normal distribution with zero mean and variance $var(\hat{d})$. It should be noted that absolute purity differences are reported when comparing purities of soil maps or map units in this chapter and in Chapters 5 and 6.

2.3 Results

2.3.1 Model building

The model building framework described in Section 2.2.4 was applied to each of the ten map units of the soil map of Drenthe. This section describes the results of the model building process for map unit *mPY*, which are shallow peat soils with a mineral surface horizon.

(1) Definition of a conceptual model of pedogenesis

Map unit *mPY* (36 000 ha) is the second largest of Drenthe. The national soil map subdivides this unit in *iW*, *zW* and *kW*, which have different topsoils due to different soil forming processes. Map unit *kW*, covering 200 ha in the northern tip of Drenthe, has a clayey topsoil that is formed by deposition of marine clay on peat in the brook valleys. The topsoils of units *iW* (22 000 ha) and *zW* (14 000 ha) are of anthropogenic origin. The spatial extent of *iW* is limited to the cultivated peatlands. The topsoil is formed by repetitive mixing of the sand cover, applied after peat excavation, with slivers of peat from the subsoil. Resulting topsoils are spatially highly variable in thickness (15-40 cm) and organic matter content (10-25%). Map unit *zW* is found in small areas within the peat colonies, along the edges of brook valleys or in depressions on the Drenthe plateau. The sandy topsoil can be formed by (1) cultivation by application of sand-rich manure, (2) leveling of the irregular surface of the Drenthe plateau during agricultural reclamation, or (3) sand application on the peaty surface to improve trafficability. The *zW* topsoil is spatially less heterogeneous than the *iW* topsoil and its organic matter content is on average 5–15%. All soils in map unit *mPY* have a sandy subsoil that may contain a podzol, depending on the position in the landscape. A podzol-B horizon in the subsoil is generally found at higher positions.

The frequency distribution of observed soil types in map unit *mPY* shows that at only 30% of the locations a shallow peat soil with a mineral topsoil was found. The podzol now is the most common soil type and is observed at 42% of the locations. At these locations the peat layer has disappeared as a result of oxidation. Oxidation rate depends on several factors. These include land use, because oxidation rate is faster under arable land than under grassland or natural vegetation; groundwater level, because oxidation rate increases as the groundwater level decreases; and peat type, because mesotrophic peat is less resistant to oxidation than oligotrophic peat. Soils of the cultivated peatlands are better drained and under more intensive agricultural use than peat soils in the brook valleys. Therefore soils in the peat colonies are expected to be more strongly affected by oxidation than soils in the brook valleys. Where peat has disappeared, earth soils (*ES*) can also be found.

Not all impurities in map unit *mPY* can be explained by peat oxidation. Part of the inclusions were present from the beginning, due to generalization errors. Confusion of soil types close to boundaries of map delineations is expected to be larger than in the centre of the delineations due to the positional accuracy of the delineations. It was therefore assumed that the probability of occurrence of soil types within an impure map delineation is also governed by the soil types of adjacent map delineations and by soil types that dominate the direct neighbourhood of a location. Generalization errors are also caused by large short-scale soil variability in the cultivated peatlands, which cannot be adequately expressed at the 1:50 000 map scale. Because of this variability, shallow peat soils (*PY*), shallow peat soil with mineral topsoil (*mPY*), deep peat soils (*P*) or deep peat soils with mineral topsoil (*mP*) can all occur in areas smaller than the minimum delineation size. Inclusion of podzols or earth soils can be found at higher and drier positions in the peat colonies and brook valleys, such as coversand ridges. These geomorphological features are in general too small to be mapped at the 1:50 000 scale.

(2) *Collection of predictors from available environmental ancillary data*

The soil forming processes and factors that cause inclusions of soil types other than *mPY* were represented by a set of 46 predictor variables (Table 2.2). These are:

1. *Land cover*. The effect of land cover on peat oxidation is represented by data layers 'recent land cover' and 'historic land cover'. Five indicator predictors were derived from both layers.
2. *Groundwater*. The effect of groundwater level on peat oxidation is represented by data layers GD, GT, GD.MHW, and GD.MLW. Three indicator predictors were derived from each layers. An ordinal categorical predictor with three levels was derived from GD and GT.
3. *Peat type*. Peat type is proxied by subsoil type as described by the soil map Finke et al. (1996). If a podzol-B horizon is present in the subsoil then it was assumed that the peat is of oligotrophic origin otherwise it was assumed that the peat is of mesotrophic origin.

4. *Oxidation risk*. Finke et al. (1996) mapped peat oxidation risk (high-low) for two map sheets of the soil map of Drenthe by combining groundwater and peat type data. Two predictors representing oxidation risk were created, one using groundwater data from the GT layer, and one using the groundwater data from the GD layer.
5. *Topsoil lithology*. One indicator predictor was derived from the soil map to represent the topsoil type.
6. *Landscape*. The extent of the cultivated peatlands was delineated from the soil and geomorphology maps. The paleogeography map was used to delineate the former highmoor landscape and the brook valley system.
7. *Elevation*. Elevation was used to map out inclusions of soil types *PZ* and *PY*.
8. *Relative elevation*. Four relative elevation layers captured local variation in elevation and can be used to identify for example local depressions or coversand ridges.
9. *Proximity to boundary of map delineations*. Two maps were generated from the soil map, indicating whether a pixel was located within 125 m or within 250 m from the boundary of the map delineation.
10. *Neighbouring soil type*. The soil type of the nearest neighbouring delineation was determined for each pixel within map unit *mPY*. The resulting map was reclassified twice based on peat thickness class.
11. *Dominant soil type*. The dominant soil type within 125, 250, and 500 m radius was determined for each pixel within map unit *mPY*. The resulting map was reclassified twice based on peat thickness class.

(3) Univariate analysis of candidate predictors

BIS contained 2 894 soil profile observations within map unit *mPY*. Each of the ten soil types is observed at least once in the map unit. Brown forest soils are observed two times and till soils three times. These two soil types were eliminated as outcome level because there were not enough observations to fit the logit functions. This implies that the probability of occurrence of these soil types in map unit *mPY* was set to zero.

Each cross-tabulation of a categorical predictor with the response variable resulted in a significant Pearson chi-square statistic. Furthermore, cross-tabulations showed that response outcome 'plaggen soil' (seventeen observations) had zero cell frequencies for several predictors (2.2). To reduce the number of candidate predictors the predictors that showed the strongest association to the response variable were selected from variable groups 'groundwater', 'recent land cover', 'historic land cover' and 'soil map'. This selection resulted in nineteen categorical and five continuous predictors (Table 2.2). A univariate MLR-model was fitted for each of the selected predictors, with soil type *mPY* as reference level. The likelihood ratio test was significant for each univariate model, indicating that all predictors are candidates for the

Table 2.2: Candidate predictors for map unit mPY.

Variable group	Description	Codes/Unit	Predictor name	Associated predictors of other groups
1	Elevation*	Absolute elevation	cm a.s.l.	ELEV
2	Relative elevation			
	search radius 250m*	cm	RELELEV250	
	search radius 500m*	cm	RELELEV500	
	search radius 750m*	cm	RELELEV750	
	search radius 1 000m*	cm	RELELEV1000	
3	Groundwater			
	GD* [†]	1=Wet,2=Moist,3=Dry	GD	PEATOX_GD PEATOX_GT PEATOX.GD
	GD wet* [†]	1=Yes, 0=No	GD.W	
	GD moist	1=Yes, 0=No	GD.M	
	GD dry	1=Yes, 0=No	GD.D	
	GD.MHW wet	1=Yes, 0=No	MHW.W	
	GD.MHW moist	1=Yes, 0=No	MHW.M	
	GD.MHW dry	1=Yes, 0=No	MHW.D	
	GD.MLG wet	1=Yes, 0=No	MLG.W	
	GD.MLG moist	1=Yes, 0=No	MLG.M	
	GD.MLG dry	1=Yes, 0=No	MLG.D	
	GT*	1=Wet,2=Moist,3=Dry	GT	PEATOX.GT PEATOX.GT
	GT wet*	1=Yes, 0=No	GT.W	
	GT moist	1=Yes, 0=No	GT.M	
	GT dry	1=Yes, 0=No	GT.D	
4	Land cover, 1997–2003			
	Permanent grassland*	1=Yes, 0=No	LC.GR	
	Permanent cropland*	1=Yes, 0=No	LC.CR	
	Gras-crop rotation	1=Yes, 0=No	LC.GRCR	
	Gras-crop, cropland*	1=Yes, 0=No	LC.ROTCR	
	Nature	1=Yes, 0=No	LC.NAT	
5	Land cover, 1900			
	Grassland*	1=Yes, 0=No	HLC.GR	
	Cropland	1=Yes, 0=No	HLC.CR	
	Heath*	1=Yes, 0=No	HLC.HEATH	
	Forest	1=Yes, 0=No	HLC.FOR	
	Nature*	1=Yes, 0=No	HLC.NAT	
6	Paleogeography			
	Brook valley system* [†]	1=Yes, 0=No	BROOKVAL	PEATTYPE
	Former highmoor area*	1=Yes, 0=No	HIGHMOOR	PEATTYPE
7	Geomorphology-soil map			
	Cultivated peatland*	1=Yes, 0=No	CULPEAT	SOILCOV GT, GT.W
8	Soil map			
	Peat type*	1=Oligotrophic 0=Mesotrophic	PEATTYPE	BROOKVAL HIGHMOOR
	Topsoil lithology*	1=Cult. peatl. 0=Sandy\clayey	SOILCOV	CULPEAT
	Distance to boundary			
	<125m* [†]	1=Yes, 0=No	DIST125	
	<250m	1=Yes, 0=No	DIST250	
	Nearest neighbour			
	2 levels	1=Peat soil 0=Mineral soil	NEIGHB.2L	
	3 levels* [†]	1=Deep peat soil 2=Shallow peat soil 3=Mineral soil	NEIGHB.3L	

continued on next page

Table 2.2: Candidate predictors for soil map unit *mPY*. (continued)

Variable group	Description	Codes/Unit	Predictor name	Associated predictors of other groups
	Domin. soil 125m radius			
	2 levels	see nearest nghb	MS125.2L	
	3 levels	see nearest nghb	MS125.3L	
	Domin. soil 250m radius			
	2 levels	see nearest nghb	MS250.2L	
	3 levels	see nearest nghb	MS250.3L	
	Domin. soil 500m radius			
	2 levels	see nearest nghb	MS500.2L	
	3 levels*	see nearest nghb	MS500.3L	
9	Groundwater-soil map			
	Oxidation risk using GD [†]	1=High, 0=Low	PEATOX_GD	GD, GD.W
	Oxidation risk using GT	1=High, 0=Low	PEATOX_GT	GD, GT GT.W

* Predictors selected after univariate analysis.

† Outcome soil type *PS* has a zero cell frequency for one of the levels of the predictor.

multivariate model. The odds ratios were generally in accordance with the presented knowledge on the soil-landscape system.

(4) Multivariate analysis of selected candidate predictors

A multicollinearity assessment confirmed the assumption that predictors within variable groups are associated. Furthermore, moderate and strong associations were found between predictors from different groups (Table 2.2). Model selection started with all 24 predictors. Again soil type *mPY* was used as reference level. All predictors except GD, GT.W and HLC.GR were selected resulting in a model that showed strong multicollinearity effects, evidenced by highly inflated coefficients and standard errors for several predictors. To eliminate the multicollinearity effects the least significant predictor of variable groups elevation, recent land cover and historic land cover were removed from the model, followed by the second least significant and so on until the two strongest predictors within these groups remained. With these predictors four competing MLR-models (Models 1–4, Table 2.3) were selected. Model 1 resulted after selecting the most significant predictor from each variable group. Models 2 to 4 are competing models in which one of the competing predictors is substituted for the other competing predictor of the same variable group. Because none of these models showed effects of multicollinearity, it was decided to keep the weakly and moderately associated predictors that belong to different variable groups in the model. Four predictors showed signs of numerical instability for the logit function of outcome level *PS*. Because eliminating an outcome level is at first less preferable than omitting predictors, an MLR-model was fitted without the predictors that caused numerical instability for outcome level *PS* (Model 4*, Table 2.3).

(5) Evaluation of adequacy of the multivariate model(s)

Summary measures of goodness-of-fit were calculated for each of the four competing multivariate models (Models 1-4, Table 2.3) and proved to be very similar. Because Model 4 performed slightly better for Pearson Chi-squared and calibration purity,

Table 2.3: Competing MLR-models with their goodness-of-fit measures. Competing predictors are indicated in bold type.

	Model 1	Model 2	Model 3	Model 4	Model 4*
<i>Variable group</i>					
1	ELEV	ELEV	ELEV	ELEV	ELEV
2	RELELEV250	RELELEV1000	RELELEV250	RELELEV250	RELELEV250
3	GT	GT	GT	GT	GT
4	LC_ROTCTR	LC_ROTCTR	LC_GR	LC_ROTCTR	LC_ROTCTR
5	HLC_HEATH	HLC_HEATH	HLC_HEATH	HLC_NAT	HLC_NAT
6	BROOKVAL	BROOKVAL	BROOKVAL	BROOKVAL	BROOKVAL
7	CULPEAT	CULPEAT	CULPEAT	CULPEAT	CULPEAT
8	PEATTYPE	PEATTYPE	PEATTYPE	PEATTYPE	PEATTYPE
	SOILCOV	SOILCOV	SOILCOV	SOILCOV	SOILCOV
	DIST125	DIST125	DIST125	DIST125	DIST125
	NEIGHB_3L	NEIGHB_3L	NEIGHB_3L	NEIGHB_3L	NEIGHB_3L
	MSOIL500_3L	MSOIL500_3L	MSOIL500_3L	MSOIL500_3L	MSOIL500_3L
9	PEATOX_GD	PEATOX_GD	PEATOX_GD	PEATOX_GD	PEATOX_GD
<i>Goodness-of-fit measure</i>					
Pearson- χ^2 (df)	19 592 (20 034)	19 824 (20 041)	19 636 (20 034)	18 716 (20 062)	24 119 (20 027)
McFadden- R^2	0.13	0.12	0.13	0.13	0.10
Calibration purity	48.4	48.0	48.1	48.4	47.6

this model was selected for the next steps in the model-building process. Model 4* performed worse than Model 4. It was therefore decided to eliminate *PS*, the plaggen soil, as outcome level because this soil was observed at only seventeen locations within map unit *mPY*. A pedological justification is that plaggen soils are unlikely to occur in map unit *mPY* as they are characteristic for the open field farming system found on coversand ridges and not for the cultivated peatlands. The MLR-model fitted without this outcome level contained the same predictors as Model 4.

The number of significant predictors differed between the six logit functions (Table 2.4). Predictors SOILCOV, HLC_NAT and NEIGHB_3L were not significant for five logit functions. Since the likelihood ratio test is significant for SOILCOV and SOILCOV is a pedologically important predictor, it was decided to retain this predictor in the model. The likelihood ratio test for HLC_NAT is not significant. Furthermore, HLC_NAT is only a moderately strong confounder of one coefficient in the logit of soil type *PZ*. HLC_NAT was therefore omitted from the model. NEIGHB_3L contributes significantly to the model and is a strong confounder of other predictors, in spite of five non-significant model coefficients. Collapsing NEIGHB_3L to two levels improves the Wald statistic: the coefficient of the binary predictor is significant in three logits. However, the likelihood ratio test suggests that the model with the three-level predictor performs better than the model with the binary predictor so NEIGHB_3L was retained. Predictors MSOIL500_3L, LC_ROTCTR and BROOKVAL were not significant for four logit functions. Collapsing MSOIL_3L into a binary predictor or omitting the predictor did not improve the model. BROOKVAL and LC_ROTCTR were kept in the model for their pedological significance although the

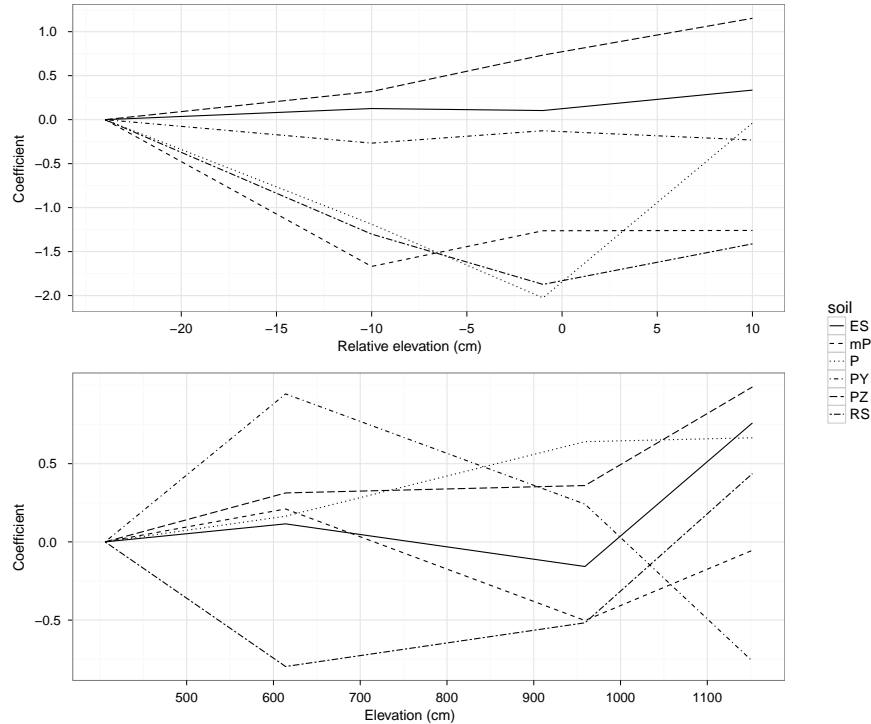


Figure 2.2: Logit graphs for predictors RELELEV250 (top) and ELEV (bottom).

likelihood ratio test of BROOKVAL did not confirm its importance. Like the odd ratios of the univariate model, the odd ratios of the multivariate MLR-model generally agree with the conceptual model of pedogenesis.

(6) *Checking the assumption of linearity in the logit*

So far the continuous predictors ELEV and RELELEV250 were treated as linear in the logit. The coefficients of both Box-Tidwell transformed predictors were not significant for four of six logit functions. The logit graphs for RELELEV250 (Fig. 2.2) show that this predictor is linear in the logit of soil types PZ, ES and RS, is somewhat linear in the logits of mP and PY, and is non-linear in the logit of P. Both logit graph and Box-Tidwell transformation for RELELEV250 suggest that this predictor can be treated as linear in the logit. The logit graph for ELEV (Fig. 2.2) shows that this predictor is linear in the logit of outcome levels PZ and P, and non-linear in the logit of mP and ES. The logit graphs of PY and RS show linearity between the second, third and fourth quartiles. The results of the logit graphs and Box-Tidwell transformation for ELEV do not convincingly support linearity in the logit, nor do they rule it out. As the fit of the model with ELEV as continuous predictor was much better than the fit with ELEV as categorical predictor, ELEV was retained as continuous.

(7) *Checking for interactions between predictors*

Three interactions remained after excluding interactions that were not pedologically plausible, that caused numerical instability or that did not improve model fit. Interactions between BROOKVAL and RELELEV250, LC_ROTCLR and RELELEV250 were statistically significant and pedologically plausible and were added to the model. Table 2.4 presents the final MLR-model for spatial prediction within map unit *mPY*.

(8) *Statistical and visual assessment of the final model*

The results of the statistical assessment of the final MLR-model for map unit *mPY* are presented in Table 2.5. The deviance of the fitted model is 13% (McFadden- R^2) smaller than the intercept-only model, which predicts the most frequently observed outcome (*PZ*) at each calibration location. Calibration purity is 49%, which is 19% larger than that of the reference map. Statistical assessment of the other nine MLR-models showed that the models explain a substantial part of the variation within the soil dataset (Table 2.5). Overall calibration purity is 66%, which is 11% larger than that of the reference map. The gain is on average about 20% for the peat map units and 3% for the mineral map units.

The soil map for map unit *mPY* did not show unexpected patterns of soil types. The area of soil type *mPY* was, as expected, greatly reduced. Podzols were predicted at 62% of the map unit area. The MLR-models for map units *P*, *ES* and *TS* were adjusted after inspection spatial distribution of the predicted soil types.

2.3.2 Model application

Ten MLR-models were used to re-map soil distribution within the ten map units of the reference map. The resulting map is shown in Fig. 2.3. The general spatial pattern of soil distribution of the updated map resembles that of the reference map: the maps correspond for 68.5% of the area. Changes are most dramatic, as expected, for the peat map units. The area with peat soils declined with 34% (33 525 ha) compared to the reference map. Roughly 60% of the soils mapped as shallow peat soils are predicted to be transformed to mineral soils: the extent of the podzol soil type increased with almost 40 000 ha. Thirty-six percent of the deep peat soils with a mineral topsoil (*mP*), typical for the cultivated peatlands, are predicted to be transformed to shallow peat soils with mineral topsoil. Fig. 2.5 clearly shows these changes. Changes within map unit *P* are less severe: only 22% is predicted to be transformed to shallow peat soils. The reason for this is that soil type *P* primarily occurs in the brook valleys where peat layers are thicker and where conditions for oxidation are less favorable compared to the cultivated peat lands. The strong decline of soil type *TS* can be explained by the fact that most till soils occur in association with podzol soils on the reference map. The majority of the observations used to calibrate the model for map unit *TS* are classified as *PZ*, which results in *PZ* as the dominant predicted soil type in map unit *TS*. The area with plaggen soils, *PS*, is reduced with 32% compared to

Table 2.4: The estimated model coefficients (β) and odds ratios (OR) of the final MLR-model for map unit mPY.

Predictor	Logit function					
	PZ		mP		ES	
	β	OR	β	OR	β	OR
Intercept	-0.63		-3.61*		-4.26*	
NEIGHB_3L=1	-0.22*	0.81	1.12*	3.08	0.08	1.08
NEIGHB_3L=2	-2.04*	0.13	0.12	1.13	0.33	1.3
MSOIL500_3L=1	-0.83*	0.44	0.81*	2.24	-1.05*	0.35
MSOIL500_3L=2	-0.36*	0.70	0.26	1.30	-0.52*	0.60
DIST_125=0	-0.24*	0.79	-0.37*	0.69	0.31	1.37
SOILCOV=0	0.12	1.13	-0.40*	0.67	-0.47	0.62
GT=1	-0.02	0.98	-0.88*	0.42	0.87*	2.38
GT=2	0.10	1.11	-0.658	0.52	0.04	1.04
LC_ROTCCR=0	-0.23*	0.80	-0.07	0.93	-0.48*	0.62
PEATOX_GD=0	-0.49*	0.61	0.46*	1.58	-0.15	0.86
PEATTYPE=0	0.12	1.12	0.28	1.32	1.50*	4.49
CULPEAT=0	0.82*	2.27	0.75*	2.12	1.85*	6.39
BROOKVAL=0	0.42*	1.52	0.52	1.69	-0.12	0.88
ELEV	0.001*	1.001	0.001*	1.001	0.002*	1.002
BROOKVAL=0 x RELELEV250	0.02*	1.02	-0.01*	0.99	0.02*	1.02
BROOKVAL=1 x RELELEV250	0.05*	1.05	-0.03*	0.97	0.02*	1.02
RLC_ROTCCR=0 x RELELEV250	0.01	1.01	-0.02*	0.98	-0.01	0.99

Predictor	Logit function					
	P		PY		RS	
	β	OR	β	OR	β	OR
Intercept	-6.91*		-3.37*		-5.70*	
NEIGHB_3L=1	0.24	1.27	-0.17	0.84	0.25	1.29
NEIGHB_3L=2	-1.02	0.36	-0.53*	0.59	0.68	1.97
MSOIL500_3L=1	1.44*	4.20	0.10	1.11	-1.60	0.20
MSOIL500_3L=2	0.64	1.89	0.49*	1.63	-0.64	0.53
DIST_125=0	0.23	1.26	0.14	1.16	-1.88*	0.15
SOILCOV=0	-0.58	0.56	-1.41*	0.24	-1.21	0.30
GT=1	0.93*	2.54	1.80*	6.04	1.25	3.49
GT=2	-0.07	0.93	0.90*	2.45	0.54	1.71
RLC_ROTCCR=0	0.60*	1.83	0.00	1.00	0.73	2.08
PEATOX_GD=0	1.02*	2.77	0.36*	1.44	0.48	1.62
PEATTYPE=0	1.85*	6.33	0.58*	1.78	0.32	1.38
CULPEAT=0	-0.10	0.91	0.91*	2.47	1.87*	6.47
BROOKVAL=0	1.23*	3.43	0.81*	2.24	-0.98	0.38
ELEV	0.001	1.001	0.001*	1.001	0.002*	1.002
BROOKVAL=0 x RELELEV250	-0.03*	0.97	0.01	1.01	0.01	1.01
BROOKVAL=1 x RELELEV250	-0.07*	0.93	-0.02	0.98	0.02	1.02
RLC_ROTCCR=0 x RELELEV250	0.02	1.02	-0.01	0.99	0.01	1.01

* Wald statistic is significant at the 0.15 level.

Table 2.5: Statistical assessment of the ten final MLR-models. The R^2 is the McFadden- R^2 .

Model	R^2	Calibration purity
P	0.31	61
mP	0.28	62
PY	0.21	51
mPY	0.13	49
BF	0.19	63
PZ	0.21	79
ES	0.21	57
PS	0.30	63
TS	0.31	75
RS	0.49	83

the reference soil map. Affected areas are the edges of *PS* map delineations (Fig. 2.5) and the plaggen soils on the Hondrug. The former is explained by the decrease in thickness of the plaggen A-horizon from the centre of the open fields towards the edges: if the thickness does not exceed 30 cm, then the soil is not classified as plaggen soil. The latter is a direct result of the observations on the Hondrug, which were all located in a relative small area (intensively surveyed area 3, Fig. 2.1). Many profile observations within map unit *PS* were classified as brown forest soils (*BF*). This is also the reason for the strong increase in area of map unit *BF* on the updated soil map compared to the reference map.

The overall theoretical purity of the updated soil map is 67% (Table 2.6). In general the theoretical purity is smaller and the uncertainty is larger for the areas mapped as peat soils on the reference map than for the mineral soils. The areas with the smallest theoretical purity are the areas that were originally mapped as shallow peat soils (*PY*, *mPY*) and earth soils (*ES*). Predictions for these areas are also the most uncertain (Table 2.6). Map unit *mPY* is characteristic for the cultivated peatlands, whereas *PY* and *ES* are mainly found along brook valley sides. Fig. 2.4 shows high entropy values for these two parts of the landscape. The soil type pattern in the cultivated peatlands is very heterogeneous by itself and is further complicated by peat oxidation. This makes soil spatial prediction challenging in this area, which is evidenced by highly uncertain predictions. The brook valley sides are topographical transition zones where gradual changes in soil weaken relationships between soil types and predictors. Uncertainty associated to the predictions will be larger in such areas than in areas with stronger relationships such as in the centre of the brook valleys and the high parts of the plateau.

Fig. 2.6 depicts a catena of predicted and mapped soil types along a 1 500 m long transect from plateau through a brook valley to plateau, as well as the change of estimated probabilities and entropy of the main predicted soil types. The location of the catena is indicated by the arrow in Fig. 2.5. The typical catena in Drenthe has plaggen soils on top of the plateau, bordered by podzols or brown forest soils,

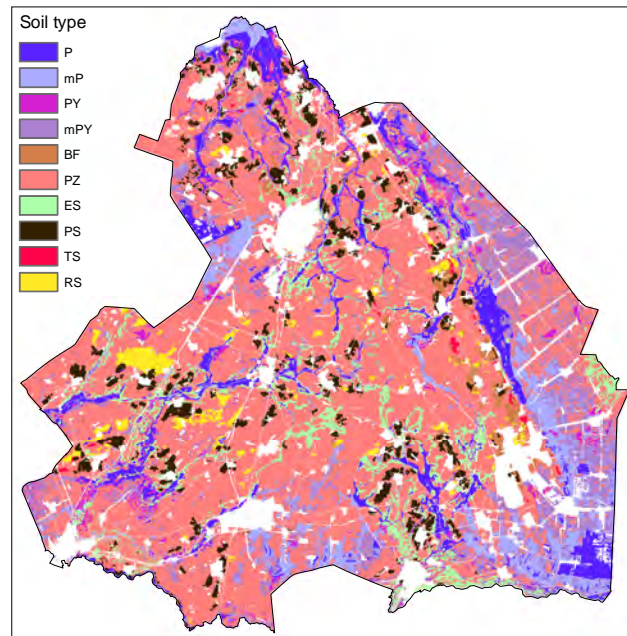


Figure 2.3: Updated soil map as predicted by the ten MLR-models.

depending on parent material. When going from plateau to brook valley one would typically encounter a gradual transition from podzols to earth soils to shallow peat soils to deep peat soils. The first difference between predicted and mapped soil types is the soil sequence from the northern plateau towards the brook valley. In the reference map plaggen soils border deep peat soils whereas the updated map shows a pedologically more realistic transition from plaggen soils to podzols to shallow peat soils to deep peat soils. The second difference is the prediction of podzols on a coversand undulation in the earth soil map unit at the southern side of the brook valley centre. These undulations are better drained than the surrounding, lower terrain, creating more favourable conditions for podzol formation. Earth soils are predicted at the sides of the undulation and podzols at the top, which is pedologically plausible. The catena was not validated but based on soil-landscape system knowledge the updated map shows a more realistic soil sequence along the transect than the reference map. The probability graph shows that the MLR-models do not have much difficulty in differentiating soil types at the topographical extremes: the lowest parts of the brook valleys, the plateaus and the top of the coversand undulation. The difference between largest and second largest estimated probability is relatively large. Small differences in estimated probabilities are found at topographical transition zones. At these locations the models easily confuse between soil types, which is evidenced by an increase in entropy at these zones.

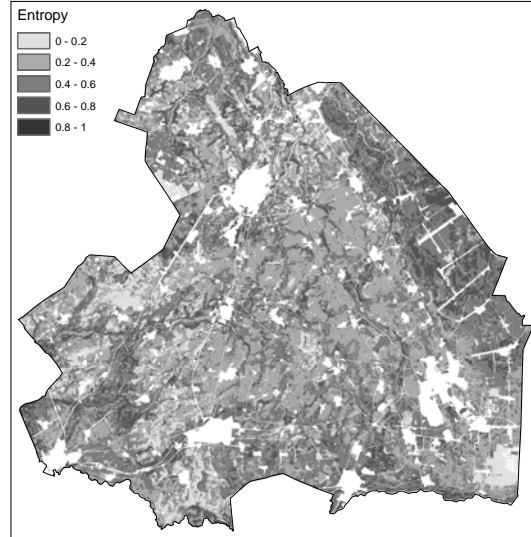


Figure 2.4: Entropy map that shows the uncertainty associated to the predictions.

Table 2.6: Theoretical purity and entropy of the updated soil map for the areas corresponding to the map units of the reference map, i.e. the areas for which the MLR-models were calibrated and applied, and for grouped map units peat (P–mP–PY–mPY) and mineral (BF–PZ–ES–PS–TS–RS).

	Theoretical purity	Entropy
Global	67	0.40
Map unit		
P	64	0.42
mP	63	0.43
PY	50	0.54
mPY	50	0.54
BF	83	0.21
PZ	79	0.34
ES	58	0.47
PS	66	0.38
TS	79	0.24
RS	77	0.24
Peat	57	0.48
Mineral	75	0.35

2.3.3 Model validation

Table 2.7 summarizes the validation results of the updated and reference soil map. The estimated actual purity of the updated map is 58%, which is 6% larger ($p =$

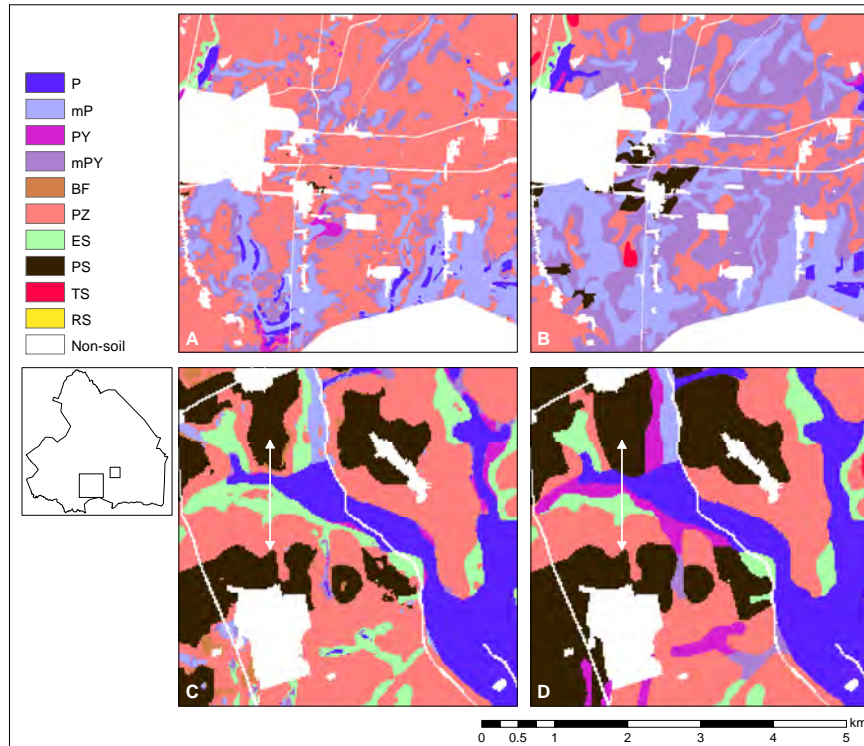


Figure 2.5: Details of the updated and reference soil maps: colonized peat updated map (A), colonized peat reference map (B), brook valley updated map (C), and brook valley reference map (D). The arrows in the bottom figures indicate the location of the north–south oriented catena shown in Fig. 2.6.

0.039) than the purity of the reference map. The purity is 9% smaller than the theoretical purity, possibly because the calibration locations are concentrated in the four areas with a detailed soil map (Fig. 2.1). Apparently, the calibrated relationships were unable to explain a similar amount of variation outside the four detailed survey area than within the four areas.

At the level of the soil strata soil spatial distribution is better represented by the MLR-model than by the reference map for soil strata *mP*, *PY*, *mPY*, *PS* and *T*. The largest increase is for stratum *PY* (35%, $p = 0.000$). Purity gains for strata *mP* (5.4%, $p = 0.318$) and *mPY* (14%, $P = 0.224$) are not significant due to the small numbers of validation locations, but they are pedologically relevant, especially the 14% purity gain for stratum *mPY*. The cause of the large purity increase of soil stratum *TS* is outlined in section 2.3.2. Purity gain of stratum *PS* is 10% ($p = 0.174$). There is no difference in actual purity between the updated and reference maps for soil strata

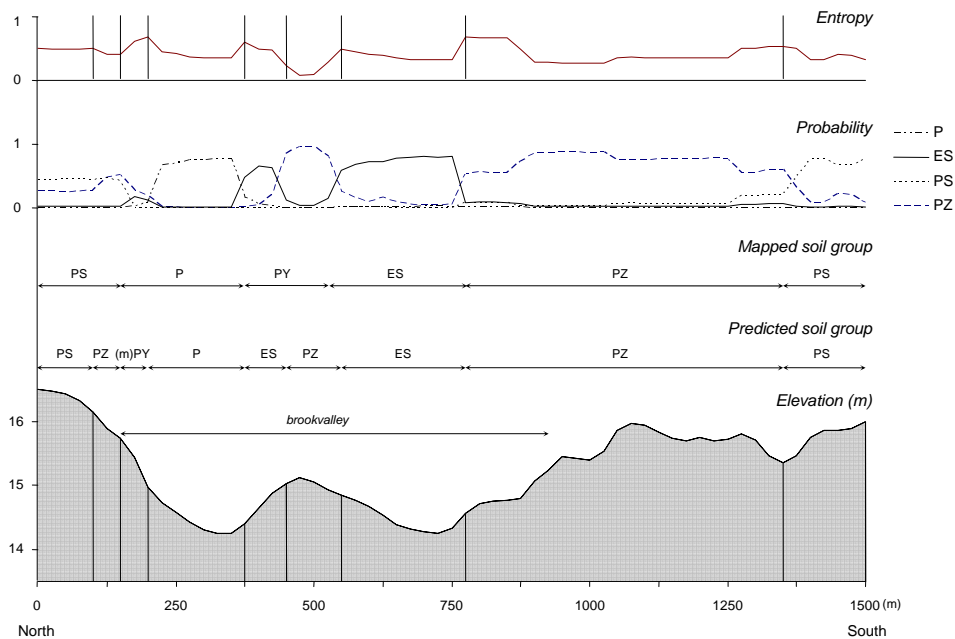


Figure 2.6: Change of the entropy, maximum probability, mapped and predicted soil types along a typical catena in the Drenthe landscape.

P, *PZ* and *RS*. The reference map better represents soil spatial distribution within strata *ES* and *BF*. The actual purity of both strata is 13% smaller ($p = 0.105$ for *ES*; $p = 0.159$ for *BF*) for the updated soil map than for the reference soil map. These figures indicate that the global increase in map purity of the updated map compared to the reference map is largely attributed to the increase in purity in the peat strata. The pooled purity increase for these strata is a pedologically relevant 11% ($p = 0.069$) whereas the pooled purity increase for the mineral strata is 2.4% ($p = 0.086$).

Table 2.8 shows the purities of the ten map units of the updated soil map. The MLR-models predict the spatial distribution of soil types *P*, *BF*, *PZ*, *PS* and *RS* fairly well, while map units *PY* and *TS* have purities close to 0. The large variation in purities can have several reasons. First, the effect of peat oxidation is underestimated: mineral soils were observed at three validation locations in map unit *PY* and at five validation locations in map unit *mPY*; shallow peat soils were observed at four validation locations in map unit *mP*. Second, the models have difficulty in predicting topsoil lithology of peat soils, which is to a large extent influenced by human activities and is not easily associated to environmental predictors. In map unit *PY*, soil type *mPY* was observed at three validation locations and soil type *P* was observed at

Table 2.7: Estimated purities of the two soil maps: overall purity (standard error) and the purities of the soil strata, grouped strata (peat–mineral) and mapping-scale strata.

	<i>n</i>	Updated map	Reference map
Overall	150	58.1 (0.04)	0.52.1 (0.04)
Soil stratum*			
P	15	45.2 (0.14)	45.2 (0.14)
mP	15	31.0 (0.11)	25.6 (0.08)
PY	9	39.5 (0.17)	4.6 (0.05)
mPY	22	49.5 (0.11)	35.9 (0.11)
BF	4	13.1 (0.13)	26.2 (0.00)
PZ	55	72.7 (0.06)	72.7 (0.06)
ES	10	42.2 (0.14)	55.4 (0.17)
PS	11	75.2 (0.15)	65.2 (0.10)
TS	4	97.5 (0.03)	–
RS	5	64.1 (0.32)	64.1 (0.32)
Peat	61	42.5 (0.07)	31.5 (0.06)
Mineral	89	69.8 (0.05)	67.4 (0.05)
Mapping-scale stratum			
Area with 1:10 000 map	26	72.4 (0.12)	
Area without 1:10 0000 map	124	56.4 (0.04)	

* Note that here the soil strata corresponds to the map units of the reference soil map. The MLR-models re-mapped the soil within these map units.

three validation locations in map unit *mP*. Third, soil types in topographical transition zones are easily confounded. In map unit *ES*, typical for such transition zones, earth soils, podzols and peat soils were observed. Fourth, sample size. The small number of validation locations in several map units results in highly uncertain purity estimates. An example is the purity of map unit *TS* that might be attributed to chance as only two validation locations were located in this map unit.

It is interesting to note that when the legend of the updated soil map is generalized to two map units (peat and mineral), the purity of the peat map unit would increase to 80%, while the pooled purity of the four separate peat map units is 43%. This indicates that the main confusion between predicted soil types within the peat map units is between the four peat soils. Thus at the locations where the MLR-models predict peat soils, it is very likely that a peat layer is present in the soil profile at these locations. However, the model is uncertain about the thickness of the peat layer and the topsoil lithology (peaty or mineral). The purity of the mineral map unit is 88%.

Table 2.8: Estimated map unit purities and class representations for ten soil types as depicted on the updated soil map.

Soil type	<i>n</i>	Map unit purity	Class representation
P	9	76.5	51.5
mP	10	28.0	55.5
PY	8	4.7	5.7
mPY	13	36.9	24.5
BF	5	71.0	39.7
PZ	82	67.1	89.6
ES	9	46.2	21.1
PS	8	79.6	47.4
TS*	2	0.0	–
RS	4	94.2	51.6

* Class representation not estimated because *TS* was not observed in the sample.

2.4 Discussion

2.4.1 Multinomial logistic regression for soil mapping

MLR is computationally a simple method compared to more demanding methods such as indicator kriging (IK) (Bierkens and Burrough, 1993) and Bayesian Maximum Entropy (BME) (Brus et al., 2008). It does not suffer from shortcomings of IK like probabilities that are outside the interval $[0,1]$ or probabilities that do not sum to 1. Nor is it as computationally demanding as BME. However, building a statistically and pedologically sound MLR-model requires careful attention as many choices have to be made, and interpretation (both statistical and pedological) of the MLR-model is not as straightforward as that of linear regression models. The framework based on (Hosmer and Lemeshow, 2000) proved a valuable guideline for building MLR-models, although the steps should be meticulously applied.

The main drawback of applying MLR to spatial data is that spatial autocorrelation during coefficient estimation and prediction is ignored. This may bias estimated effects of the predictors on the response variable. Hengl et al. (2007b) showed that MLR did not perform as well as methods that incorporate spatial autocorrelation in soil spatial prediction. Autologistic regression can account for spatial autocorrelation in the response variable and is a popular method in spatial ecology (Augustin, 1996; Smith, 1994). Unfortunately, the autologistic regression model can only handle binomially-distributed data. There are no examples that extend the autologistic model to the multinomial case. Another drawback is that certain structures in the calibration data, such as predictors that completely separate outcome levels or predictors that lack observations for one or more of its levels, cause non-existent maximum likelihood estimates or infinite odds ratio estimates (Hosmer and Lemeshow, 2000). This is a serious drawback of MLR as it limits the number of outcome levels

and the number of predictor levels. Even with only ten soil types outcome levels of the response variable were often forcibly omitted. Furthermore, predictors that cause complete separation or that have zero cell counts are in theory strong predictors as they suggest that certain soil types do not occur under certain conditions. Several candidate predictors that were highly correlated with the response variable could not be used as predictors because of complete separation or zero cell counts.

2.4.2 Soil spatial prediction

The most frequently observed soil types within a map unit are over-represented on the predicted soil maps (e.g. soil types *mPY* and *PZ* in map unit *mPY*), which was also reported by Hengl et al. (2007b). These authors argue that this is caused by weak association of the predictors with some of the soil types. The odds ratios of the predictors included in the logits of *PY* and *ES*, however, show that the several predictors are strongly related to soil types *PY* and *ES* (Table 2.4). There is no evidence that weak associations with less frequently observed soils cause over-representation of the most frequently observed soils. The predictors strongly influence the estimated probabilities of less frequently observed soil types, but apparently this is not enough to exceed the estimated probabilities of the most frequently observed soil types.

At the reason why the clustered distribution of the calibration locations causes the 9% discrepancy between actual and theoretical purities can only be hinted. The four survey areas can be regarded as reference areas used to obtain predictive relationships. These relationships are then extrapolated across the entire survey area. This resembles the digital soil mapping approaches presented by Lagacherie et al. (1995), Bui and Moran (2003) and Grinand et al. (2008). Lagacherie et al. (1995) state that the reference area approach works when similar soil forming processes act in the two areas, creating similar soil patterns. If soil forming processes were similar for the areas with and without a detailed soil map, as expected, then similar purities for these areas would be expected. But validation indicates that the actual purities of the two mapping scale strata differ (Table 2.7). The estimated actual purity of the updated map for the areas with detailed soil maps is 16% larger than for the area without the detailed maps. Note, however, that this estimate is very uncertain given the large standard errors of the purities. This might indicate that the modelled relationships are not so easily extrapolated across the province. Under the assumption that the natural soil forming processes within and outside the detailed survey areas are similar, the difference between theoretical and actual purities and the difference between purities within and outside the detailed soil survey areas might be attributed to human influence on soil formation. Since peat soils are much more sensitive to human interventions in the landscape than mineral soils, it would be expected that the discrepancy between theoretical and actual purity to be larger for peat soils than for mineral soils. This is supported by the results in Tables 2.6 and 2.7. For the pooled peat soils this discrepancy is 14% and for the mineral soils 5%.

The influence of human activities on soil distribution might also partly explain the 27% difference in overall purity of the peat soil-strata compared to mineral soil-strata (Table 2.7). Human influence can sometimes be proxied by predictors, for example human influence on peat oxidation can be proxied by land cover and groundwater. In many cases, however, this influence is hard to proxy with biophysical data. For example, the decision to apply a sand cover on a peaty topsoil to improve trafficability is made on the scale of individual agricultural fields. Furthermore, the topsoil of the soils in the cultivated peatlands exhibits large short-distance variation as a result of the reclamation method used. Peat soils with organic and mineral topsoils occur in association. This means that topsoil lithology of peat soils is not easily predicted. If topsoil lithology would be disregarded, then the actual purity of the peat soil-stratum of the updated soil map (Table 2.7) would increase from 42.5% to 56.6% and global overall purity from 58.1% to 64.1%.

2.4.3 Legacy soil data

The utility of legacy soil data for updating soil maps in landscapes with highly dynamic soils such as peat soils, greatly depends on the age of legacy data. Roughly 60% of the calibration locations located in the four detailed survey areas was already twelve years old at the time this study was carried out. The other 40% was between three and six years old. Twelve-year-old observations on peat soils are also becoming outdated as peat oxidation continues. Omitting these data would have greatly reduced the size of the dataset for calibrating the models for the peat map units and hence was not an option. Furthermore, the soil survey for the reference soil map was conducted at least fifteen years before the detailed soil surveys that yielded almost all calibration observations. This means that these observations might still contain useful information for updating. Nevertheless, use of decade-old observations on peat soils for updating might overestimate the area of (shallow) peat soils on the predicted soil map, although this is not strongly supported by the validation. Age of the calibration observations might also have contributed to the difference between theoretical and actual overall purity.

2.4.4 Validation of soil maps

Accuracy assessment of soil maps is imperative for any soil mapping study; conventional or digital. Commonly used statistics to quantify accuracy of soil type maps are the purity and the kappa index (Grinand et al., 2008; Hengl et al., 2007b; Li and Zhang, 2007). Another statistic for accuracy assessment that provides valuable information is the class representation of the soil types depicted on the map. This statistic is often used in image classification studies (Foody, 2002) (reported as producer's

accuracy or sensitivity) but is hardly reported for (digital) soil maps. Class representation of the updated soil map are presented in Table 2.8. An example of the merit of this statistic is given by map unit 'deep peat soils' (P). This map unit has a high purity (77%), which tells the user that soil type P is found at 77% of the area predicted as soil type P . However, the class representation of soil type P is 52%, meaning that only 52% of the true area of soil type P is mapped as P . One can question the usefulness of a soil map, for example for a carbon stock inventory, if the area with peat soils is heavily underestimated, which is not indicated by the purity. Class representation for peat soils (disregarding topsoil lithology and thickness) as depicted by the updated map is 72%.

Pedological knowledge was used for model-building. Such knowledge could also be exploited for error analysis or validation of the model. One could argue if predicting a site which is P as mP (two different deep peat soils) is equally wrong as predicting it as PZ (a mineral soil) and vice versa. Whether all errors are equal or different, clearly depends on the application for which a soil type map is used. Predicting a site which is P as mP would not have a large impact on estimates of carbon stock but it would have for the organic matter content of the topsoil; while predicting a site which is mP as PZ would have a large effect on carbon stock estimates but not on the topsoil organic matter content. Pedological knowledge was not used for estimation of the validation statistics or for the evaluation of the classification tables, but it would be worthwhile to investigate this in future mapping studies.

2.5 Conclusions

Legacy soil data in combination with high-resolution environmental ancillary data can be used to update a soil map. The updated soil map had 6% larger purity than the existing soil map. Updating proved to have more effect for the peat map units than for the mineral map units. The presented framework provides a systematic approach for building MLR-models and allows integration of expert knowledge during model selection and evaluation. When calibration locations are clustered in small areas within the survey area such as in this study, the transferability of the calibrated relationships might be limited, especially when there is strong human influence on soil development. This might lead to overestimation of the theoretical map purity, which emphasizes the importance of validation of soil maps with an independent probability sample.



Chapter 3

Pedometric mapping of soil organic matter using a soil map with quantified uncertainty

In this chapter three models are compared that use soil type information from point observations and a soil map to map the topsoil organic matter content for the province of Drenthe in the Netherlands. The models differ in how the information on soil type is obtained: Model 1 uses soil type as depicted on the soil map for calibration and prediction; Model 2 uses soil type as observed in the field for calibration and soil type as depicted on the map for prediction; and Model 3 uses observed soil type for calibration and a pedometric soil map with quantified uncertainty for prediction. Validation with independent probability sample data showed that Model 3 outperformed Models 1 and 2 in terms of mean squared error. However, Model 3 over-estimated the prediction error variance and so was too pessimistic about prediction accuracy. Model 2 performed the worst: it had the largest mean squared error and the prediction error variance was strongly under-estimated. This confirmed that calibration on observed soil type is only valid when uncertainty about soil type at prediction locations is explicitly accounted for by the model. It is concluded that whenever information about uncertainty associated to the soil types depicted on a soil map is available, and both soil property and soil type are observed at sampling locations, Model 3 can be an improvement over the conventional Model 1.

Based on: B. Kempen, G.B.M. Heuvelink, D.J. Brus, J.J. Stoorvogel
European Journal of Soil Science 61 (2010): 333–347

3.1 Introduction

Soil maps typically describe the spatial variation of soil types, but are also an important source of information on spatial variation of quantitative soil properties (Bregt and Beemster, 1989; van Meirvenne et al., 1994; Liu et al., 2006). A commonly used method for spatial prediction of soil properties is to use a 'representative' value for each map unit of the soil map (Leenhardt et al., 1994; Brus et al., 1996; Voltz et al., 1997; Webster and Oliver, 2007). These values are obtained from soil profile descriptions in soil survey reports or are provided by experienced soil surveyors. Goovaerts and Journel (1995) pointed out two weaknesses of this mapping approach: i) it results in abrupt transitions in the predicted value from one soil type to another, which are unrealistic in landscapes where lateral changes in soil are gradual and ii) it ignores spatial variation of the soil property within the map units. Marsman and de Gruijter (1986) found that within-map unit variation of soil properties ranged between 65% and 80% of the total variation in a 1 600 ha sandy area in the Netherlands.

Several studies showed that combining maps of soil type with soil property information from point observations can improve spatial prediction of soil properties compared with prediction from a soil map only (Voltz and Webster, 1990; Heuvelink and Bierkens, 1992; Goovaerts and Journel, 1995; Brus et al., 1996; Heuvelink, 1996; Oberthür et al., 1999). Typically, a regression model is calibrated by using mapped soil type as predictor followed by kriging of residuals (Hengl et al., 2004; Liu et al., 2006). The strength of the relationship between soil property and mapped soil type depends on the accuracy of the soil type map. Large impurities in the map units might result in a weak relationship, even when the relationship between soil property and the observed 'true' soil type is strong.

When both soil property and soil type are observed at sampling locations, the regression model can be calibrated by using observed soil type (e.g. Meersmans et al., 2009). This can improve the regression because the relationship between soil property and observed soil type is not confounded by impurities in the soil map units. For spatial prediction, however, the calibrated relationship has to be applied to the soil map because the true soil type is unknown at unsampled prediction locations. Using hard information (observed soil type) for calibration and soft information (mapped soil type) for prediction has the undesirable consequence that the prediction error variance will be under-estimated because the uncertainty in the map units of the soil map is not accounted for by the model. Furthermore, a strong, informative, relationship between soil property and soil type is applied to a soil map that is less informative about the soil property. Thus to take advantage of a calibration on true soil type for spatial prediction, the true soil type at the prediction locations must be used. The true soil type is typically unknown but it can be represented with a probability model. When a probability model is available, then this model may be used

for spatial prediction of the soil property. Probability distributions of soil type can be obtained from soil maps produced using pedometric methods such as indicator kriging (Goovaerts and Journel, 1995; Oberthür et al., 1999), Bayesian maximum entropy (Brus et al., 2008) or multinomial logistic regression (Debelli-Gilo and Etzelmüller, 2009, and Chapter 2). Soil survey reports accompanying traditional soil maps often provide areal estimates of the soil types occurring within the map units (Soil Survey Division Staff, 1993). Such information may be used to define a frequency distribution for each map unit that can serve as a probability distribution at any location within that map unit.

In the present chapter a geostatistical model is presented that uses a soil type map with quantified uncertainty as a covariate for spatial prediction of a quantitative soil property. This map represents the true soil type at each location with a probability distribution. A case study in the province of Drenthe will illustrate the use of this model for spatial prediction of the organic matter content of the topsoil (0–30 cm). For this province a pedometric soil map is available with location-specific probability distributions of ten major soil types (Chapter 2), as well as a large data set with point observations on soil organic matter (SOM) and soil type. The prediction model is calibrated with soil type information from these point observations and is subsequently applied to the pedometric soil type map for spatial prediction of the SOM content. Model performance is validated with an independent probability sample and compared with results from i) the conventional model that uses mapped soil type (i.e. the soil type with the largest probability) for calibration and prediction and ii) the theoretically flawed model that uses observed soil type at sampling locations for calibration and the soil type as depicted on the map for prediction.

3.2 Theory

Let us assume the following model of spatial variation of soil property Z :

$$Z(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})\epsilon, \quad (3.1)$$

where $a(\mathbf{s})$ is the mean or trend at location \mathbf{s} , $b(\mathbf{s})$ is the standard deviation and ϵ is a zero-mean, second-order stationary spatially correlated (standardized) residual with unit variance. Note that in Eq. 3.1 the mean and standard deviation of Z are a function of the spatial coordinates \mathbf{s} , i.e. they vary in space.

Suppose that Z is measured at n locations within a two-dimensional spatial domain D , yielding values $z(\mathbf{s}_k)$, $k = 1, 2, \dots, n$. At each sampling location the residual is computed by subtracting mean $a(\mathbf{s}_k)$ from the measured soil property value $z(\mathbf{s}_k)$. The residual is standardized by division with the standard deviation $b(\mathbf{s}_k)$. Next the spatial structure of the standardized residuals is modelled with a variogram. The

sill of the variogram is known to be 1, and the remaining variogram parameters are estimated from the sample of residuals. Residuals at unsampled locations can then be predicted using simple kriging. Next the soil property Z at prediction location \mathbf{s}_0 is predicted by back-transformation of the kriged residual using $b(\mathbf{s}_0)$ and $a(\mathbf{s}_0)$ (Goovaerts, 1997):

$$\hat{z}(\mathbf{s}_0) = b(\mathbf{s}_0) \left[\sum_{k=1}^n \lambda_k \epsilon(\mathbf{s}_k) \right] + a(\mathbf{s}_0), \quad (3.2)$$

where $\hat{z}(\mathbf{s}_0)$ is the predicted soil property value at \mathbf{s}_0 , $\epsilon(\mathbf{s}_k)$ is the observed residual at sampling location and λ_k is the simple kriging weight assigned to the residual at location \mathbf{s}_k . The prediction error variance at \mathbf{s}_0 is calculated as:

$$\begin{aligned} V(\mathbf{s}_0) &= \text{var}(b(\mathbf{s}_0)[\hat{\epsilon}(\mathbf{s}_0) - \epsilon(\mathbf{s}_0)]) \\ &= b^2(\mathbf{s}_0) \text{var}[\hat{\epsilon}(\mathbf{s}_0) - \epsilon(\mathbf{s}_0)], \end{aligned} \quad (3.3)$$

where $\text{var}[\hat{\epsilon}(\mathbf{s}_0) - \epsilon(\mathbf{s}_0)]$ is the simple kriging variance.

Prediction of soil property Z at each location \mathbf{s}_0 requires estimates of $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$. I will now discuss three models for prediction of Z where these model parameters are a function of the categorical explanatory variable "soil type". The models differ in the source from which the soil type information is obtained to estimate $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$.

In this chapter it is assumed that $a(\mathbf{s}_0)$, $b(\mathbf{s}_0)$ and the variogram are estimated without error: $\hat{a}(\mathbf{s}_0)$ and $\hat{b}(\mathbf{s}_0)$ are substituted into Eqs. 3.2 and 3.3 as if these are the true values to obtain the prediction of soil property Z and the prediction error variance. Ignoring the estimation error of the model parameters simplifies modelling dramatically but generally is unrealistic and affects the model outcome. This assumption will be addressed further on in this chapter in section 3.4.

Model 1: Parameter estimation and prediction using a soil type information from a map

Suppose that a soil map is available that distinguishes M soil types, $m = 1, 2, \dots, M$, and that specifies the soil type $c_{map}(\mathbf{s})$ at each location $\mathbf{s} \in D$. If soil type as read from the soil map is the only predictor of soil property Z , then $a(\mathbf{s})$ is a stepwise function with levels equal to the number of soil types. The mean of a soil type is estimated from all observations contained in the map unit (Webster and Oliver, 2007). Thus, the mean at prediction location \mathbf{s}_0 is estimated by :

$$\hat{a}(\mathbf{s}_0) = \frac{1}{d} \sum_{k=1}^n \delta[c_{map}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)] z(\mathbf{s}_k), \quad (3.4)$$

where $c_{map}(\mathbf{s}_k)$ is the mapped soil type at sampling location \mathbf{s}_k with $k = 1, 2, \dots, n$, \mathbf{s}_0 is the mapped soil type at prediction location \mathbf{s}_0 , $\delta[c_{map}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)]$ is the Kronecker delta that equals 1 if $c_{map}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)$ and 0 otherwise, and d is the number of sampling locations for which the mapped soil type equals the mapped soil type at location \mathbf{s}_0 . The population variance at location \mathbf{s}_0 is estimated by:

$$\hat{b}^2(\mathbf{s}_0) = \frac{1}{d-1} \sum_{k=1}^n \delta[c_{map}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)] [z(\mathbf{s}_k) - \hat{a}(\mathbf{s}_0)]^2. \quad (3.5)$$

The standard deviation at \mathbf{s}_0 is taken as the square root of the variance. Likewise $\hat{a}(\mathbf{s}_k)$ and $\hat{b}^2(\mathbf{s}_k)$ are estimated at sampling locations \mathbf{s}_k . Once these are known, the geostatistical analysis is done as described above.

Model 2: Parameter estimation using soil type information from point observations and prediction using the soil map

Model parameters $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$ can also be estimated using the observed ('true') soil type at sampling locations \mathbf{s}_k , $k = 1, 2, \dots, n$. This is attractive because the relationship between soil property and true soil type is not confounded by map errors (i.e. impurities in the map units) and is presumably stronger than the relationship between soil property and mapped soil type.

These parameters are estimated as in *Model 1*, the only difference being that averaging is done over observations with the same observed soil type instead of over observations with the same mapped soil type. Thus $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$ are estimated by:

$$\hat{a}(\mathbf{s}_0) = \frac{1}{d} \sum_{k=1}^n \delta[c_{true}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)] z(\mathbf{s}_k), \quad (3.6)$$

and

$$\hat{b}^2(\mathbf{s}_0) = \frac{1}{d-1} \sum_{k=1}^n \delta[c_{true}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)] [z(\mathbf{s}_k) - \hat{a}(\mathbf{s}_0)]^2, \quad (3.7)$$

where $c_{true}(\mathbf{s}_k)$ is the observed soil type at sampling location \mathbf{s}_k , $\delta[c_{true}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)]$ is the Kronecker delta that equals 1 if $c_{true}(\mathbf{s}_k) = c_{map}(\mathbf{s}_0)$ and 0 otherwise, and d is now the number of sampling locations for which the observed soil type equals the mapped soil type at location \mathbf{s}_0 . Note that here mapped soil type at \mathbf{s}_0 is used to estimate $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$ because true soil type is unknown at unsampled locations. Note also that this approach requires that each soil type read from the soil map is observed at least twice in the sample.

Although calibration of $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$ based on observed soil types is attractive for reasons explained above. However, applying the estimated model parameters to a

soil type map for spatial prediction of these parameters has the undesirable consequence of under-estimating the prediction error variance, because the uncertainty about the soil type prevailing at unsampled locations is not accounted for by the model. Furthermore, a strong, informative, relationship between soil property and soil type is applied to a soil map that is less informative about the soil property, i.e. *Model 2* overrates the usefulness of the information of the soil map. Thus, use of *Model 2* for spatial prediction of soil property Z is essentially flawed. However, if a soil map with quantified uncertainty is available, then one could benefit from parameter estimation based on observed soil type without the drawbacks described here. The following section describes how such model is calibrated and how the kriging prediction and prediction error variance are computed.

Model 3: Parameter estimation using soil type information from point observations and prediction using a soil map with quantified uncertainty

As for *Model 2*, *Model 3* uses observed soil type at sampling locations to estimate model parameters $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$. However, unlike *Model 2*, which uses mapped soil type at \mathbf{s}_0 as if it were the true soil type, *Model 3* represents true soil type at \mathbf{s}_0 with a probability distribution.

Let us assume that a vector of length M containing the probabilities of occurrence of each soil type $p_m(\mathbf{s}_0)$, $m = 1, 2, \dots, M$, is available at each location $\mathbf{s} \in D$. Model parameters $a(\mathbf{s}_0)$ and $b(\mathbf{s}_0)$ are now stochastic variables denoted as $A(\mathbf{s}_0)$ and $B(\mathbf{s}_0)$, which have discrete probability distributions:

$$P(A(\mathbf{s}_0) = a_m) = P(B(\mathbf{s}_0) = b_m) = p_m(\mathbf{s}_0), \quad m = 1, 2, \dots, M. \quad (3.8)$$

The expected value and variance of $A(\mathbf{s}_0)$ are given by:

$$E[A(\mathbf{s}_0)] = \sum_{m=1}^M p_m(\mathbf{s}_0) a_m, \quad (3.9)$$

and

$$\text{var}[A(\mathbf{s}_0)] = \sum_{m=1}^M p_m(\mathbf{s}_0) (a_m - E[A(\mathbf{s}_0)])^2, \quad (3.10)$$

where a_m is the mean of soil type m . Similarly, the expected value and variance of $B(\mathbf{s}_0)$ are given by:

$$E[B(\mathbf{s}_0)] = \sum_{m=1}^M p_m(\mathbf{s}_0) b_m, \quad (3.11)$$

and

$$\text{var}[B(\mathbf{s}_0)] = \sum_{m=1}^M p_m(\mathbf{s}_0)(b_m - E[B(\mathbf{s}_0)])^2, \quad (3.12)$$

where b_m is the standard deviation of soil type m . The most obvious predictor of soil property Z at unsampled location \mathbf{s}_0 is:

$$\hat{z}(\mathbf{s}_0) = E[A(\mathbf{s}_0)] + E[B(\mathbf{s}_0)]\hat{\epsilon}(\mathbf{s}_0). \quad (3.13)$$

It should be noted that this predictor might not necessarily be the optimal predictor, i.e. the best linear unbiased predictor. However, for reasons of simplicity it is used here.

Computation of the expected value and variance of the prediction error is more complex than in the previous models because the model parameters are now stochastic, whereas in *Models 1* and *2* they are fixed. For brevity, the stochastic variables are denoted as $A(\mathbf{s}_0)$ as X , $B(\mathbf{s}_0)$ as Y , $\epsilon(\mathbf{s}_0)$ as U , $\hat{\epsilon}(\mathbf{s}_0)$ as V , and the deterministic variables $E[A(\mathbf{s}_0)]$ as μ_x and $E[B(\mathbf{s}_0)]$ as μ_y . Here X and U , X and V , Y and U , and Y and V are pairwise independent. The prediction error associated with Eq. 3.13 has the expectation:

$$\begin{aligned} E[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)] &= E[X + YU - \mu_x - \mu_y V] \\ &= E[X] + E[Y]E[U] - E[\mu_x] - E[\mu_y]E[V] \\ &= \mu_x + \mu_y \cdot 0 - \mu_x - \mu_y \cdot 0 = 0. \end{aligned} \quad (3.14)$$

The prediction error variance satisfies:

$$\begin{aligned} V(\mathbf{s}_0) &= \text{var}[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)] \\ &= \text{var}[X + YU - \mu_x - \mu_y V] \\ &= \text{var}[X] + \text{var}[YU] + \mu_y^2 \text{var}[V] + \\ &\quad 2\text{cov}[X, YU] - 2\mu_y \text{cov}[X, V] - 2\mu_y \text{cov}[YU, V]. \end{aligned} \quad (3.15)$$

Making use of the unbiasedness of the kriging residual, of $\text{var}[U] = 1$, and of independence between variables, it follows that:

$$\begin{aligned}
V(\mathbf{s}_0) &= \text{var}[X] + (\text{var}[Y] + \mu_y^2) + \mu_y^2 \text{var}[V] + 0 - 2\mu_y(\mu_y \text{cov}[U, V]) - 0 \\
&= \text{var}[X] + (\text{var}[Y] + \mu_y^2) + \mu_y^2 \text{var}[V] \\
&\quad - 2\mu_y^2 \left(\frac{1}{2} \{ \text{var}[U] + \text{var}[V] - \text{var}[U - V] \} \right) \\
&= \text{var}[X] + \text{var}[Y] + \mu_y^2 \text{var}[U - V] \\
&= \text{var}[A(\mathbf{s}_0)] + \text{var}[B(\mathbf{s}_0)] + E[B(\mathbf{s}_0)]^2 \text{var}[\hat{\epsilon}(\mathbf{s}_0) - \epsilon(\mathbf{s}_0)]. \tag{3.16}
\end{aligned}$$

where $\text{var}[\hat{\epsilon}(\mathbf{s}_0) - \epsilon(\mathbf{s}_0)]$ is the simple kriging variance. The third term on the right-hand side of Eq. 3.16 is similar to Eq. 3.3 that is used to compute the prediction error variance of *Models 1* and 2. The first and second term represent the variance caused by uncertainty about the prevailing soil type at prediction location \mathbf{s}_0 . In *Model 1* this uncertainty is accounted for by $b^2 \mathbf{s}_0$ in Eq. 3.3 while model *Model 2* ignores this uncertainty.

Using *Model 3* for spatial prediction of soil property Z requires estimates of a_m and b_m . Soil type mean a_m is estimated by (Goovaerts, 1997):

$$\hat{a}_m = \frac{1}{d_m} \sum_{k=1}^n \delta[c_{true}(\mathbf{s}_k) = m] z(\mathbf{s}_k), \tag{3.17}$$

where d_m is the number of sampling locations with observed soil type m , and $\delta[c_{true}(\mathbf{s}_k) = m]$ is the Kronecker delta that equals 1 if observed soil type at sampling location \mathbf{s}_k equals m and 0 otherwise. Soil type standard deviation b_m is estimated by:

$$\hat{b}_m = \sqrt{\frac{1}{d_m - 1} \sum_{k=1}^n \delta[c_{true}(\mathbf{s}_k) = m] [z(\mathbf{s}_k) - \hat{a}_m]^2}. \tag{3.18}$$

3.3 Case study

The province of Drenthe (2 680 km²) is situated in the northeastern part of the Netherlands (Fig. 1.1). The landscape in Drenthe is characterized by a glacial till plateau that is dissected by an extensive brook valley system. The glacial till is covered with poor aeolian sand deposits that can be up to two metres thick. Cultivated peatlands border the plateau to the east and south. Podzols formed in the aeolian sand deposits. In loamy parent material, brown forest soils formed. Plaggen soils surround medieval settlements on the plateau. Peat soils dominate the centres of the brook valleys. Hydromorphic earth soils are found in the brook valley-plateau transition

zone. Raw sand soils are found in the drift-sand complexes. Peat soils of the cultivated peatlands are characterized by strong human disturbances to a considerable depth due to deep cultivation. Their topsoil consists of sand mixed with small peat fragments.

3.3.1 The Data

Soil map. A raster soil map with 25-m resolution of the province of Drenthe was used. This map distinguishes ten major soil types (Fig. 2.3). The soil map is an updated and generalized version of the existing 1:50 000 national soil map of Drenthe (Chapter 2). The main reason for updating the 1:50 000 map was the large areal decline of peat soils because of intensive agriculture since the original map at that scale was completed (Finke et al., 1996; van Kekem et al., 2005). Location-specific probability distributions of the ten soil types were modelled using multinomial logistic regression, and are available on a grid with 25-m resolution. The mapped soil type is the soil type with the largest probability. Fig. 2.4 shows an entropy map of the probability distributions. The entropy was used to quantify uncertainty associated to the predictions. Entropy values vary between 0 (no uncertainty; one of the soil types has probability 1) and 1 (maximum uncertainty; uniform probability distribution). Note that here the entropy is a normalized score by choosing the base of the logarithm equal to the number of soil types.

Point data. Two sets of point data were available for this study. These sets contained soil profile descriptions at 24 845 locations in Drenthe. Two-thirds of these points were also used to calibrate the multinomial logistic regression models that produced the pedometric soil map (Chapter 2).

- *Data set 1.* 23 785 soil profile descriptions: 95,7% of the total data. These data were collected during five detailed soil surveys (scale 1:10 000) in 1983, 1996-1997, 2002, 2003, and 2006. The five survey areas cover 15% of the area of Drenthe. Sampling density in these areas varies between 0.33-1 locations per ha. The sampling locations in the 1:10 000 survey areas were selected by purposive sampling.
- *Data set 2.* 1 060 soil profile descriptions collected throughout Drenthe during various projects between 1955 and 2008. Average sampling density is four locations per 1 000 ha. The sampling locations were selected by both purposive ($n = 633$) and probability sampling ($n = 427$). Fig. 3.1 shows the locations of the sampling locations of data set 2 and the five areas with detailed soil surveys.

At each sampling location the soil profile was described and classified according to the Dutch soil classification system for detailed soil surveys (ten Cate et al., 1995). Each class was recoded to one of the ten soil types read from the soil map. The

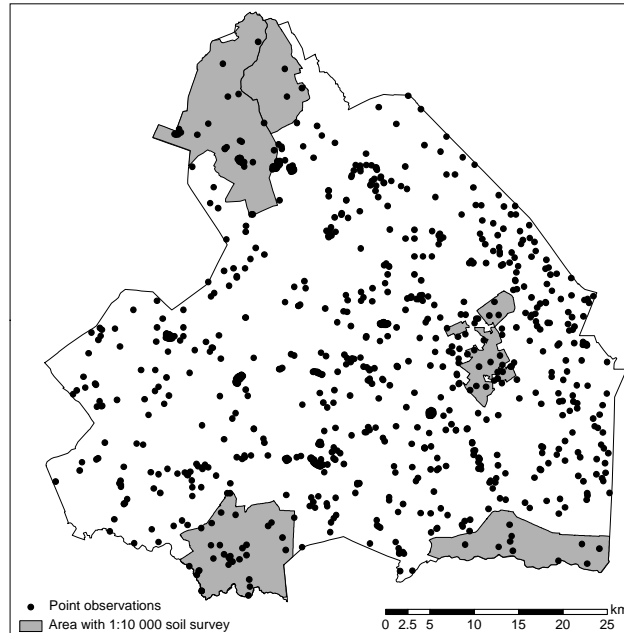


Figure 3.1: Locations of the five areas with detailed soil surveys that yielded point data set 1 and the sampling locations of point data set 2.

soil profile descriptions included the SOM content for each horizon in the soil profile, which was generally determined in the field by hand estimation. The hand estimates were calibrated during the field surveys with a limited number of laboratory measurements. Laboratory measurements of SOM were available for only a few hundred locations. Hand-estimates were not distinguished from laboratory measurements during data analysis and modelling.

SOM is a dynamic property that changes in time. Using a validation set from a different period (2008, see also section 3.3.4) than the calibration set might result in biased predictions. Loss of SOM particularly affects the reclaimed peat soils in the peat colonial landscape. Oxidation of SOM in these soils has resulted in a substantial areal decline (de Vries and Brouwer, 2006). Soil profile descriptions of reclaimed peat soils dating before 1993 were therefore not included in the data sets described above. Profile descriptions of fen peat soils in the brook valleys and of peat soils in contemporary highmoor areas from before 1993 were kept. Oxidation of SOM was assumed to be limited for these soils because of less favourable conditions. Land use in the brook valleys generally is natural grassland with extensive grazing. Highmoor areas are large natural areas. Groundwater levels in both brook valleys as highmoor areas are generally close to the surface. Profile descriptions of mineral soils were all included irrespective of the description date. Hanegraaf et al. (2009) and Reijneveld

et al. (2009) assessed trends in SOM content for agricultural soils in the Netherlands. These authors estimated changes in SOM content of between 0.1% and 0.4% per decade for mineral soils in the north-eastern Netherlands. These changes are fairly small compared with the spatial variation of SOM in mineral soils. It was therefore assumed that the effect of changes of SOM content in mineral soils on calibration bias was negligible.

Soil organic matter content. The SOM content of the topsoil (defined as 0–30 cm), was computed from the SOM contents of the soil horizons according to:

$$SOM = \frac{\sum_{i=1}^n T_i \rho_i SOM_i}{\sum_{i=1}^n T_i \rho_i}, \quad (3.19)$$

where T_i is the thickness of soil horizon i within 30 cm depth (m), ρ_i is the bulk density of the soil horizon (kg m^{-3}) and SOM_i is the soil organic matter content (%). A weighted average based on bulk density was used because mineral and organic horizons occur together in the topsoils in the study area. Such horizons have highly contrasting SOM contents and bulk densities. In order to compute an average SOM content (expressed on a % mass basis) for a topsoil that contains such contrasting horizons, differences in bulk density must be accounted for. Bulk density values of mineral horizons ($SOM < 15\%$) were derived from look-up tables with profile characteristics of the Dutch soils. A pedotransfer function was used to compute bulk density of organic horizons (de Vries, 1999):

$$\rho_i = (1 - ts_i) \frac{100}{\left(\frac{SOM_i}{1.47}\right) + \left(\frac{100 - SOM_i}{2.66}\right)}, \quad (3.20)$$

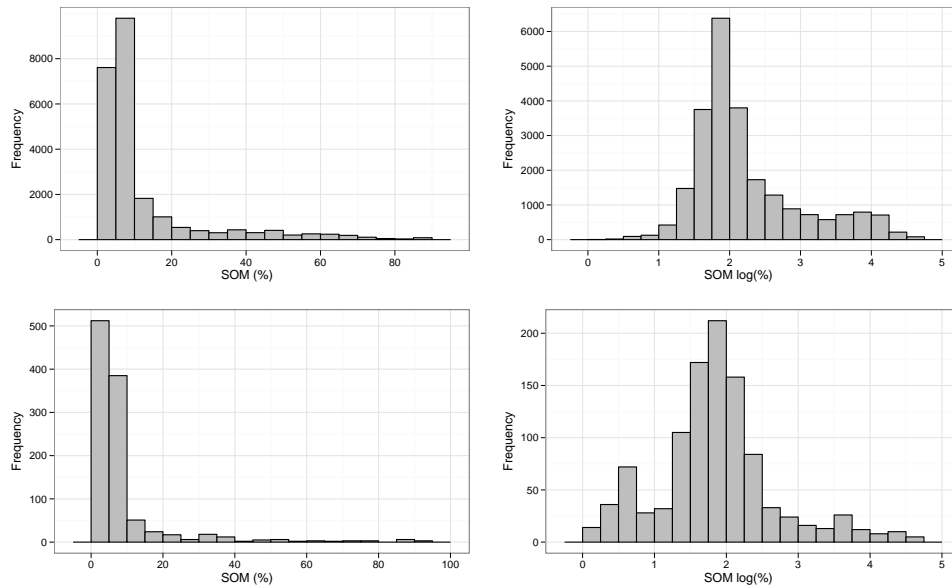
where ts_i is a variable that takes a value between 0.72 and 0.89 depending on the type of organic material and the position of the organic horizon in the profile.

3.3.2 Exploratory data analysis

Table 3.1 lists the summary statistics of topsoil organic matter content of both data sets and Fig. 3.2 shows the histograms. The difference in mean SOM content between the data sets is caused by i) the relatively larger number of peat soils included in data set 1 and ii) larger mean SOM contents for three of the four peat soil types in data set 1. The data of both sets showed strong positive skew. After natural logarithm transformation of the (SOM+1) values, the data were still mildly positively skewed. The peak in the left tail of the distribution of the transformed values of data set 2 is caused by 'raw sands' soil type that is characterized by very small SOM contents in the topsoil. The log-transformed data were used for modelling.

Table 3.1: Summary statistics for raw and log-transformed topsoil organic matter content of the two data sets.

	Mean	Median	Variance	Skew
Data set 1				
SOM (%)	12.4	6.2	233	2.3
SOM log(%)	2.23	1.97	0.58	1.1
Data set 2				
SOM (%)	8.6	5.2	167	4.0
SOM log(%)	1.85	1.82	0.67	0.6

**Figure 3.2:** Histogram of the SOM and log-transformed SOM contents of data set 1 (left plots) and data set 2 (right plots).

3.3.3 Mapping the soil organic matter content

The soil type means, standard deviations and variogram parameters were estimated by using only data from data set 2 ($n = 1060$). Use of both data sets for estimation of the model parameters would result in estimates that are strongly dominated by data from data set 1 which are clustered in only a small part of the study area. However, both data sets were used for kriging the residuals.

Estimation of the soil type means and standard deviations. Two sets of soil type means and

standard deviations of the log-transformed SOM content were estimated from the data: one set was derived from a calibration with mapped soil type and one set from a calibration with observed soil type. Standardized residuals at sampling locations of data set 2 were computed for each set by subtracting the estimated means from the observed log-transformed SOM content followed by division with the standard deviation.

Variogram modelling. Variogram parameters were estimated for the residuals of both calibrations. The variogram parameters were estimated using maximum-likelihood estimation (ML) for its advantages over method-of-moments estimation as discussed by Pardo-Igúzquiza (1998) and Lark (2000). The procedure for parameter estimation by ML as outlined by Lark and Cullis (2004) was followed. To obtain initial values for the variogram parameters the variogram of the residuals was initially estimated by the method-of-moments estimator. After the first run, ML estimation was repeated with different initial values to ensure convergence to the global instead of a local optimum. Parameters were estimated for the spherical and exponential covariance models since these are commonly used to model the spatial structure of soil variables (e.g. McBratney and Webster, 1986; van Meirvenne et al., 2003; Lark and Cullis, 2004; Rivero et al., 2007). The variogram model with the largest log-residual likelihood statistic was then selected for simple kriging of the residuals.

Spatial prediction. Two standardized residuals were computed at each sampling location in data set 1: one for each calibration. These were used together with the standardized residuals from data set 2 for interpolation by simple kriging at the nodes of a regular grid with 25-m resolution. This resulted in two rasters with interpolated residuals: again one for each calibration. Model parameters were estimated at the nodes of a 25-m grid using the three models discussed in the theory section. These estimated parameters were used together with the interpolated residuals to map the SOM content (Eq. 3.2). Note that the interpolated residuals resulting from a calibration with mapped soil type were used for prediction of SOM content with *Model 1*, while those resulting from a calibration with observed soil type were used for prediction with *Models 2* and *3*. Maps of the prediction error variance were produced by using Eq. 3.3 for *Models 1* and *2* and Eq. 3.16 for *Model 3*.

3.3.4 Validation

Sampling design. The performance of the three models was assessed with an independent, stratified simple random sample (de Gruijter et al., 2006). In Chapter 2 a detailed description of the stratification method, sample allocation and sampling protocol is provided (section 2.2.6). A total of 150 locations were sampled in April and May 2008. At each location a soil sample was taken at a depth of 30 cm using a soil auger. Soil samples were oven-dried at 103° C for six hours, and then sieved and crushed. The soil organic matter content of a dry sample was determined with

the weight loss-on-ignition method. The samples were combusted at 550° C for three hours. The SOM content was determined from the weight difference before and after combustion.

Statistical Inference. For each model three validation indices were estimated from the observed and predicted log-transformed SOM content at the 150 validation locations. These were: the mean error (ME), which is a measure for the bias of the predictions; the mean squared error (MSE) which is a measure for the accuracy of the predictions; the mean squared deviation ratio (MSDR) which is a measure of the goodness of the theoretical estimate of the prediction error variance (PEV) (Voltz and Webster, 1990). The MSDR should be close to 1. Values larger than 1 indicate under-estimation of the PEV, while values smaller than 1 indicate over-estimation. The ME of a stratified random sample is estimated by (de Gruijter et al., 2006):

$$ME = \sum_{h=1}^H a_h \frac{1}{n_h} \sum_{i=1}^{n_h} [\hat{z}(\mathbf{s}_{hi}) - z(\mathbf{s}_{hi})], \quad (3.21)$$

where $\hat{z}(\mathbf{s}_{hi})$ is the predicted log-transformed SOM content at validation location \mathbf{s}_i , $i = 1, 2, \dots, n_h$, in stratum h , $h = 1, 2, \dots, H$, $z(\mathbf{s}_{hi})$ is the natural log transform of the measured (SOM+1) content, and a_h is the relative area of stratum h . The MSE is estimated by:

$$MSE = \sum_{h=1}^H a_h \frac{1}{n_h} \sum_{i=1}^{n_h} [\hat{z}(\mathbf{s}_{hi}) - z(\mathbf{s}_{hi})]^2. \quad (3.22)$$

The MSDR is the ratio between actual squared error and the theoretical estimate of the PEV and is estimated by:

$$MSDR = \sum_{h=1}^H a_h \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{[\hat{z}(\mathbf{s}_{hi}) - z(\mathbf{s}_{hi})]^2}{\hat{V}(\mathbf{s}_{hi})}, \quad (3.23)$$

where $\hat{V}(\mathbf{s}_{hi})$ is the PEV as defined by Eq. 3.3 for *Models 1* and *2* and Eq. 3.16 for *Model 3*.

To compare the performance of two models, the difference in the squared error obtained by two models, $\hat{d}(\mathbf{s}_{hi})$, was computed at each validation location. The mean squared error difference, $\hat{\hat{d}}$, was estimated by replacing the quantity between square brackets in Eq. 3.21 with $\hat{d}(\mathbf{s}_{hi})$. Under the null hypothesis that the expected value of the estimated mean difference is zero, it was assumed that $\hat{\hat{d}}$ follows a normal distribution with zero mean and variance (de Gruijter et al., 2006):

$$\text{var}(\hat{d}) = \sum_{h=1}^H a_h^2 \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} [\hat{d}(s_{hi}) - \hat{d}_h]^2, \quad (3.24)$$

where \hat{d}_h is the mean squared error difference of stratum h .

3.4 Results and discussion

3.4.1 Mapping the soil organic matter content

Estimated soil type means and standard deviations

Table 3.2 presents the summary statistics for the log-transformed SOM content derived from calibrations with mapped and observed soil type. The estimated within-soil type variances show that there is considerable variation in SOM content within the map units of the soil map, especially within the four peat map units. The larger variation in the peat map units compared with the mineral map units is caused by the lesser purity of these units (section 2.3.3), and the larger SOM content range in the topsoil of peat soils (15–100% versus 0–15% for mineral soils). Calibration on observed soil type greatly reduces the residual variation within the peat soil types. As expected, calibration on observed soil type results in a stronger relationship between SOM content and soil type than a calibration on mapped soil type: observed soil type explains 77% of the total variance while mapped soil type explains 54%.

Table 3.3 shows the summary statistics for the unstandardized residuals from *Model 1* and *Model 2*. The variance of the residuals of *Model 2* is half of that of *Model 1* indicating a much stronger relationship between SOM content and observed soil type than between SOM content mapped soil type. The skewness statistic of both sets of residuals is close to zero suggesting a symmetrical distribution around the mean.

Variogram Modelling

The exponential model was selected for both prediction models because the log-likelihoods of the exponential models were larger than those of the spherical models. Fig. 3.3 shows the experimental variogram (dots) of residuals from *Model 1* and *Model 2*. The line represents the variogram model fitted by maximum likelihood. The estimated parameters of these are in Table 3.3. The sill of both models is close to 1 as a result of the standardization. The part of the residual variance that is spatially structured, indicated by the spatial dependence ratio, is similar for both models. This is surprising as less structure in the residuals from *Model 1* would have been

Table 3.2: Soil type means and variances of the topsoil organic matter content $\log(\%)$ for a model calibrated using mapped soil type and for a model calibrated using observed soil type. The quantity between brackets is the estimated standard error of the mean

Soil type	Model 1: mapped soil type			Model 2: observed soil type		
	n	mean (se)	variance	n	mean (se)	variance
P	45	3.53 (0.12)	0.68	69	3.81 (0.05)	0.19
mP	20	2.80 (0.24)	1.13	13	2.29 (0.13)	0.23
PY	24	2.49 (0.21)	1.02	40	3.03 (0.07)	0.20
mPY	143	2.38 (0.07)	0.60	76	2.20 (0.05)	0.21
BF	17	1.69 (0.08)	0.10	22	1.66 (0.06)	0.08
PZ	510	1.81 (0.02)	0.21	468	1.77 (0.02)	0.15
ES	92	1.77 (0.06)	0.32	123	1.87 (0.03)	0.14
PS	78	1.90 (0.03)	0.07	86	1.95 (0.03)	0.07
TS	3	1.52 (0.25)	0.19	11	1.59 (0.19)	0.40
RS	128	0.63 (0.03)	0.13	153	0.67 (0.03)	0.15

Table 3.3: Summary statistics of the unstandardized residuals and maximum likelihood estimates of variance parameters for models with different calibrations of the model parameters.

	Model 1	Model 2
<i>Summary statistics</i>		
Mean	0.00	0.00
Median	-0.004	-0.005
Variance	0.30	0.15
Skew	0.18	0.02
<i>Variogram parameters</i>		
Partial sill	0.60	0.55
Nugget	0.44	0.45
Spatial dependence [†]	0.58	0.55
Range (m)	1019	710

[†] The spatial dependence is calculated as [partial sill/(nugget+partial sill)] (Lark and Cullis, 2004)

expected because of the larger explanatory power of observed soil type. The range of spatial correlation is shorter for *Model 2* than for *Model 1*.

Spatial Prediction

Fig. 3.4 shows the maps of the log-transformed SOM content obtained by the three models and the associated prediction error variances. The spatial pattern of SOM content in all maps resembles the soil map with large values in the brook valleys,

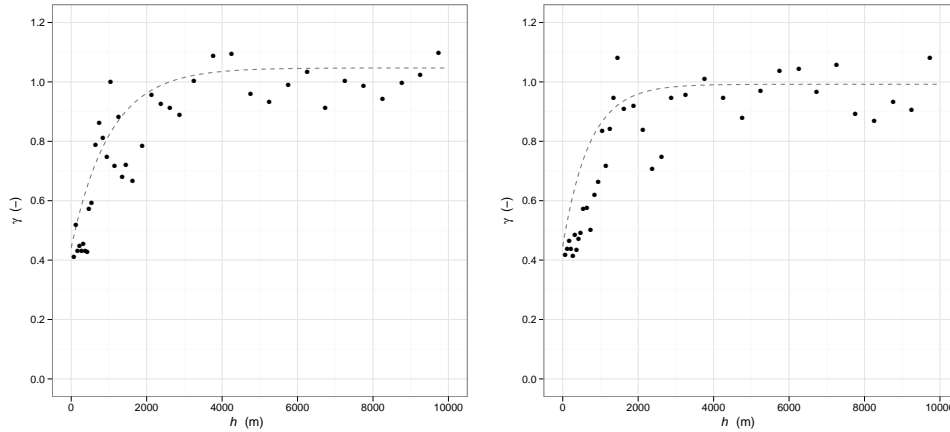


Figure 3.3: Experimental variogram (dots) and maximum likelihood fit of the variogram model (dotted line) for residuals from calibration on mapped soil type (left plot) and calibration on observed soil type (right plot).

which are dominated by peat soils P and PY ; medium values in the peat-colonial landscape, which is dominated by peat soils mP and mPY ; and small values on the plateau with its mineral sandy soils. The maps with *Model 1* and *Model 2* predictions show both sharp and gradual changes across soil type boundaries, which is an effect of the varying sampling density across the study area (Fig. 3.5). In areas with low sampling density sharp changes in SOM content, typical of a discrete model of spatial variation (Heuvelink, 1996), dominate the spatial pattern. Only close to the sampling locations there is a smoothing effect of kriging as a result of the relatively short ranges. In the areas with high sampling density the effect of kriging is much stronger and both sharp and gradual changes occur, which are typical of a map predicted with a mixed model of spatial variation, which combines a discrete model (for the trend) and a continuous model (for the error) to describe the spatial variation of an attribute (Heuvelink, 1996). Thus, in order to take advantage of the kriged residual for spatial prediction a sufficiently dense pattern of observations is required, especially when the range of spatial autocorrelation is short. The map obtained with *Model 3* appears smoother than those obtained with the other models (Fig. 3.5), because here soil type is not represented with discrete entities but with probability distributions. Sharp changes in SOM content only occur where there are sharp changes in probabilities. Furthermore, the *Model 3* map shows much more short-distance variation in SOM content. This is a result of using location-specific probability distributions to represent soil type instead of discrete entities. As a consequence, the predicted SOM content varies between locations (pixels on the map). *Models 1* and *2* are not able to model local uncertainty about soil type, that is uncertainty within map units. Here the SOM content, as predicted by the trend part of the

model, is equal for all locations within a map unit. The variation of SOM content that can be observed within the areas of the *Model 1* and *2* maps that correspond to the map units of the soil map, is attributed to the effect of the kriged residual.

The PEV of *Models 1* and *3* have similar magnitudes (the average of *Model 1* is 0.32; that of *Model 3*, 0.29) but the spatial patterns differ (Fig. 3.4). The PEV pattern of *Model 3* shows much more short-distance variation than that of *Model 1*. This is again caused by the location-specific probability distributions of soil type. As a result the variance of model parameters A and B (Eqs. 3.10 and 3.12) varies between prediction locations. This means that the uncertainty about the predicted SOM content is larger at locations with large uncertainty about soil type than at locations with small uncertainty. In *Model 1* the variance attributed to uncertainty about soil type is not modelled explicitly but is accounted for by the within-map unit variance that is constant within a map unit. This explains the similar magnitude of the PEV for *Models 1* and *3* despite their different spatial patterns. The PEV of *Model 2* is much smaller than those of *Models 1* and *3* (average PEV is 0.14) because this model does not account for the variance resulting from uncertainty about soil type. Fig. 3.6 shows the percentage of the prediction error variance of *Model 3* that is attributed to uncertainty about the trend. Not surprisingly, the map correlates well with the entropy map (Fig. 2.4): large entropy values (large uncertainty about the prevailing soil type) generally result in PEVs that are dominated by the variance of the trend part of the model.

For reasons of simplicity some assumptions were made for the estimation of the model parameters, which may have resulted in too optimistic estimates of the prediction error variance. First, the SOM means per soil type the soil type means were estimated by OLS, hence assuming independence of the point observations. Lark and Cullis (2004) state that OLS estimates of model coefficients from dependent observations may result in biased estimates of the variances of the coefficients. Variances obtained by OLS estimation are under-estimated. Second, it was assumed that model parameters $a(s_0)$ and $b(s_0)$ and the variogram parameters were estimated without error. The estimated parameters were then plugged into the prediction equations as if these were the true values, hereby ignoring uncertainty in these parameters.

Uncertainty about the estimated trend coefficients can be taken into account for variogram modelling and spatial prediction, for example through universal kriging. In that case residual maximum likelihood (REML) estimation is most appropriate for simultaneously estimating the trend coefficients and variogram parameters (Lark and Cullis, 2004; Minasny and McBratney, 2007b). While incorporating uncertainty about the estimated trend coefficients in parameter estimation and prediction is fairly straightforward, addressing the uncertainty about the estimated standard deviation and variogram parameters is not. This requires more advanced methods. Pardo-Igúzquiza et al. (2009) use maximum likelihood inference to assess variogram parameter uncertainty, while Marchant and Lark (2004) use simulation tests. Diggle and Ribeiro Jr. (2007) advocate the use of a Bayesian approach to assess parameter

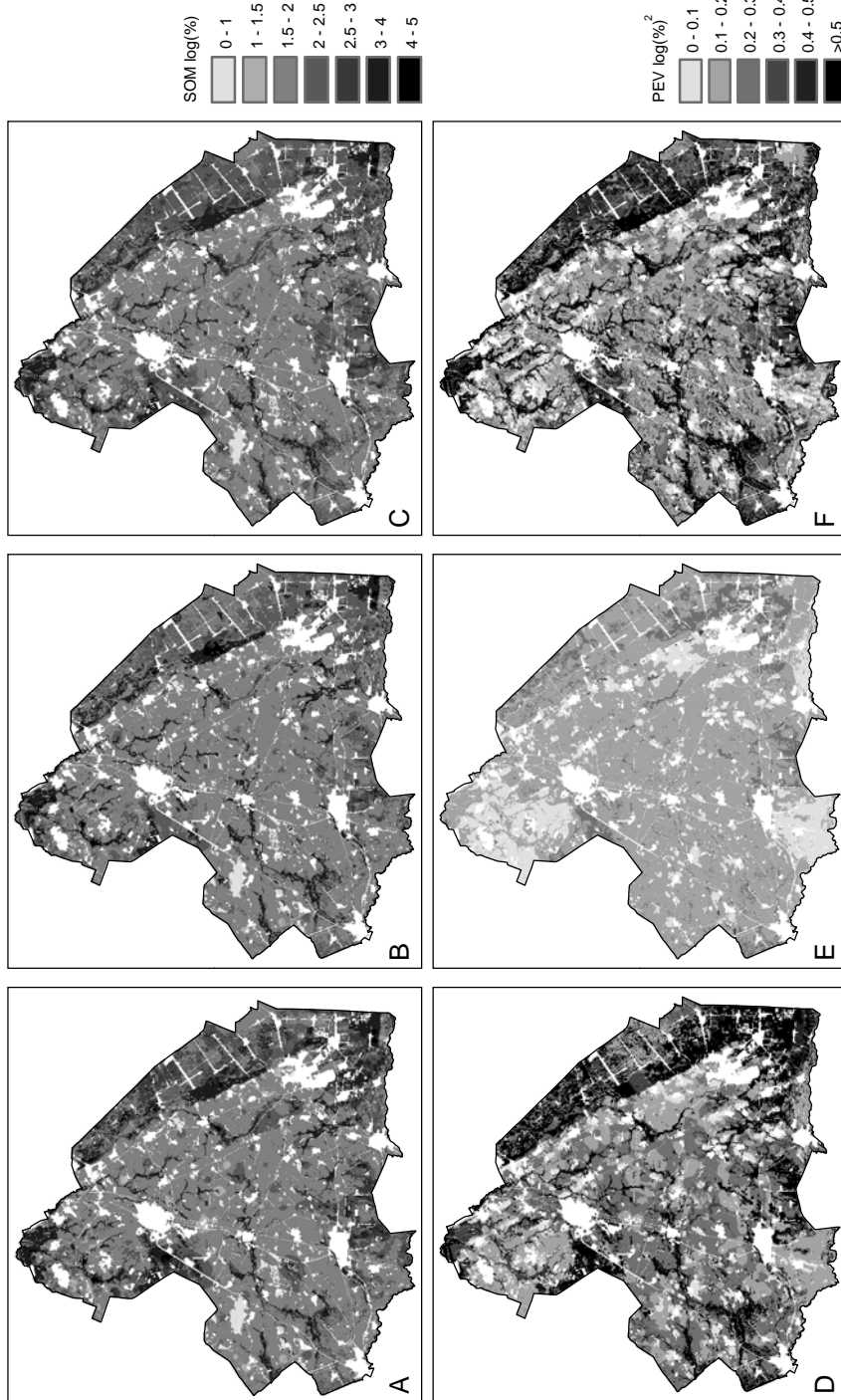


Figure 3.4: Predicted log-transformed SOM contents with Model 1 (A), Model 2 (B) and Model 3 (C) and the prediction error variances associated to the predictions by Model 1 (D), Model 2 (E) and Model 3 (F).

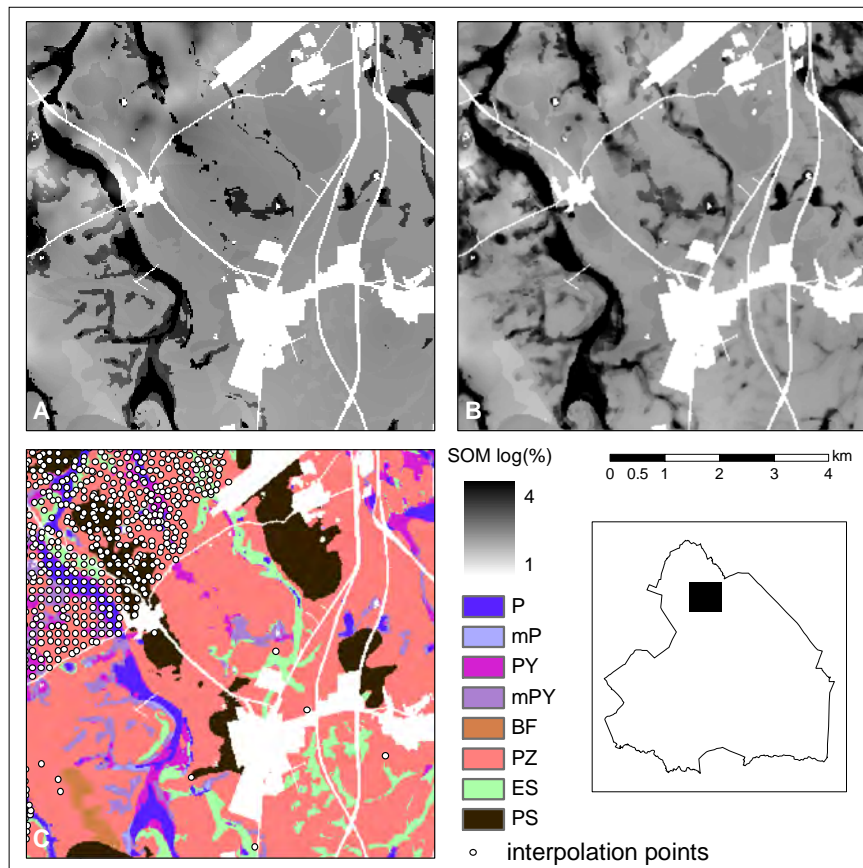


Figure 3.5: Detail of the SOM map obtained by Model 1 (A) and Model 3 (B) that show different spatial patterns. The soil map with the interpolation points are shown in (C).

uncertainty. Use of these advanced methods was not the aim of this chapter.

3.4.2 Validation

Table 3.4 lists the validation results. The mean errors of the models are close to zero, suggesting unbiased predictions. Apparently using calibration and validation data sets from different periods did not cause biased predictions. The mean squared errors show that *Model 3* outperforms *Model 2* ($p = 0.018$) and performs slightly better than *Model 1* ($p = 0.240$). *Model 2* has the largest MSE. This shows that model calibration on observed soil type only improves prediction accuracy when the model explicitly accounts for uncertainty resulting from impurities in the soil map units.

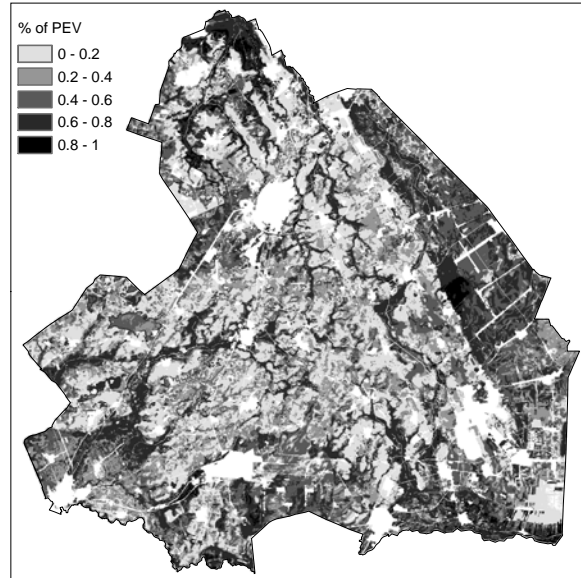


Figure 3.6: Percentage of the prediction error variance of Model 3 that is attributed to uncertainty about the trend part of the model.

Table 3.4: Validation indices of the three prediction models: ME (mean error), MSE (mean squared error), and MSDR (mean squared deviation ratio).

	ME	MSE	MSDR
Model 1	-0.003	0.259	0.91
Model 2	-0.035	0.292	1.98
Model 3	0.005	0.236	0.74

The PEV of *Model 1* is 0.33, that of *Model 2* is 0.14 and that of *Model 3* is 0.30. These values are close to the spatial means of the PEV maps. Also on basis of the PEV, *Model 3* performs better than *Model 1* ($p = 0.009$). The PEV of *Model 3* is composed of the simple kriging variance of *Model 2* and the variances of the soil type means and standard deviations (Eq. 3.16). The latter proved to be negligible, while the variance of the means accounts on average for 39% of the PEV. This is an advantage of *Model 3* over *Model 1*: it allows to quantify the variance that is attributed to uncertainty about soil type. In this particular case, it turns out that eliminating the uncertainty about the true soil type at prediction locations would decrease the PEV by one-third. The small PEV of *Model 2* compared to that of the other models is caused by ignoring the uncertainty about the soil type prevailing at the prediction locations. This is reflected by the MSDR of *Model 2*, which is 1.98, indicating severe underestimation of the PEV. *Model 1* performs well in terms of MSDR, with only a small overestimation

of the actual squared error. Apparently the model does not suffer from ignoring the variances of the model parameters. *Model 3* does not seem to provide an accurate estimate of the PEV: the MSDR deviates strongly from the expected value 1 ($p = 0.003$), indicating an over-estimation of the PEV.

A possible explanation for the bias in MSDR with *Model 3* could be that the multinomial logistic regression models used to produce the pedometric soil map did not take spatial autocorrelation into account when predicting the probability distributions. These distributions were not conditioned on the observations at locations in the neighbourhood of the prediction location, as for example in Bayesian Maximum Entropy prediction (Brus et al., 2008). Consequently, the SOM content as predicted by the trend part of the model is predicted independent of the SOM content at neighbouring locations, while field observations on SOM content (such as used for validation) are correlated. This hypothesis was tested using simulations. Soil type was simulated at sampling and validation locations by drawing from the multinomial probability distributions. Residuals were simulated at both locations through unconditional sequential Gaussian simulation using the estimated variogram model of prediction *Model 2* (Table 3). Next the 'true' SOM content was computed at the validation locations using the simulated soil type and residuals and the soil type data in Table 3.2 (right part). Residuals were predicted at the validation locations by kriging the simulated residuals at the sampling locations, followed by prediction of the SOM content with *Model 3* and computation of the PEV with Eq. 3.16. Finally, the predicted SOM content was validated with the simulated (true) SOM content and the MSDR computed. The average MSDR of 1000 simulations was 0.994 and the 95% confidence interval was 0.985-1.004. These results indicate that it is likely that lack of autocorrelation in the soil type map caused over-estimation of the PEV.

3.4.3 Applicability

Like any other pedometric mapping approach, the models presented in this chapter require data, specifically soil type maps and observations with recorded soil types. Soil type maps are available for most countries of the world although cartographic scales greatly differ between countries (Rossiter, 2004). Here an updated and generalized version of the 1:50 000 soil map of the Netherlands was used. Medium and large scale soil type maps (scales larger 1:250 000) are (partially) available for many countries all over the world¹. An overview provided by Jones et al. (2005) shows that 36 European countries have a national coverage at a scale of 1:250 000 and that one third of these countries have national coverage at 1:25 000–1:50 000 scale. Most countries have thousands of analyzed and described soil profiles in their soil information systems. The SSURGO soil map covers the entire eastern and central US and

¹<http://www.itc.nl/~rossiter/research/rsrch.ss>; *A Compendium of Online Soil Survey Information*

large parts of the western US at 1:12 000–1:63 600 scale². The US soil database contains about 20 000 descriptions of pedons (Rossiter, 2004). Canada has coverage at 1:10 000–1:125 000 scale of significant agricultural areas³. Selvaradjou et al. (2005a,b) give an overview of available (scanned) soil maps of Africa and Asia. This shows that application of *Model 3* is possible in large parts of the world although pedometric soil type maps are rare. When such maps are not available, point observations with recorded soil types can be used to define frequency distributions of soil types within the mapping units of the soil type map, which can then be used to proxy probability distributions for soil type.

3.5 Conclusions

Calibration of the model trend on observed soil type resulted in a much stronger predictive relationship between SOM and soil type than calibration on mapped soil type, which is done conventionally. Such calibration implies, however, that soil type as depicted on a soil map cannot be simply used as a covariate. The true soil type at prediction locations must be used, which is unknown. However, when a probability model of the true soil type is available, then this model may be used. This approach was followed in this chapter and obtained more accurate predictions of SOM with smaller prediction error variance (PEV) than the conventional method. *Model 2*, which is theoretically flawed, performed worst. It had the largest MSE and the PEV was strongly underestimated. Use of this model for pedometric mapping of soil properties should be avoided.

Model 3 gives a more realistic representation of the PEV than *Model 1*. It uses location-specific probability distributions to account for uncertainty about the prevailing soil type at prediction locations. Hence, uncertainty associated to the predicted SOM content is larger at locations where we are more uncertain about the prevailing soil type than at locations where we are less uncertain. *Model 1* is not capable of modelling this because in this model the uncertainty about the prevailing soil type is implicitly accounted by the map unit standard deviations and is therefore constant within a map unit. In addition, *Model 3* makes it possible to quantify the variance attributed to uncertainty about soil type at prediction locations.

Quantitative information about the uncertainty of existing maps of soil type is rare, although soil survey reports or experienced soil surveyors might provide estimates. Still, pedometric soil mapping is developing rapidly and is increasingly applied for producing and updating of soil maps from field scale to global scale. Maps of soil type with quantified uncertainty should become more readily available in the near future. Use of such maps for mapping soil properties is advocated.

²<http://soils.usda.gov/survey/geography/surgo>; *Soil Survey Geographic (SSURGO) Database*

³<http://sis2.agr.gc.ca/cansis>; *Canadian Soil Information Service*



Chapter 4

Three-dimensional mapping of soil organic matter content using soil type-specific depth functions

In this chapter a method is proposed for mapping depth functions of soil organic matter (SOM) that combines general pedological knowledge with geostatistical modelling. A digital soil map that represents soil type at any location with a probability distribution formed the starting point. For each of the ten soil types depicted on this map a depth function structure was defined that describes the SOM depth profile based on knowledge of soil profile morphology. Five depth function building blocks were defined, referred to as 'model horizons'. For each soil type a depth function structure was obtained by stacking a subset of model horizons. The parameters of the ten soil type-specific depth functions were calibrated with data from soil profile descriptions and spatially interpolated using environmental covariates as predictors. From the predicted parameters and the soil type-specific depth function structures the depth functions were reconstructed for each soil type at each prediction location. By combining the soil type-specific depth functions with the soil type map a probability distribution of depth functions was obtained at each location in the study area. The soil type-specific depth functions and their associated probabilities were used to map the SOM stock for depth intervals 0–30 cm, 30–60 cm, 60–90 cm, and 0–90 cm. The mapped SOM stocks were validated with independent probability sample data. Validation results showed adequate predictions for the topsoil but poor predictions for the subsoil. Additionally, prediction performance of the pedometric depth functions was compared to that of conventional depth functions.

Based on: B. Kempen, D.J. Brus, J.J. Stoorvogel
Geoderma 162 (2011): 107–123

4.1 Introduction

Several attempts have been made recently to use pedometric methods to map the three-dimensional variation of soil properties using depth functions (Minasny and McBratney, 2006b; Malone et al., 2009; Meersmans et al., 2009; Mishra et al., 2009). This typically involves the use of splines or exponential decay functions to describe the variation of soil properties down a profile. The parameters of these functions are estimated at observation locations, correlated with environmental covariates and then spatially interpolated across the area of interest. Use of splines or exponential decay functions is based on the premise that soil properties vary continuously with depth (Bishop et al., 1999; Ponce-Hernandez et al., 1986). Although this might be true for large areas in the world, in areas where there has been strong human influence on soil formation or where the soil profile contains highly contrasting parent materials, sharp discontinuities in the depth distribution of soil properties occur. Furthermore, besides an anthropogenic or geologic origin, discontinuities can also have a pedologic origin such as the sharp boundary between the eluvial and illuvial horizons in a podzol. Fig. 4.1 gives some examples of soil profiles with discontinuities of different origins.

The Netherlands has intensively managed landscapes. Soils are often disturbed or are completely man-made. Besides, many soils have different parent materials in their profile. For example, soils of the cultivated peatlands—reclaimed highmoor swamps—have a sandy topsoil that covers peat remains in the subsoil. In the north of the Netherlands, thick layers of loamy glacial till are covered with sand deposits, whereas in the river floodplains alternating sand, loam, clay and peat layers are found in the soil profile as a result of a changing depositional environment in the past. This makes purely continuous variation of soil properties with depth the exception rather than the norm in large parts of the Netherlands. Discontinuities encountered in depth profiles of soil properties must be explicitly considered in depth functions used to model these profiles.

Depth functions are typically calibrated using information from soil profile samples or descriptions at point locations (Malone et al., 2009; Mishra et al., 2009). In addition, soil maps are available for many areas. Although soil maps typically describe discrete spatial variation of soil types, they also provide insight in the profile build-up of the different soil types and representative average properties for the different soil layers (Bregt and Beemster, 1989; van Meirvenne et al., 1994; Liu et al., 2006). Hence, soil type maps can be considered discrete three-dimensional models of soil properties. If available, such maps may be used for three-dimensional mapping of soil properties. Soil horizons then act as carriers of soil property information. If a dataset with field descriptions and classifications of soil profiles at point locations is available then these data can be used to describe a depth function for each soil type depicted on the soil type map.

The type of depth function can vary between different soil types depending on the natural and anthropogenic soil forming processes. Different soil-forming processes result in different soils with different soil property depth profiles and therefore require different depth functions to model these. Meersmans et al. (2009) were first to take anthropogenic influence (tillage) on soil formation and profile morphology into account for modelling the vertical distribution of soil organic carbon (SOC). They used a function with a constant SOC concentration until tillage depth. The vertical distribution of the SOC concentration below the tillage layer was modelled with an exponential decay function.

In this chapter the current methodology for three-dimensional mapping of soil properties is extended by integrating pedological knowledge on soil property depth profiles and geostatistical modelling by i) characterizing a depth function of soil organic matter (SOM) for each of the ten soil types occurring in a study area in the province of Drenthe, ii) mapping the soil-type specific depth functions across the study area and iii) using the mapped functions in combination with a digital soil type map, which represents soil type at any location in the study area with a probability distribution, to predict the SOM stock at four depth intervals. The results were validated with data from an independent probability sample.

4.2 Material and methods

4.2.1 Study area

The 125 km² study area is situated in the south-central part of the province of Drenthe (Fig. 4.2). The area surrounds the village of Oosterhesselen (52.75 N, 6.72 E) and corresponds to map sheet 17G of the 1:25 000 Dutch Topographical Map. It is a typical part of the Drenthe plateau with large agricultural complexes formed by the open-field farming system, early 20th-century heath reclamations, fen peat-filled brook valleys and forested drift-sand complexes. The area includes a small part of the man-made cultivated peatlands in the southwest. Major land use is agriculture with cropland (potatoes, wheat, maize) dominating the open-field complexes and the cultivated peatlands; a mosaic of cropland and intensive pasture systems is found in the heath reclamations, while extensive pasture systems and natural grasslands characterize land use in the brook valleys. One third is under nature (forests, heath and highmoor). Major soil types include podzols, hydromorphic earth soils, plaggen soils and peat soils.

Podzols formed in poor aeolian sands in the drier parts of the landscape and dominate the heath reclamations and parts of the cultivated peatlands, while brown forest soils formed in more loamy sands. Hydromorphic earth soils, which are soils with

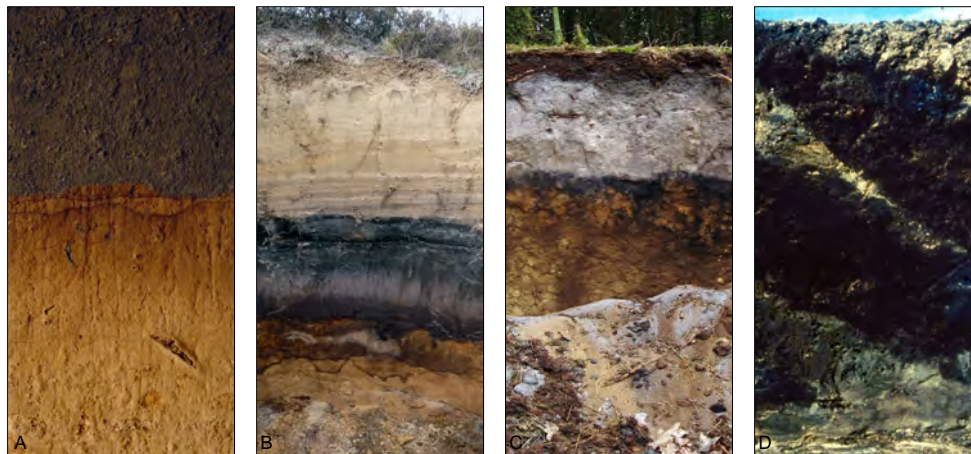


Figure 4.1: Soil profiles showing discontinuities in the vertical distribution of soil organic matter caused by different processes: A) anthropogenic: plough layer on top of podzol remains, B) geologic: Holocene drift sand covering Pleistocene sand with podzol, C) pedologic: eluvial E-horizon above an illuvial podzol-B horizon, D) anthropogenic: the soil of the cultivated peatland showing dislocated peat (dark material) and sand (light material) layers resulting from deep cultivation.

a humic topsoil overlying a C-horizon with gleyic features, are found in hydrologically intermediate positions between well-drained podzols and poorly-drained peat soils. Plaggen soils which have a deep (>30 cm) anthropogenic humic topsoil are found in the open-field complexes. Raw sand soils (arenosols) are found in the drift-sand complexes. Deep peat soils are found in the centres of the brook valleys. Peat soils in the cultivated peatlands are marked by their strong human disturbances to a considerable depth because of deep cultivation and their man-made, sandy topsoil.

4.2.2 Soil profile data

Soil profile data from 2 111 locations situated in the province of Drenthe (Fig. 4.2) were obtained from *BIS*. Ninety-one of these are situated in the study area. The remaining profile descriptions are situated elsewhere in the province at locations with similar soil conditions. These profile data were collected during various projects between 1955 and 2009 that used both purposive ($n = 1\,492$) and probability sampling designs ($n = 619$). At each sampling location the soil profile was described, typically to a depth of 120 to 150 cm, and classified according to the Dutch soil classification system for detailed soil surveys (ten Cate et al., 1995).

The profile descriptions included the SOM content (in mass%) for most of the horizons, horizon thicknesses (cm) and horizon parent material. Most SOM contents

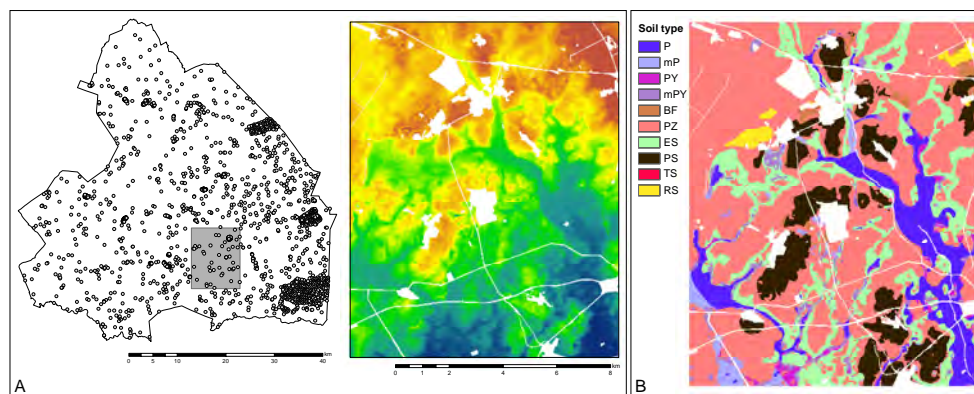


Figure 4.2: (A) Location of the study area in Drenthe. The points indicate locations at which soil profile descriptions are available. The right plot shows the elevation in the study area (10–22 m above sea level) with the lowest positions in blue and the highest positions in red-brown. (B) Soil map of the study area. The depicted soil type at each location is the soil type with the largest probability. P = deep peat soils (organic layer >40 cm), mP = deep peat soils with mineral surface horizon, PY = shallow peat soils, mPY = shallow peat soils with mineral surface horizon, BF = brown forest soils, PZ = podzols, ES = hydromorphic earth soils, PS = plaggen soils, TS = glacial till soils (till is present within 40 cm from the surface), RS = raw sand soils.

were estimated by hand in the field; approximately 500 soil profiles have partial or complete laboratory measurements (no distinction was made between these for data analysis and modelling). In most survey projects hand-estimates of soil properties are calibrated with laboratory measurements, and consequently it is expected that use of hand-estimates does not lead to substantial prediction bias when such data is used for (geostatistical) modelling. Twenty percent of the 11 000 soil horizons in the dataset missed a value for the SOM content. The SOM content for these horizons, mainly mineral C-horizons, was obtained from similar horizons in representative soil profile descriptions (de Vries, 1999).

For this study depth functions of volumetric SOM content (kg m^{-3}) were constructed and predicted so that these functions can be used to predict SOM stock for a specific depth interval. SOM content in mass% was therefore converted to volume basis:

$$C_V = \frac{C_M}{100} \rho_s, \quad (4.1)$$

where C_V is the SOM content on volume basis (kg m^{-3}), C_M is the SOM content in mass% ($\text{kg kg}^{-1} \cdot 100$) and ρ_s is the bulk density (kg m^{-3}). Bulk densities were derived from a look-up table where soil bulk density is a step-wise function of parent material (geological deposit), SOM class and position in the profile (topsoil or

subsoil). The look-up table was compiled by experienced soil surveyors. There is no information about the accuracy of the bulk density values derived from the look-up table, but there are no have reasons to assume these would cause bias in the model results.

4.2.3 Soil map

The soil map is a raster soil map of 25-m resolution that distinguishes ten major soil types (Fig. 4.2). The soil map is an updated and generalized version of the Drenthe part of the 1:50 000 national soil map. Soil types are represented through location-specific probability distributions, as predicted by multinomial logistic regression. The mapping process is described in detail in Chapter 2.

4.2.4 Environmental covariates

Ten datasets were available from which 26 environmental covariates were derived. These can be used to improve spatial prediction of the depth function parameters. Table 4.1 provides an overview of the covariates, including the classes of the categorical ones. The primary datasets are:

- *Digital elevation model (DEM)*. A 25-m resolution DEM, constructed from LIDAR measurements. Three relative elevation grids were derived from the DEM by subtracting the local mean elevation within search radii of 250, 500, and 750 m from the elevation at each grid cell. These grids capture local relief.
- *Groundwater class map*. Finke et al. (2004) constructed a 25-m resolution that distinguishes eleven groundwater table classes. These were grouped into three drainage classes: poorly, moderately, and well drained soils.
- *Land cover maps*. This set contains five grids depicting land cover in 1900 (50 m), 1940 (25 m), 1960 (25 m), 1980 (25 m) and 2003 (25 m) (Clement and Kooistra, 2003; Knol et al., 2003, 2004; Hazeu, 2005). The 1900 grid distinguishes ten classes and was reclassified into a grid with two classes and into a grid with three classes. The 2003 grid distinguishes 23 classes and was reclassified into six grids, each with a different combination and number of classes. The 1900, 1940, 1960, 1980 and 2003 grids were combined into a map that represents reclamation period. Six grids were derived from the reclamation period map, each with a different combination and a different number of periods.
- *National Soil Map* (Steur and Heijink, 1991). The extent of the cultivated peatlands was digitized using digital and paper versions of the Drenthe part of the

1:50 000 soil map. Furthermore, a layer representing mechanical soil disturbance was extracted from the attribute data of the map polygons. Both data layers were converted to grids of 25-m resolution.

- *Paleogeography map* (Spek, 2004). This 1:50 000 map represents a reconstruction of the geography of Drenthe around 1000 AD, and distinguishes 12 geographical units. The map was converted to a grid of 25-m resolution. Three new grids were extracted from this grid: one with the former extent of fen peat coverage, one with the former extent of highmoor coverage and one with the former total extent of peat coverage. Furthermore two grids were created from combinations of peat types and the extent of cultivated peatlands.
- *Geomorphology map* (Koomen and Maas, 2004). A 1:50 000 map representing landform units. The map was converted to a grid of 25-m resolution. A layer with the extent of the land dunes was extracted from the geomorphology grid.

Table 4.1: List of environmental covariates.

Description	Levels	Code
<i>Digital elevation model</i>		
Relative elev, 250 m	-	REL250
Relative elev, 500 m	-	REL500
Relative elev, 750 m	-	REL750
<i>Groundwater class maps</i>		
Drainage condition	1 = poor, 2 = moderate, 3 = well	GD
<i>Land cover maps</i>		
Land cover 1900, 2 cl	1 = agriculture, 2 = natural	LC19002CL
Land cover 1900, 3 cl	1 = grassland, 2 = cropland, 3 = natural	LC19003CL
Current land cover, 2 cl	1 = agriculture, 2 = natural	COV2CL
Current land cover, 2 cl	1 = highmoor, 2 = other	COV2aCL
Current land cover, 3 cl	1 = agriculture, 2 = forest, 3 = rangeland/highmoor	COV3CL
Current land cover, 3 cl	1 = grassland, 2 = cropland, 3 = natural	COV3aCL
Current land cover, 4 cl	1 = agriculture, 2 = forest, 3 = rangeland, 4 = highmoor	COV4CL
Current land cover, 4 cl	1 = grassland, 2 = cropland, 3 = forest, 4 = rangeland/highmoor	COV4aCL
Current land cover, 5 cl	1 = grassland, 2 = cropland, 3 = forest, 4 = rangeland, 5 = highmoor	COV5CL
Reclamation period, 2 cl	1 = <1940, 2 = >1940	RECLAM2CL
Reclamation period, 3 cl	1 = <1940, 2 = 1940-2003, 3 = >2003	RECLAM3CL
Reclamation period, 3 cl	1 = <1940, 2 = 1940-1980, 3 = >1980	RECLAM3aCL
Reclamation period, 3 cl	1 = <1900, 2 = 1900-1940, 3 = >1940	RECLAM3bCL
Reclamation period, 4 cl	1 = <1900, 2 = 1900-1940, 3 = >1940-2003, 4 = >2003	RECLAM4CL
Reclamation period, 6 cl	1 = <1900, 2 = 1900-1940, 3 = >1940-1960, 4 = 1960-1980, 5 = 1980-2003, 6 = >2003	RECLAM6CL
<i>Soil map, (scale 1:50 000)</i>		
Peat-colonial landscape	1 = Yes, 0 = No	PEATCOL
Disturbed soils	1 = Yes, 0 = No	DISTURBED
<i>Paleogeography map</i>		
Former fenpeat coverage	1 = Yes, 0 = No	FENPEAT
Former highmoor coverage	1 = Yes, 0 = No	HIGHMOOR
Former peat coverage	1 = Yes, 0 = No	PEAT
<i>Paleogeography & Soil maps</i>		
Peatlands geography	1 = Peat-colonies, 2 = Highmoor, non peat-colonial, 3 = other	PEATGEO
Former highmoor, non peat-colonial	1 = Yes, 0 = No	MOOROUTPEATCOL
<i>Geomorphology map</i>		
Land dunes	1 = Yes, 0 = No	DUNES

4.2.5 Defining the depth functions

Depth function structure and parameters

For each of the ten soil types in the study area a depth function structure was defined that describes the depth distribution of SOM based on a conceptual model of pedogenesis, explained below. Each soil type-specific depth function is constructed from building blocks, hereafter referred to as 'model horizons'. Based on the pedological model, it was assumed that five different model horizons were needed to define the structures of the SOM depth profiles of the soil types in the study area:

1. organic topsoil horizon with constant SOM over depth.
2. mineral topsoil horizon with constant SOM over depth.
3. organic subsurface horizon with constant SOM over depth.
4. mineral subsurface horizon with constant SOM over depth.
5. horizon with SOM exponentially decreasing over depth.

Note that these model horizons share similarities with functional layers or units, i.e. pedogenic horizons with comparable soil properties (Finke et al., 1992; Cosandey et al., 2003).

For each soil type a depth function structure was obtained by stacking a subset of model horizons. The stacking order of the model horizons is fixed and follows the above order. For example model horizon 1 is always on top of 3 and model horizon five is always the last horizon in the stack. In addition, only a single topsoil horizon is allowed. Each model horizon has two associated parameters that characterize the depth distribution of SOM. This means that the SOM depth profiles of the ten soil types in the study area can be modelled with ten unique parameters. For model horizons 1 to 4 these parameters are SOM content (kg m^{-3}) which shall be denoted as C_i and thickness (m) which shall be denoted as d_i . Here i indicates the model horizon. The SOM content in model horizon 5 is modelled by a negative exponential depth function with parameters C_a and k :

$$C_V(z^*) = C_a \exp(-kz^*) , \quad (4.2)$$

where $C_V(z^*)$ is the SOM content volume basis (kg m^{-3}), z^* is the depth from the top of the model horizon (m), C_a is the SOM content at the top of the model horizon, k (m^{-1}) is the rate of SOM decrease with depth. The depth function for a soil type can then be constructed using the depth function structure and the associated parameters.

Pedological models for soil type-specific depth functions

The soil type-specific depth function structures—the stacking order of the model horizons—are based on expert knowledge on the depth profiles of SOM of the soils in the study area:

- Deep peat soil *P* generally has a well-decomposed organic topsoil (model horizon 1) overlaying a less-decomposed organic horizon that has a higher SOM content (model horizon 3) than the topsoil. When mineral material is encountered within observation depth then the transition from organic to mineral material is sharp. SOM content in the mineral subsoil decreases with depth. The soil horizons in the mineral subsoil make up model horizon 5. The organic topsoil with decomposed material can be absent in areas under natural vegetation ($d_1 = 0$). For soil type *P* the SOM depth profile can be modelled by stacking model horizons (1), 3 and 5.
- The SOM depth profile of soil type *mP* is similar to that of soil type *P*, only the topsoil is mineral instead of organic and is always present. The depth profile can be modelled by stacking model horizons 2, 3 and 5.
- The SOM depth profile of shallow peat soil *PY* can be modelled with the same model horizon sequence as *P*, although the organic subsoil is often absent ($d_3 = 0$).
- The SOM depth profile of soil type *mPY* is similar to that of soil type *mP* but with a thinner organic subsoil. The mineral subsoil is always present within observation depth. The depth profile can be modelled by stacking model horizons 2, 3 and 5. Soil types *mP* and *mPY* are characteristic of the cultivated peatlands. Deep cultivation of these soils after reclamation often dislocated sand and peat layers in the subsoil (Fig. 4.1d). Inclusions of irregular sand bodies in the predominantly organic subsoil are therefore common. To keep the modelling of the depth functions straightforward these inclusions were considered part of model horizon 3.
- Soil type *PS* is characterized by a humic topsoil 30–50 cm thick. Generally one or two A-horizons are distinguished within the topsoil: the Aap-horizon, which is the ploughed anthropogenic surface horizon (30–35 cm deep), and the Aa-horizon, which is a buried anthropogenic A-horizon. In the conceptual depth function model the former is regarded as model horizon 2, while the latter is regarded as model horizon 4. Below the plaggen layer remains of a podzol (without the E-horizon) or brown forest soil are found, the soil horizons of which make up model horizon 5.
- The SOM depth profile of soil types *BF*, *PZ*, *ES* and *TS* can be modelled by stacking model horizons 2 and 5. It is assumed that each has a mineral topsoil

with a constant SOM depth profile regardless the land use. Soil under agriculture have a constant SOM because of long-term cultivation. Soils under forest were ploughed before afforestation in the early 20th century. These soils are characterized by a disturbed, heterogeneous topsoil, of which the SOM profile can best be modelled with a constant. Soils under heath virtually lack an A-horizon. The topsoil typically is classified as AE or Eh. Also for these topsoils a constant SOM profile was assumed. Subsoil morphology of podzols (*PZ*) and brown forest soils (*BF*) is characterized by soil horizon sequence Bh or Bw–BC–C, which has a decreasing SOM profile. The podzol E-horizon is absent in cultivated soils. Most soils under forest also lack an E-horizon because the original E material is often incorporated in the topsoil; mixed with A and B material. The hydromorphic earth soils (*ES*) and till soils (*TS*) lack a B-horizon. Their subsoil horizon sequence generally is AC–C with a gradually decreasing SOM profile. The subsoil SOM depth profile in soil types *BF*, *PZ*, *ES* and *TS* was therefore modelled with an exponential decay function. The subsoil horizons in these soils will be allocated to the model horizon 5 building block.

- Soil type *RS* is mainly found in the drift-sand complexes. In this highly dynamic environment the material of the original podzols was blown out and redeposited in the surroundings. The land dunes were afforested with pine in the early 20th century. Once the material was kept in place by vegetation, soil forming processes could start again. Initial signs of soil formation, if present, can now be observed within 5–15 cm from the soil surface. These signs include a shallow Ah, AE or AC horizon or a very shallow podzol profile (<15 cm thick). In this study these layers were regarded as topsoil layers and were modelled by model horizon 2. Subsoil morphology is extremely heterogeneous as a result of the aeolian processes that created the drift-sand complexes. The soils can contain layers of humus-poor drift sand, podzolic horizons, buried A-horizons, glacial till and peat remains. The occurrence of these layers and their thickness is spatially (both lateral and vertical) highly variable. It is far beyond the potential of the proposed method to model this variation. A simple and pragmatic solution is to model the SOM depth profile in the subsoil of soil type *RS* with a constant (model horizon 4).

4.2.6 Mapping the depth functions

This section describes the methodological framework for mapping the soil type-specific depth functions of SOM. A flowchart summarizing the mapping process is presented in Fig. 4.3.

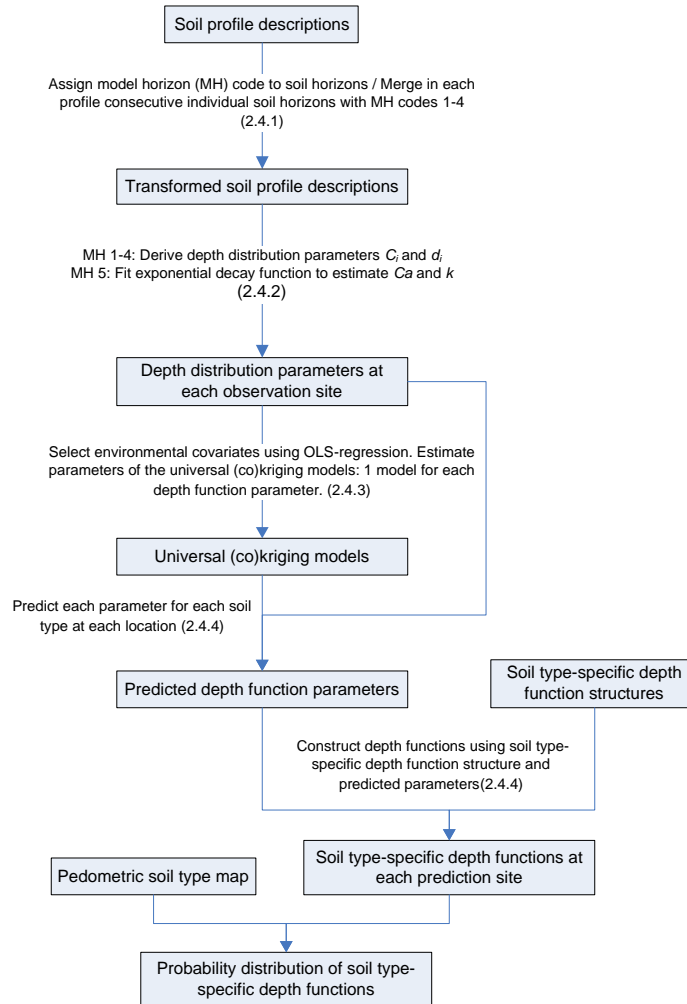


Figure 4.3: Methodological flowchart.

Allocating soil horizons to model horizons

The soil profile dataset contained 2 111 profile descriptions with almost 11 000 soil horizons. In total 230 unique horizon codes were used to classify the soil horizons based on morphogenetic properties. In order to facilitate allocating soil horizons to model horizons, the soil horizons were recoded into twelve master soil horizons (Table 4.2). A model horizon number was then attributed to each soil horizon in the dataset based on master soil horizon code and geological deposit. This process was guided by expert decisions.

Table 4.2: Master soil horizon classes used to transform soil horizons to model horizons.

Soil horizon	Description
A1	Organic topsoil
A2	Mineral, humic topsoil
A3	Anthropogenic, humic soil horizon
Ab	Buried A-horizon
AT	Transition horizon: AE or AC
E	Eluviation horizon
Bs	Illuvial horizon with SOM accumulation
BC	Transition horizon
C1	Parent material layer: peat
C2	Parent material layer: sand, till
MM	Mixed mineral material (A,B,E,C2)
MO	Mixed mineral, organic material (A,B,E,C1)

Allocating soil horizons to model horizons may result in model horizons that consisted of multiple soil horizons. For model horizons 1 to 4 that consisted of several consecutive soil horizons, the average SOM content was computed as a thickness-weighted average of the SOM contents of the individual soil horizons. To obtain the thickness of a model horizon the thicknesses of the individual soil horizons were simply summed. The individual soil horizons that make up model horizon 5 were retained as the SOM contents of the individual soil horizons are required to fit the exponential decay function.

Deriving the depth function parameters of model horizons

The parameters SOM content and thickness, C_i and $d_{i,i}$, of model horizons 1 to 4 were directly derived from the transformed soil profile descriptions. The parameters C_a and k of model horizon 5 were estimated by fitting the exponential function Eq. 4.2 to the individual soil horizons that make up model horizon 5, using non-linear least squares.

The parameters C_a and k were fitted by minimizing the sum of squared differences between the observed and predicted SOM stocks of individual soil horizons within the model horizon:

$$O = \sum_{i=1}^n \left[\left(\frac{C_a}{k} [\exp(-kz_{L_i}^*) - \exp(-kz_{U_i}^*)] \right) - C_{I_i} \right]^2, \quad (4.3)$$

where $z_{L_i}^*$ is the depth of the lower boundary of soil horizon i in a soil profile with respect to the top of the model horizon, n is the number of soil horizons within the model horizon, $z_{U_i}^*$ is the depth of the upper boundary soil horizon i with respect to the top of the model horizon, and C_{I_i} is the observed SOM stock of horizon i . Note that this approach has similarities with the equal-area smoothing spline

(Bishop et al., 1999; Ponce-Hernandez et al., 1986) where for each sample layer (horizon) the area to the left of the fitted spline is equal to the area to the right of the fitted spline so that the spline represents the average value of the soil property for each sample layer. But unlike the spline, minimizing Eq. 4.3 does not always guarantee mass-conservation. Nevertheless, it has better mass-conserving properties than a decay function fitted to the mid-points of the soil horizons. Also note that at least three observations of SOM within a model horizon 5 building block are required to fit the decay function. Soil profiles in the dataset with a model horizon 5 building block that comprises less than three soil horizons were not used to estimate and interpolate the decay parameters of this horizon.

Selecting environmental covariates and calibrating the geostatistical models

The aim is to map soil-type specific depth functions of SOM. So far a model horizon sequence was defined for each soil type in the study area, the parameters were defined that describe the SOM depth profiles and the values of these parameters were derived from soil profile descriptions at point locations. The next step is to predict the depth function parameters at the nodes of a 25-m grid that covers the study area. For each depth function parameter one spatial model was defined, using observed soil type at the point locations as one of the covariates in the model because parameter values are expected to differ between soil types.

For each depth function parameter a set of environmental covariates was selected using ordinary-least-squares (OLS) regression and the Akaike Information Criterion (Webster and McBratney, 1989) as a selection criterion. Interactions between covariates were not included to keep the models simple. The depth function parameters were predicted by universal kriging with variance models estimated by residual maximum likelihood. A detailed description of this method can be found in Lark et al. (2006) and Lark and Webster (2006). Correlated depth function parameters were interpolated using universal co-kriging with auto- and cross-covariance models estimated by the methods-of-moments estimator. Profile descriptions of peat soils dating before 1993 were not used for covariate selection and spatial interpolation of the parameters of organic model horizons 1 and 3 in order to avoid a possible bias in the predictions because of changes in SOM content over time.

Mapping the depth function parameters

As explained above, observed soil type (considered to be the true soil type) was used as covariate in the the universal (co)kriging models of the depth function parameters. This implies that actual soil type at unsampled locations is required for prediction (Chapter 3). Actual soil type, however, is unknown at these locations but can be represented with a probability model. In this case the probability model is

the digital soil type map: at each location—a 25-m pixel—it provides a probability distribution of ten soil types. This implies that at each prediction location the depth function parameters of all model horizons of the soil types with a probability greater than zero must be predicted. From the predicted soil type-specific parameters and the soil type-specific depth function structures the depth function of SOM can be reconstructed for each soil type at each prediction location. By combining the soil type-specific depth functions with the probability distributions from the soil map a probability distribution of depth functions was obtained for each location in the study area.

4.2.7 Application and validation of the depth functions

Mapping SOM stocks

The probability distributions of depth functions can be used to generate a variety of maps of which the SOM stock for a specific depth interval is the most obvious one. Malone et al. (2009) provide illustrated examples of the use of SOC and available water capacity depth functions to answer three scenario-based queries.

For the purpose of validation the soil-type specific depth functions were used to compute the SOM stock (in kg m^{-2}) for depth intervals 0–30 cm, 30–60 cm, 60–90 cm and 0–90 cm by integrating the predicted depth functions for these layers. This results in four soil type-specific SOM stocks at each prediction location. The SOM stock for a pixel is then predicted by :

$$C_I(\mathbf{s}) = \sum_{k=1}^K p_k(\mathbf{s}) C_{Ik}(\mathbf{s}) , \quad (4.4)$$

where $C_I(\mathbf{s})$ is the SOM stock at location \mathbf{s} , $p_k(\mathbf{s})$ is the probability of occurrence of soil type k , $k = 1, 2, \dots, K$, and $C_{Ik}(\mathbf{s})$ is the computed SOM stock for a user-defined depth interval for soil type k . The predicted SOM stocks at the four depth intervals were compared with measured SOM stocks at validation locations.

Sampling for validation

The predicted SOM stocks were validated with data from an independent stratified simple random sample (Brus et al., 2011). The strata were three landforms: brookvalleys, coversand ridges and coversand plains, which were manually delineated from the DEM. These strata were used to improve the spreading of the sample locations over the study area. A total of 50 locations were allocated to the strata proportionate to their surface areas. Locations where permission was denied or proved otherwise impossible to sample were replaced with locations from a reserve list. The locations

were sampled in May 2008. At each location an auger bore was made, the soil profile was described and classified and soil samples were taken at depths: 0–30 cm, 30–60 cm, and 60–90 cm.

The samples were oven-dried at 103° C for 6 h, and then sieved and crushed. The SOM content (in mass%) of a dry sample was determined with the weight loss-on-ignition method. The samples were combusted at 550° C for 3 h. The SOM content was determined from the weight difference before and after combustion. The measured SOM content on mass basis was converted to volume basis by multiplying with the bulk density of the soil layer. For this bulk density values were assigned to each soil horizon in the soil profile using the same look-up table that was used for assigning bulk density values to the soil horizons in the calibration dataset. If a validation layer was composed of more than one soil horizon then a depth-weighted average of the bulk density was computed from the bulk densities of the individual soil horizons. The SOM stock was computed by multiplying the volumetric SOM content with the thickness of the sampled soil layer. Total SOM in the 0–90-cm layer was computed by adding the SOM stocks of the three individual layers.

Statistical inference

Two validation indices were computed from the observed and predicted SOM stocks at the three depth intervals at the 50 validation locations. These were: the mean error (ME), which is a measure for the bias of the predictions and the root mean squared error (RMSE), which is a measure for the accuracy of the predictions. The global mean of a validation index is estimated from a stratified random sample by (de Gruijter et al., 2006):

$$\hat{y} = \sum_{h=1}^H a_h \hat{y}_h, \quad (4.5)$$

where a_h is the relative area of stratum h , $h = 1, 2, \dots, H$, and \hat{y}_h is the sample mean of the validation index of stratum h . The stratum sample mean is estimated by:

$$\hat{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}, \quad (4.6)$$

where y_{hi} is the value of the validation index at sampling location i in stratum h , and n_h is the number of sampling locations in stratum h . For the ME y_{hi} is computed as:

$$y_{hi} = \hat{C}_{Ihi} - C_{Ihi}, \quad (4.7)$$

and for the MSE as:

$$y_{hi} = [\hat{C}_{Ihi} - C_{Ihi}]^2, \quad (4.8)$$

where \hat{C}_{Ihi} is the predicted SOM stock at sampling location i in stratum h , and C_{Ihi} is the measured SOM stock. The RMSE is computed as the square root of \hat{y}_{MSE} . The sampling variance of the mean of the validation index is estimated by (de Gruijter et al., 2006):

$$\hat{V}(\hat{y}) = \sum_{h=1}^H a_h^2 \hat{V}(\hat{y}_h), \quad (4.9)$$

where $\hat{V}(\hat{y}_h)$ is the sampling variance of the stratum mean \hat{y}_h , which is estimated by:

$$\hat{V}(\hat{y}_h) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \hat{y}_h)^2. \quad (4.10)$$

4.3 Results

4.3.1 Soil type-specific depth function structures

Table 4.3 presents the model horizon stacks that define the SOM depth function structures of the ten soil types in the study area. It shows that for each soil type two or three model horizons were used. Note that some soil types have the same depth function structure. Soil type-specific depth function were constructed using the soil type-specific depth function structures and the associated parameters (Table 4.3). For instance, the depth function for soil type mP is given by:

$$C_V(z) = \begin{cases} C_2 & \text{for } z \leq d_2 \\ C_3 & \text{for } d_2 < z \leq d_2 + d_3 \\ C_a \exp(-k[z - (d_2 + d_3)]) & \text{else,} \end{cases} \quad (4.11)$$

where $C_V(z)$ is the SOM content (kg m^{-3}) at depth z from the soil surface, C_2 , C_3 , d_2 , d_3 are the SOM contents and depths of model horizons 2 and 3, and C_a and k are the parameters of the exponential decay function. In total six parameters are required to model the depth distribution of SOM for soil type mP . Similar functions were constructed for the other nine soil types.

4.3.2 Transforming soil profile descriptions

Table 4.4 shows some illustrative examples of allocating master soil horizons (Table 4.2) to model horizons. Table 4.5 shows the transformed soil profile descriptions presented in Table 4.4 after recalculation of the SOM content and thickness for model horizons 1 to 4 that comprised multiple, consecutive soil horizons. Now the parameters of model horizons 1 to 4 can be directly derived from the profile descriptions.

Table 4.3: Model horizon sequence used to define the structure of the soil type-specific depth functions of SOM and the associated parameters that describe the functions.

Soil type*	Model horizon					Depth function parameters
	1	2	3	4	5	
<i>P</i>	x		x		x	$C_1, d_1, C_3, d_3, C_a, k$
<i>mP</i>		x	x		x	$C_2, d_2, C_3, d_3, C_a, k$
<i>PY</i>	x		x		x	$C_1, d_1, C_3, d_3, C_a, k$
<i>mPY</i>		x	x		x	$C_2, d_2, C_3, d_3, C_a, k$
<i>BF</i>		x			x	C_2, d_2, C_a, k
<i>PZ</i>		x			x	C_2, d_2, C_a, k
<i>ES</i>		x			x	C_2, d_2, C_a, k
<i>PS</i>		x		x	x	$C_2, d_2, C_4, d_4, C_a, k$
<i>TS</i>		x			x	C_2, d_2, C_a, k
<i>RS</i>		x		x		C_2, d_2, C_4, d_4

* *P* = deep peat soils, *mP* = deep peat soils with mineral surface horizon, *PY* = shallow peat soils, *mPY* = shallow peat soils with mineral surface horizon, *BF* = brown forest soils, *PZ* = podzols, *ES* = earth soils, *PS* = plaggen soils, *TS* = glacial till soils, *RS* = raw sand soils.

The five soil horizons that make up model horizon 5 of profile 1 in Table 4.5 were used to fit the C_a and k parameters of the exponential decay function for this profile. The parameters of this model horizon in the other soil profile descriptions were fitted likewise.

4.3.3 Deriving the depth function parameters

The parameters of the exponential decay function were estimated for each model horizon 5 building block in the dataset by minimizing Eq. 4.3. To evaluate the predictive capability of the decay functions, each function was used to predict the SOM stock of the model horizon 5 building block for which it was fitted. The predicted stock was then compared to the stock computed from the soil horizons that make up the building block; see for example profile 3 in Table 4.5. Fig. 4.4 (left panel) shows a scatter plot with predicted versus observed stocks ($R^2 = 0.98$). Each point represents a model horizon 5 building block in the soil profile dataset. The ME of the predicted stocks is -0.5 kg m^{-2} and the RMSE is 2.1 kg m^{-2} . This is a considerable improvement—less bias and larger accuracy—compared to the results one would obtain from an exponential function fitted to SOM contents (kg m^{-3}) assigned to the midpoints of the soil horizons: Fig. 4.4 (right panel), $R^2 = 0.91$, ME = 1.2 kg m^{-2} , RMSE = 6.2 kg m^{-2} .

Fig. 4.5 displays boxplots of the ten depth function parameters and Table 4.6 lists the soil type parameter means. These results show that the depth function parameter values and their spreads differ between soil types, even when the functions are modelled with identical model horizon sequences. This is what one would expect

since different environments in which the soils formed result in different parameter values. For instance, the large values for the C_2 parameter of soil types mP and mPY are a result of yearly mixture of some subsoil peat through the topsoil during ploughing. The intermediate values for soil types ES and PS are an effect of poor drainage (ES), which results in less favourable conditions for SOM decomposition, and the centuries long application of a SOM-rich mixture of sheep manure and heath sods that resulted in the formation of plaggen soils (PS). The parameters in each model horizon sequence were checked for mutual correlations. The decay function parameters C_a and k were correlated ($r = 0.60$). The other parameters were mutually uncorrelated.

Table 4.4: Six examples of allocating model horizons to soil horizons.

Depth (cm)	Horizon	Master horizon	SOM (kg m^{-3})	Model horizon	Depth (cm)	Horizon	Master horizon	SOM (kg m^{-3})	Model horizon
<i>1. PZ - cultivated peatlands^a</i>					<i>2. PZ - forest</i>				
0–20	Ap	A3	138	2	0–2	Che	C2	14	2
20–30	Cw	C1	200	5	2–50	A/B/E	MM	47	5
30–45	Eu	E	77	5	50–60	Bhs	Bs	16	5
45–65	Bhe	Bh	54	5	60–100	Cu1	C2	5	5
65–90	BCe	BC	16	5	100–200	Cu2	C2	3	5
90–150	Ce	C2	5	5					
<i>3. mP - cultivated peatlands</i>					<i>4. mP - cultivated peatlands^b</i>				
0–20	Ap	A3	42	2	0–25	A3	A2	70	2
20–55	Cw1	C1	225	3	25–50	Cu	C1	280	3
55–60	Cw2	C1	88	3	50–70	Ce	C2	30	3
60–65	Ahb	Ab	83	5	70–110	Cu	C1	298	3
65–75	Bhe	Bs	65	5	110–135	Cer	C2	5	5
75–85	BCe	BC	8	5	135–150	Cr	C2	5	5
85–140	Ce1	C2	5	5					
140–180	Ce2	C2	0	5					
<i>5. P - highmoor area^c</i>					<i>6. RS - drift-sand complex^d</i>				
0–12	Cu1	C1	225	3	0–3	Cem	C2	7	2
12–30	Cu2	C1	180	3	3–30	Cbm	C2	17	2
30–86	Cr1	C1	180	3	30–60	Cg	C2	8	4
86–140	Cr2	C1	228	3	60–65	Cu	C1	120	4
140–150	Cr	C2	31	5	65–90	Cgr	C2	25	4
					90–110	Cr	C1	245	4
					110–120	Eub	E	23	4
					120–140	Bhb	Bs	186	4
					140–155	Bhe	Bs	78	4
					155–200	Ce	C2	0	4

^a Soils classified as *PZ* (podzol) in the cultivated peatlands sometimes have peat remains between the sandy topsoil and podzolic subsoil. Such layer was considered a part of model horizon 5.

^b Deep peat soil with mineral topsoil which has a sandy layer (50–70 cm) in the peat layer because of deep cultivation. Such inclusion is considered part of model horizon 3.

^c Peat soil in a highmoor area. Here the A1 topsoil (well-decomposed organic horizon) is absent. The value of depth function parameter d_1 is 0 for this profile.

^d Raw sand soil from a drift-sand complex. This profile shows the heterogeneity of the subsoil. The first two horizons are regarded as the topsoil (model horizon 2) based on morphology: there is formation of a micro-podzol in the drift sand indicated by the suffix 'm' in the horizon code. The third horizon is unaltered drift sand that overlays a shallow organic horizon (C1), possibly an old forest floor. Below this horizon another horizon with unaltered drift sand is found. At the bottom of the profile another organic horizon (C1) is found that overlays an alluvial horizon that is part of the original podzol profile that was found at the soil surface.

Table 4.5: Transformed soil profile descriptions from Table 4.4. Descriptions like these were used to derive the parameters of depth functions.

Depth (cm)	Model horizon	SOM (kg m^{-3})	Depth (cm)	Model horizon	SOM (kg m^{-3})
1. PZ - cultivated peatlands			2. PZ - forest		
0-20	2	138	0-2	2	14
20-30	5	200	2-50	5	47
30-45	5	77	50-60	5	16
45-65	5	54	60-100	5	5
65-90	5	16	100-200	5	3
90-150	5	5			
3. mP - cultivated peatlands			4. mP - cultivated peatlands		
0-20	2	42	0-25	2	70
20-60	3	208	25-110	3	229
60-65	5	83	110-135	5	5
65-75	5	65	135-150	5	5
75-85	5	8			
85-140	5	5			
140-180	5	0			
5. P - highmoor area			6. RS - drift-sand complex		
0-140	3	202	0-30	2	16
140-150	5	31	30-120	4	73

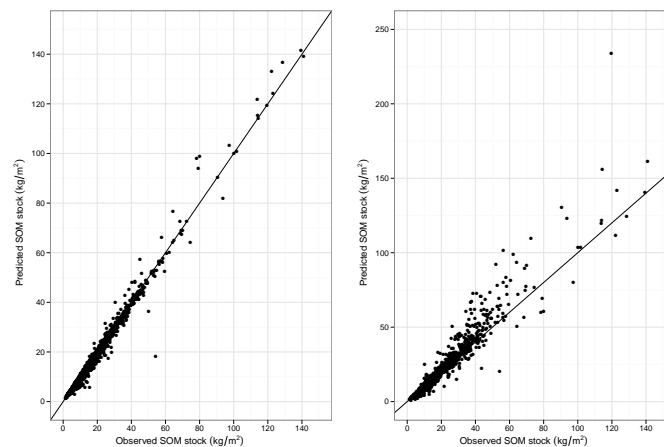


Figure 4.4: Observed and exponential decay function fitted SOM stock at the data locations.

Some attention should be given to model horizon 4, which is used as building block in the depth function of soil types *PS* (plaggen soils) and *RS* (raw sand soils). For soil type *PS* the model horizon 4 building block is used to model the SOM depth profile of buried A-horizons and for soil type *RS* the extremely heterogeneous subsoil (see section 4.2.5). Consequently, the soil horizons that comprise this building block are morphologically very different (contrary to the soil horizons in other model horizon building blocks). This has implications for the mapping of the depth function parameters of the model horizon 4 building block. Because of the highly contrasting landscapes in which soils *RS* (drift-sand complexes) and *PS* (agricultural, open-field complexes) occur and the differences in soil forming processes it is likely that the

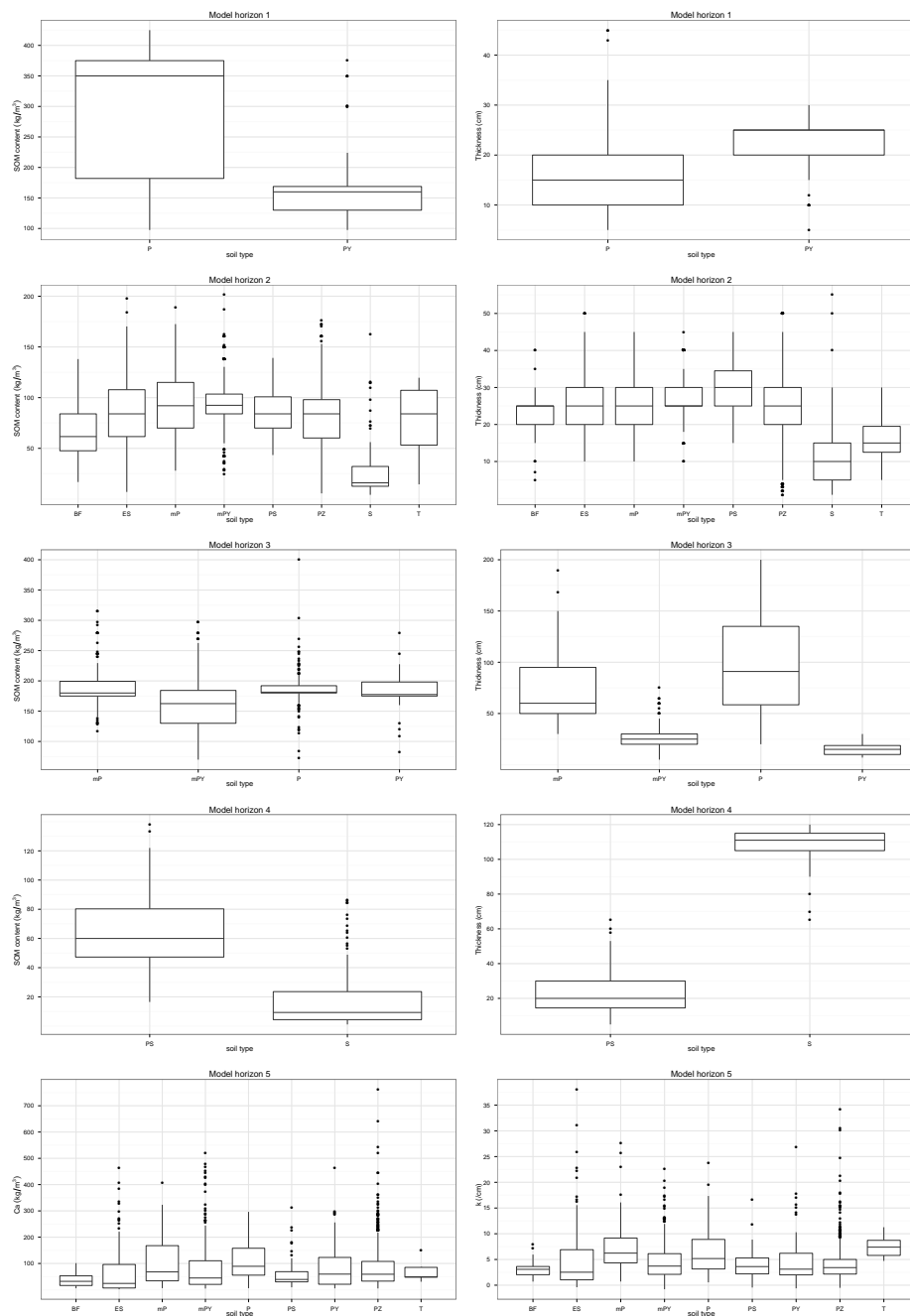


Figure 4.5: Boxplots of the soil type-specific distributions of the ten depth function parameters.

Table 4.6: Soil type means of the parameters of the depth function. Parameters C_i and d_i are the SOM content (kg m^{-3}) and depth (cm) of model horizon i , and C_a (kg m^{-3}) and k (m^{-1}) are the parameters of the exponential decay function. The number between brackets is the number of data used to compute the associated mean, and n is the total number of data on the depth function parameters in the dataset.

Model horizon	Parameter	n	Soil type									
			P	mP	PY	mPY	BF	PZ	ES	PS	TS	RS
1	C_1	252	289(171)		165(81)							
	d_1	272	15(191)		22(81)							
2	C_2	1777		96(161)		95(364)	65(33)	79(770)	89(183)	86(107)	78(11)	26(148)
	d_2	1777		26(161)		27(364)	23(33)	24(770)	26(183)	30(107)	17(11)	12(148)
3	C_3	741	187(190)	190(161)	183(26)	166(364)						
	d_3	797	95(191)	74(161)	5(81)	26(364)						
4	C_4	228								66(71)		18(157)
	d_4	107								15(107)		
5	C_a	1473	111(58)	106(49)	85(74)	85(285)	39(31)	84(731)	69(142)	58(98)	73(5)	
	k	1473	6.5(58)	8.0(49)	4.9(74)	4.8(285)	3.1(31)	4.1(731)	5.0(142)	3.9(98)	7.6(5)	

spatial distribution of the depth function parameters is controlled by different factors and processes. These soils therefore require different sets of covariates for spatial prediction. Furthermore, the assumption that the depth function parameters of the model horizon 4 building block of soil type PS are spatially correlated with those of soil type RS seems implausible from a pedologic point of view. Also, the spatial dependency structure of depth function parameters will likely be very different between the two soil types. Therefore, it was decided to use soil type-specific models for spatial prediction of the depth function parameters of model horizon 4, instead of defining one geostatistical model for each parameter in which soil type is used as covariate like what is done for the other model horizons.

The thickness parameter d_4 of the depth function for soil type RS has no pedologically defined boundary: it is the remaining depth of the profile below the topsoil. This parameter was therefore not spatially predicted but computed as 120 cm minus the predicted thickness of the overlying model horizon 2. This implies that horizon data up to a depth of 120 cm were used to compute the SOM content parameter C_4 for soil type RS (see profile 6 in Tables 4.4 and 4.5).

4.3.4 Selection of environmental covariates and geostatistical modelling

Ten unique depth function parameters model the SOM depth profiles of the ten soil types in the study area. A geostatistical model was constructed for each of the parameters of model horizons 1, 2, 3, and 5. The two parameters of model horizon 4 were modelled separately for soil types that use this horizon as depth function building block.

Table 4.7 shows for each depth function parameter the environmental covariates that were selected in the OLS regression models. The residuals of the regression models

Table 4.7: Environmental covariates used in the regression models for the parameters of the model horizons. The R^2 is the adjusted coefficient of determination of the regression model, and n is the number of data used for model selection and spatial prediction.

Model horizon	Parameter	n	R^2	Covariates
1	C_1	252	0.68	SOILTYPE+COV5CL+RECLAM2CL
	d_1	272	0.52	COV5CL+LC19003CL+MOOROUTPEATCOL
2	C_2	1777	0.35	SOILTYPE+LC19003CL+PEATCOL+DISTURBED+RECLAM3CL
	d_2	1777	0.27	SOILTYPE+COV4CL+DISTURBED+DUNES+PEATCOL
3	C_3	741	0.09	SOILTYPE+COV4CL+RECLAM2CL
	d_3	797	0.64	SOILTYPE+RECLAM3CL+MOOROUTPEATCOL
4 (PS)	C_{4ps}	71	0.18	GD+PEATCOL
	d_{4ps}	107	0.12	COV2CL
4 (S)	C_{4s}	157	0.17	DISTURBED+PEAT+PEATCOL
	d_{4s}	-	-	Not modelled with covariates but computed as $120 - \hat{d}_{2s}$
5	$\log(C_a)$	1473	0.09	SOILTYPE+PEATLAND+RECLAM4CL+GD
	$\log(k)$	1473	0.06	SOILTYPE+COV2CL+RECLAM3bCL+GD+FENPEAT+REL250

for parameters C_a and k showed strong positive skew. The raw data values were transformed to natural logarithms and regression models were fitted again to the transformed data values. The residuals of the other parameters were weakly or mildly skewed and were kept on their original scale. The correlation between the covariates and parameters of the topsoil model horizons (1 and 2) was generally much stronger than that between the covariates and parameters of the subsurface horizons (3, 4 and 5) with the exception of the depth parameter of model horizon 3. Land cover and reclamation period were found to be strong predicting covariates for most depth function parameters. Also geographical covariates such as the cultivated peatlands and the former highmoor areas outside the current peat colonial landscape proved to be strong predictors. Soil type was a significant covariate for all parameters of model horizons 1, 2, 3, and 5, except parameter d_1 . Here the effect of soil type is accounted for by the other covariates.

The parameters C_i and d_i were spatially predicted using universal kriging with a variance model estimated by residual maximum likelihood, while the correlated log-transformed parameters C_a and k were predicted using universal co-kriging with a linear model of co-regionalization estimated by the method-of-moments estimator. The trend models included the covariates that were selected by OLS-regression. In total nine universal kriging models and one universal co-kriging model were constructed for spatial interpolation of the depth function parameters. Table 4.8 presents results for parameter estimates of the variance models for each depth function parameter. Autocorrelation of the residuals is relatively strong for parameters C_1 , d_1 , C_2 , C_{4ps} , C_{4s} ; with 40 to 90% of the residual variance spatially structured at short to medium distances. The selected covariates for parameter d_3 explained a large part of the observed variance in the data. The weak autocorrelation of the residuals of this parameter indicates that the important predictors were included in the trend model. The variogram of parameter C_3 is almost pure nugget. The covariates in the trend

Table 4.8: Parameter estimates of the variance models of the depth function parameters. Parameters C_i and d_i are estimated by REML and C_a and k by the method-of-moments estimator using a linear model of co-regionalization.

Model horizon	Parameter	Model type	Variance	a^a (m)	Spatial dependence ^b
1	C_1	Sph	3864	2855	0.44
	d_1	Sph	45.6	4374	0.41
2	C_2	Exp	859	903	0.53
	d_2	Sph	59.3	3645	0.14
3	C_3	Sph	1380	272	0.07
	d_3	Sph	685	2731	0.27
4 (PS)	C_{4ps}	Sph	527	2881	0.77
	d_{4ps}	Sph	214	867	0.91
4 (S)	C_{4s}	Exp	403	537	0.59
	$\log(C_a)$	Sph	1.01	10000	0.02
5	$\log(k)$		0.39		0.14
	$\log(C_a) * \log(k)$		0.04		0.00

^a Distance parameter of the exponential model (practical range is $3a$) and range of the spherical model

^b The spatial dependence is the ratio between spatially structured variance and total variance [partial sill/(nugget+partial sill)](Lark and Cullis, 2004).

model explained only a very small part of the observed variance. This indicates that most of the variation in the data is spatially unstructured and cannot be predicted by geostatistical methods. This is not surprising given the source of the data on model horizon C_3 . Roughly 70% of the data was derived from soil profile descriptions from soils in the cultivated peatlands. Here the SOM content of the organic horizon in the subsoil is very heterogeneous both laterally and vertically because of strong anthropogenic disturbance as explained in section 4.3.1 and shown by Fig. 4.1d. The auto-variogram of the residuals of the log-transformed C_a parameter and the cross-variogram of the log-transformed C_a and k parameters were pure nugget variograms.

4.3.5 Spatial prediction of the depth functions

The universal (co)kriging models were used to predict soil type-specific depth function parameters, which was accomplished by applying the universal kriging model for prediction of a given parameter as many times as there are soil types whose depth distribution is partly described with this parameter. For instance, the universal kriging model of parameter C_1 was applied twice: once given the soil type at each prediction location is P and once given the soil type is PY . In this way, a total of 50 soil-type specific parameters were predicted at each location: 25 soil type-model horizon combinations (Table 4.3) times two parameters per combination. With the 50 predicted parameters and the soil type-specific depth function structures, the depth function was reconstructed for each of the ten soil types at each prediction location in the study area. Each of the ten depth functions at each location has an associated probability, which was derived from the digital soil type map.

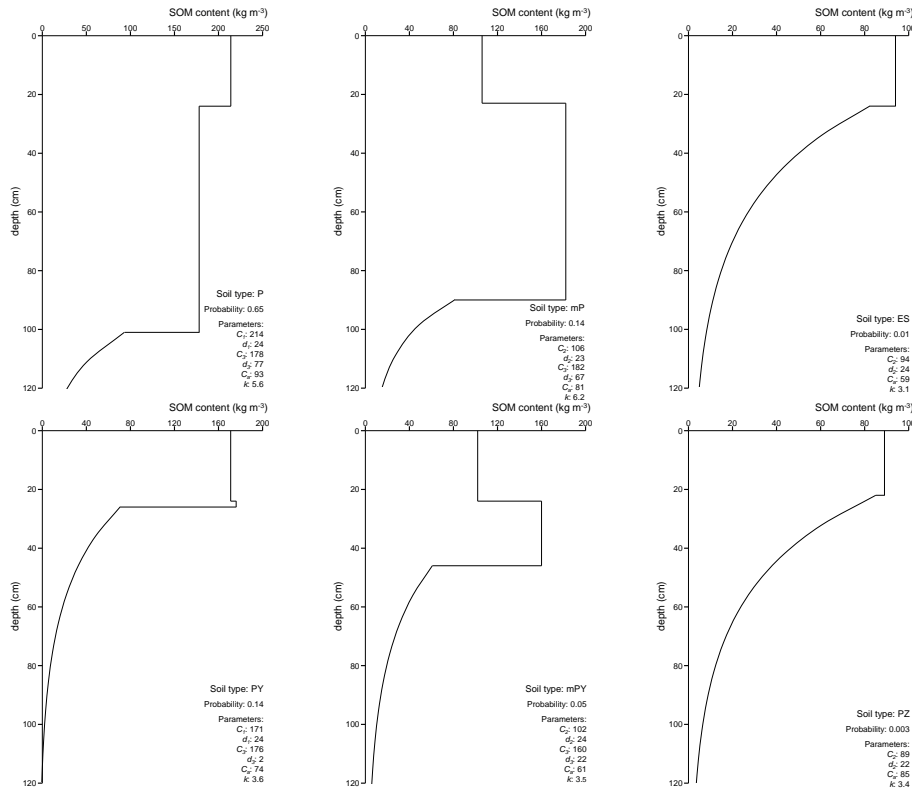


Figure 4.6: Predicted depth functions for soil types with positive probability at one location.

As an example, Fig. 4.6 shows the depth functions for six soil types at one location in the study area with the associated probabilities and predicted depth function parameters. Four peat soils (*P*, *mP*, *PY* and *mPY*) and two mineral soils (*ES* and *PZ*) have a probability greater than zero at this location. The predicted depth functions are as one would expect on the basis of knowledge on the profile morphology of these soils. The topsoil SOM content (C_1) of deep peat soil *P* is larger than that of the subsurface organic layer (C_3) which can be explained by the larger bulk density of the former. The very shallow subsurface organic layer in shallow peat soil *PY* is a modelling artifact in this example, and will generally not be distinguished from the topsoil horizon in the field. The mineral soils *ES* (hydromorphic earth soils) and *PZ* (podzols) differ in topsoil SOM content (C_2) and the SOM content in the upper part of the subsoil. Topsoil SOM is larger for the earth soil than for the podzol, which is explained by the generally wetter conditions under which the former occurs. The larger SOM content in the upper subsoil of the podzol compared to the earth soil, is explained by SOM accumulation in the podzol-B horizon, which slowly decreases with depth. The earth

soil lacks this horizon. Instead, it often has an AC-transition horizon at this location in the profile, which has a lower SOM content than the podzol-B horizon.

4.3.6 Application and validation of the depth functions

Mapping SOM stocks

Fig. 4.7 shows the predicted SOM stocks for the four depth intervals. The spatial pattern of SOM is clearly controlled by the soil type with the largest probability at a prediction location (Fig. 4.2B). The largest stocks are found in the brook valleys where peat soils prevail. Other areas with large SOM stocks are in the northwest and southwest corners. These areas were once covered with peat and peat remains can still be found in the soils. The SOM stock for the 30–60-cm layer is the most variable. This is explained by the profile morphology of the soils. Within the mineral soil group the soil profiles differ most around this depth: podzols typically have the podzol-B horizon at this depth, brown forest soils a brown B-horizon containing some moder humus, earth soils an AC- or C-horizon, plaggen soils a dark, buried A-horizon and raw sand soils a C-horizon. It is therefore not surprising that at this depth interval the spatial variation is largest. Also within the peat soil group this is the depth with the largest between-profile variation. Soils *P* and *mP* generally have peat around this depth, soil *PY* mineral material and soil *mPY* can have peat or mineral material. The topsoil of the mineral soils is much more homogenous than the subsoil as a result of intensive land use (regular tillage and organic manure application). This results in small between-soil type variation; especially for the most common mineral soils in the study area, which are the podzol, earth soil and plaggen soil (see Table 4.6). Furthermore, most mineral soils will have (sandy) C-material below 60 cm depth. SOM contents will therefore vary little between these soils at depths exceeding 60 cm.

The average mapped SOM stock in the soils in the study area for the 0–30-cm layer is 28 kg m^{-2} ; for the 30–60-cm layer, 18 kg m^{-2} and for the 60–90-cm layer, 10 kg m^{-2} . The stock for the 0–90 cm-layer ranged between 13 and 182 kg m^{-2} . The largest stocks are predicted in areas mapped as peat soils (Fig. 4.2B), i.e. where peat soils have the largest probability of occurrence. For mapped soil type *P* the average predicted stock for the 0–90-cm layer is 141 kg m^{-2} ; for soil type *mP* 124 kg m^{-2} ; for soil type *PY* 71 kg m^{-2} ; and for soil type *mPY* 73 kg m^{-2} . Medium stocks, on average around 45 kg m^{-2} in areas mapped as soil types *PS*, *PZ* and *ES* and 35 kg m^{-2} in the areas with mapped soil types *BF* and *TS*. The lowest stocks (21 kg m^{-2}) are predicted in the two drift-sand complexes in the northern part of the study area where raw sand soils dominate.

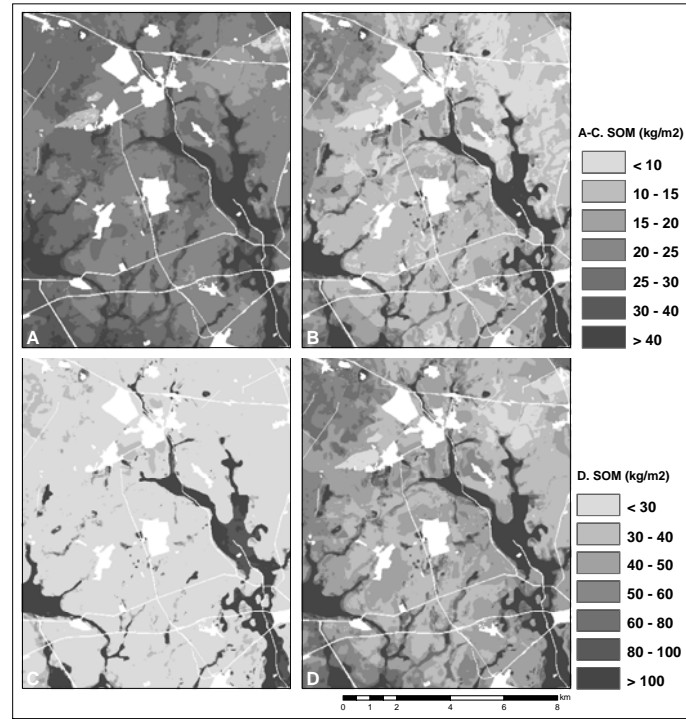


Figure 4.7: Predicted SOM stock (kg m^{-2}) at 0–30 cm (A), 30–60 cm (B), 60–90 cm (C) and 0–90 cm (D) depth intervals.

Validation

Table 4.9a shows the validation results of the predicted SOM stocks for the four validation layers. There is modest positive bias (over-estimation) for the 0–30-cm layer and small positive bias for the subsurface layers. The RMSE is largest for the 30–60-cm layer. This is expected given the strong between-soil type variation for this depth interval. The predicted stocks are probability weighted averages of predicted soil type-specific stocks. Large uncertainty in the soil type occurring at a validation location will predominantly affect the accuracy of the predictions of the 30–60-cm layer. The R^2 -values were computed from a scatter plot of observed versus predicted and decrease from the top layer to the bottom layer. These values show good results for the upper 30 cm but poor results for the bottom two layers. The R^2 -value of the 0–90-cm layer is a modest 0.46.

4.4 Discussion

This study provides an example how general pedological knowledge can be used to define soil-type specific depth functions of SOM. This is in contrast with previous, more data-driven studies (Minasny et al., 2006; Malone et al., 2009; Mishra et al., 2009) in which one type of function, typically a spline or exponential decay, is used irrespective of soil type and under the assumption that the depth-wise variation of soil properties is entirely continuous. Using pedological knowledge for defining depth function structures might result in a more realistic representation of the depth distribution of soil properties when there are large differences in profile morphology between soil types or when there has been strong human influence on profile development. Furthermore, the proposed method explicitly uses a soil type map, which is a valuable source of information on the three-dimensional spatial variation of soil properties. Conventional soil maps are available for many areas in the world. If such maps can be upgraded with information on the areal proportions of soil types within the mapping units, then the method presented in this chapter enables the exploitation of these conventional maps for three-dimensional mapping of soil properties. This is in contrast to the more data-driven methods mentioned before.

The presented method demands a rich soil dataset. It requires complete soil profile descriptions and a soil type map, which may limit application in data-poor environments. Care should be taken that the number of depth function parameters remains within practical limits. Ten parameters were used to model the soil type-specific depth functions (Table 4.6). This required nine universal kriging models and one universal co-kriging model to predict these parameters. Soil type was used as covariate in these models to account for the effect of soil type on parameter values. Hereby assuming that the covariate effects on parameter values were similar for all

Table 4.9: Estimated validation indices for the predicted SOM stocks for pedometric and conventional depth functions. ME is the mean error, MSE the mean squared error, RMSE the root mean squared error and R^2 the coefficient of determination. The quantity between brackets is the standard error of the estimates.

Soil layer	ME	MSE	RMSE	R^2	Soil layer	ME	MSE	RMSE	R^2
<i>a. Pedometric depth functions</i>					<i>c. Conventional depth functions</i>				
0–30 cm	4.5 (1.2)	98 (23)	9.9	0.75	0–30 cm	1.1 (1.3)	81 (26)	9.0	0.66
30–60 cm	1.1 (1.9)	179 (72)	13.4	0.23	30–60 cm	-4.1 (2.0)	218 (93)	14.8	0.17
60–90 cm	1.1 (1.4)	94 (43)	9.7	0.09	60–90 cm	-2.4 (0.8)	40 (13)	6.4	0.19
0–90 cm	6.7 (3.3)	585 (206)	24.2	0.46	0–90 cm	-5.4 (3.2)	553 (178)	23.5	0.46
<i>b. Pedometric depth function, actual soil type</i>					<i>d. Conventional depth function, actual soil type</i>				
0–30 cm	4.0 (1.1)	82 (15)	9.1	0.78	0–30 cm	-0.2 (1.1)	70 (18)	8.4	0.73
30–60 cm	-0.2 (1.4)	103 (41)	10.1	0.54	30–60 cm	-6.2 (1.7)	187 (78)	13.7	0.35
60–90 cm	0.1 (1.4)	101 (43)	10.0	0.11	60–90 cm	-3.5 (0.8)	49 (16)	7.0	0.19
0–90 cm	3.9 (2.8)	424 (120)	20.6	0.59	0–90 cm	-9.9 (2.3)	431 (116)	20.8	0.67

soil types. This might not necessarily be true. An alternative would be including covariate interactions in the trend parts of the geostatistical models but for reasons of simplicity this was not done. Another alternative would be to calibrate one geostatistical model for each soil type-specific parameter. For this study (with only ten soil types) this would already have required calibration of 50 geostatistical models (Table 4.2). Besides, this would have complicated estimation of trend coefficients and variogram parameters as for soil types *BF* and *TS* only a few dozen profile descriptions were available.

For each soil type one depth function structure was defined (Table 4.3), based on a conceptual model of pedogenesis. It was assumed, for reasons of simplicity, that this structure was valid for all soils classified as such. There are exceptions, however, to the defined structures, especially within the podzol soil type. Three general podzol types occur in the study area with different profile morphologies: podzols under agriculture on the Drenthe plateau, forested podzols on the Drenthe plateau, and the podzols in the cultivated peatlands. The depth distribution of SOM of the agricultural podzol, the most common podzol in the study area, is adequately modelled by the depth function used in this study. The forested podzols have horizons that can be disturbed up to depths of 1 m below the surface as a result of deep ploughing before afforestation in the early 20th century. Deeply disturbed subsurface horizons in profile descriptions disturbed the fit of the exponential decay function to the horizon data and resulted in estimates of the C_a parameter that are outside the physical boundaries. Furthermore, the SOM content does not decrease exponentially with depth in such horizons, which was assumed by the depth function model. Forested podzol profiles disturbed deeper than 50 cm below the surface were, therefore, excluded from modelling. This means that the depth function model of soil type *PZ* is not valid in areas with deeply disturbed forested podzols. The podzols in the cultivated peatlands are also characterized by disturbed profiles that may contain inclusions of peat remnants. Thus, in order to properly model the SOM depth profile of the podzol unit, three depth function models would have been required.

Validation with an independent dataset showed that the depth functions provided good estimates of the SOM stock in the upper part of the soil profile. However, the functions performed poorly for the soil subsurface. These results agree with findings of Minasny et al. (2006) and Malone et al. (2009) who also reported a decreasing accuracy of prediction with depth for SOC and available water capacity in a study area in southeast Australia. Vasques et al. (2010) modelled the SOC content at four fixed depth intervals in a watershed in Florida. They also found better results (lower mean absolute error and RMSE) for the 0–30-cm layer than for the 30–60-cm and 60–120-cm layers. Furthermore, they found that environmental covariates explained little spatial variation of organic carbon in the lower part of the profile. In this study, covariates correlated well with topsoil depth function parameters but poorly with subsurface parameters (Table 4.7). Vanwallegem et al. (2010) spatially modelled the vertical variation of the soil horizons of a loess-derived profile in Belgium. They only

found absent or weak correlations between horizon thicknesses and environmental covariates resulting in poor predictions of the spatial variation of these thicknesses. These examples from study areas in different parts of the world with different soils and where different methods and different sets of environmental covariates were used to model the depth profiles of soil properties, illustrate that there is a general challenge of capturing subsurface variation in soil properties or soil horizons by pedometric models.

There is modest bias in the predicted SOM stocks. With a digital soil type map it is possible to quantify the contribution of the uncertainty of the predicted soil types (expressed with probability distributions) to the ME (bias) and MSE. Chapter 3 showed that the uncertainty in the soil map is overestimated, which might affect estimates of soil properties when the soil map is used as predictor. The contribution of soil type uncertainty to the ME and MSE was assessed by validating the SOM stocks predicted by the depth function of the soil type that was observed at the validation locations. Results are shown in Table 4.9b. For the 0–90-cm layer the bias was reduced by 42%, with unbiased predictions for the 30–60 cm and 60–90 cm layers. The MSE of the 0–90-cm layer was reduced by 28% and that of the 30–60-cm layer even with 42%. This indicates that the between-soil type variation is largest for this depth interval. The R^2 -value of this layer improved from 0.23 to 0.54 and that of the 0–90-cm layer from 0.46 to 0.59. These results show that the contribution of the uncertainty in the soil map to the prediction errors is considerable. Other important causes of error are the hand-estimated SOM contents with which the models were calibrated and the expert-knowledge based bulk density values used to convert SOM content from mass to volumetric basis. Unfortunately, there was no information about the accuracy of the hand-estimates and bulk density values so their contribution to the MSE could not be quantified.

The presented approach is closely related to the conventional approach that represents depth profiles of soil properties with representative profile descriptions—the soil horizons of which are characterized by typical values for several soil properties—associated to the map units of a soil type map. The soil type-specific depth functions defined in this chapter, however, are more flexible than conventional functions. The parameters of the functions can vary in space, depending on environmental conditions that can be represented by a set of covariates. In addition, the depth functions are knowledge-based and describe both discontinuous (stepped) and continuous depth-wise variation within a soil profile, which likely better represents the true SOM depth profile of the soil in the study area. Furthermore, a probability distribution of depth functions was obtained, which provides information on the uncertainty of the depth distribution of SOM at each prediction location. This information is typically not available in the conventional approach.

The question now is if the presented pedometric approach to three-dimensional mapping of soil properties indeed performed better than the conventional approach?

To address this question, representative profile descriptions were obtained of the soil types depicted on the soil map (de Vries, 1999). SOM stocks were computed for each validation depth interval for each profile. Eq. 4.4 was applied to predict SOM stock of each depth interval at each validation location. The predicted stocks were compared with the measured stocks (Table 4.9c). The performance of the two approaches was assessed by comparing the squared errors. The conventional approach gave better predictions for the 0–30-cm ($p = 0.06$) and 60–90-cm ($p = 0.06$) layers, while, interestingly, the pedometric approach gave a better prediction for the 30–60-cm layer ($p = 0.11$). This is typically the layer with the largest within-soil type and between-soil type variation of SOM stock. For the 0–90-cm layer there was no difference in performance of the two methods on the basis of the MSE. Predicted SOM stocks with the pedometric approach were less biased for the 30–60-cm and 60–90-cm layers, while the bias for the 0–90-cm layer was slightly lower for the conventional approach. For the conventional approach it was also assessed how well the depth function of the soil type observed at each validation location, predicts the SOM stock (Table 4.9d). Comparing the validation indices from Table 4.9d with those from Table 4.9b, shows that the bias in the SOM stocks of the 30–60-cm, 60–90-cm and 0–90-cm layers as predicted by the pedometric approach was much smaller than for the conventional approach. In addition the pedometric approach gave much better predictions for the 30–60-cm layer ($p = 0.03$). The conventional approach better predicted the SOM stocks of the observed soil type for the 0–30-cm ($p = 0.14$) and 60–90-cm ($p = 0.08$) layers. For the 0–90-cm layer the difference in prediction accuracy was negligible. Although the validation is not conclusive, the modelled depth functions might give a more realistic representation of the vertical variation of soil properties than the discrete (stepped) functions based on representative soil profile descriptions and thus might provide better predictions of the SOM content at smaller depth intervals. A full validation of the predicted depth functions requires many soil samples per validation location, the costs of which readily become prohibitive. Validation for three depth intervals as done here is less costly, but also less conclusive as a large part of the vertical variation can be averaged out within the depth increments.

4.5 Conclusions

This chapter presented and illustrated a method for mapping depth functions of soil organic matter, which integrates general pedological knowledge about SOM depth profiles of the soil types in the study area (for defining the structure of soil type-specific depth functions) and geostatistical modelling (for spatial interpolation of the depth function parameters). When a soil type map is available then this can be a valuable source of information and an attractive starting-point for three-dimensional modelling of soil properties as the soil type map itself can be considered a simple representation of the three-dimensional variation of soil properties. Soil type proved

to be a strong predictor for each of the depth function parameters.

Validation with an independent dataset shows that the depth functions provided adequate estimates of the SOM stock in the 0–90-cm soil layer. However, the depth functions performed much better for the top layer than for subsurface layers. Similar results were reported for other studies, which illustrates that there is a general challenge of capturing subsurface variation in soil properties by pedometric models.

A comparison of the presented pedometric approach with the conventional approach for mapping the three-dimensional variation of SOM showed that there is little difference in performance on basis of the validation results, although the pedometric depth functions presumably give more realistic representations of the vertical variation than the discrete (stepped) functions of the conventional approach. This comparison illustrates that it is not obvious that state-of-the-art pedometric methods provide higher quality soil maps than conventional methods as is often claimed but rarely verified.



Chapter 5

Digital soil type mapping with the generalized linear geostatistical model

This chapter introduces the generalized linear geostatistical model (GLGM) for soil type mapping and investigates if spatial prediction with this 'spatial' model results in a soil map of greater accuracy than a map obtained using a non-spatial model, i.e. a model that ignores spatial dependence in the soil type variable. The GLGM is central to the framework of model-based geostatistics, which is considered state-of-the-art in digital soil mapping. A pragmatic approach was adopted in which the five soil types in a case study area in the cultivated peatlands were modelled separately with a binomial logit-linear GLGM. Prediction with soil type-specific GLGMs resulted in five binomial probabilities at each prediction location, which were scaled to multinomial probabilities. A soil map was created from the predicted probabilities by selecting the soil type with maximal probability. In addition two non-spatial models were used to map soil type. These were the multinomial logit model and the generalized linear model for Bernoulli distributed data. Validation with independent probability sample data showed that the spatial model did not give more accurate predictions than the two non-spatial models.

Based on: B. Kempen, D.J. Brus, G.B.M. Heuvelink
Submitted to Geoderma

5.1 Introduction

Incorporating spatial dependence in statistical prediction models for categorical soil attributes such as soil type or texture class is not as straightforward as for continuous attributes. Because of this, many applications focus on methods that ignore spatial dependence in the categorical soil variable, such as multinomial logistic regression (Hengl et al., 2007b; Debella-Gilo and Etzelmüller, 2009, Chapter 2), artificial neural networks (Behrens et al., 2005) and classification and regression trees (Minasny and McBratney, 2007a; Nelson and Odeh, 2009; Stum et al., 2010). Among the methods that accommodate spatial dependence in predictive mapping of categorical soil attributes, perhaps indicator kriging (IK) is the most widely used (e.g. Bierkens and Burrough, 1993; Oberthür et al., 1999; Goovaerts, 2001). However, IK is known to have some theoretical and practical shortcomings such as probabilities that are outside the interval $[0,1]$ or probabilities that do not sum to unity (D'Or and Bogaert, 2004; Papritz, 2009). More recently Bayesian maximum entropy was introduced as a statistically valid alternative to IK and allows integration of different data sources of various quality (D'Or and Bogaert, 2004; Brus et al., 2008). A drawback of Bayesian maximum entropy is that the method quickly becomes computationally prohibitive when the number of outcome classes is large. Modelling categorical data as a Markov random field is a promising approach but applications in the environmental sciences in general (Norberg et al., 2002; Kasetkasem, 2005; Hou et al., 2011) and soil science in particular (Hartman, 2006; Heuvelink, 2007) are still limited. Besides, application of the Markov random field model has its drawbacks. The model is numerically complex, estimation of its parameters can be difficult given the potentially large number of model parameters and statistical validity of the model is not easily verified (Heuvelink, 2007).

A potentially interesting alternative to the methods above is the generalized linear geostatistical model (GLGM) (Diggle et al., 2002; Ben-Ahmed et al., 2010). The GLGM was introduced by Diggle et al. (1998) and is central to the model-based geostatistical framework. This framework extends the linear geostatistical (kriging) method for Gaussian-distributed data to situations in which the stochastic variation in the data is known to be non-Gaussian. In model-based geostatistics one assumes that the observed responses are realizations of a spatial random process. Estimation of the parameters of this process is based on an explicitly declared stochastic model, that is supposed to have generated the responses (Lark and Cullis, 2004; Diggle and Ribeiro Jr., 2007). For example, for continuous soil pollution data the log-normal model might be the natural candidate for the stochastic process that generated the data. The joint probability density function of the data is then the multivariate log-normal distribution, which is completely defined by the mean function and covariance function of the log-transformed variable. The model parameters, in this example the mean and covariance parameters, are estimated by likelihood-based methods. For soil mapping, model-based geostatistics has so far only been used for

mapping continuous soil attributes (e.g. Lark and Cullis, 2004; Lark et al., 2006).

Soil type data can be considered multinomial data when the number of outcome categories, the soil types, is larger than two. The natural candidate model for multinomial data is the multinomial logit model (MLM). A logical choice for modelling spatially correlated multinomial data is an MLM with spatially correlated random effects, but to our best knowledge there are no examples in the literature that describe such model. Nor are we aware of readily available software to fit spatial MLMs. An alternative to the MLM is the Poisson log-linear model. It is well-known in the statistical literature that multinomial data can be modelled with the Poisson log-linear model with extra parameters (one per sampling object) added to the linear predictor (Palmgren, 1981; McCullagh and Nelder, 1989). These extra parameters ensure that the probabilities of the outcome categories (soil types, in this case) sum to unity. Chen and Kuo (2001) prove that multinomial and Poisson log-linear fixed effect models are likelihood equivalent. They assert that this equivalence holds when the log-linear model is extended with random effects, but do not provide a proof. We have reservations about this likelihood equivalence in the presence of random effects and therefore propose to model soil types as Bernoulli-distributed random variables. Each individual soil type is then considered a spatially dependent Bernoulli variable that is modelled with a GLGM for Bernoulli-distributed data. This yields a statistically valid spatial model for each soil type that predicts its probability of occurrence. Next a pragmatic approach can be taken to obtain a multinomial probability distribution from the estimated soil type-specific Bernoulli distributions. It is this approach that is taken in this work.

The Netherlands has full soil map coverage at scale 1:50 000. This map is the primary source of soil information and is used for a wide variety of purposes. Recent studies have shown that the 1:50 000 soil map has become outdated for areas with peat soils as peat layers are decomposing as a result of intensive land use (de Vries et al., 2010). In Chapter 2 the 1:50 000 soil map for the province of Drenthe was updated using the MLM. In this chapter the GLGM is used to re-map soil type in a cultivated peatland area and it is investigated if soil type mapping with a spatial model results in a map that is of greater accuracy than a map obtained by a non-spatial model.

The objectives of this chapter are to i) use the GLGM to update the soil map of a peatland area using a large set of recent point data, ii) assess the accuracy of the soil map with independent probability sample data and iii) compare the results with soil type maps as predicted with two non-spatial prediction models.

5.2 Theory

Here a brief review is given of the GLGM, the estimation of its parameters and how the model can be used for spatial prediction. In addition it is shown how the GLGM

can be used for spatial prediction of categorical soil variables. For a comprehensive exposition of the GLGM see Diggle et al. (2003) and Diggle and Ribeiro Jr. (2007).

5.2.1 The generalized linear geostatistical model

The linear mixed model (LMM) extends the traditional linear model for Gaussian responses by introducing random effects into the model structure in addition to fixed effects. In the context of model-based geostatistics it is assumed that the random effects are spatially correlated, Gaussian distributed random variables (Lark et al., 2006). This means that spatial dependence in the response variable is modelled by introducing spatial dependence in the random effect of the model. The LMM for geostatistical data can then be written as:

$$Y(\mathbf{x}) = m + U(\mathbf{x}) + \epsilon(\mathbf{x}) , \quad (5.1)$$

where the $Y(\mathbf{x})$ are the response variables at spatial location $\mathbf{x} \in D$, $D \subset \mathfrak{R}^2$, m is a fixed effect which can be a constant or a trend (for the latter m depends on spatial location \mathbf{x}), $U(\cdot) = \{U(\mathbf{x}) : \mathbf{x} \in D\}$ is a second-order stationary, Gaussian distributed, spatial process with zero mean and variance σ^2 , and the $\epsilon(\mathbf{x})$ are mutually independent, Gaussian distributed random variables with zero mean and variance τ^2 . In geostatistics the term ϵ is the nugget effect, which represents measurement error and short-range variation (Lark et al., 2006).

The generalized linear mixed model (GLMM) extends the classical LMM to non-Gaussian error distributions that belong to the exponential family, including Bernoulli and Poisson (McCullagh and Nelder, 1989). In similar fashion the LMM for geostatistical data can be extended to accommodate modelling of non-Gaussian distributed responses. Diggle and Ribeiro Jr. (2007) refer to this model as the generalized linear geostatistical model. The GLGM has the following components:

1. *Signal process* $S(\cdot)$. A real-valued Gaussian spatial process defined on D , with $E[S(\mathbf{x})] = \mathbf{d}(\mathbf{x})^T \boldsymbol{\beta}$, $\text{Var}[S(\mathbf{x})] = \sigma^2$, correlation function $\rho(\mathbf{h}) = \text{Corr}\{S(\mathbf{x}), S(\mathbf{x}')\}$, where \mathbf{h} is the Euclidean distance between \mathbf{x} and \mathbf{x}' , and $\text{Cov}\{S(\mathbf{x}), S(\mathbf{x}')\} = \sigma^2 \rho(\mathbf{h})$. Here $\mathbf{d}(\mathbf{x})^T \boldsymbol{\beta}$ is a spatial trend, where $\mathbf{d}(\mathbf{x})$ is a vector of explanatory variables at location \mathbf{x} and $\boldsymbol{\beta}$ is a vector of trend coefficients (usually the first component of $\mathbf{d}(\cdot)$ is taken as a constant, and if this is the only component then $S(\cdot)$ has constant mean). The signal process $S(\mathbf{x})$ can be written as:

$$S(\mathbf{x}) = \mathbf{d}(\mathbf{x})^T \boldsymbol{\beta} + U(\mathbf{x}) , \quad (5.2)$$

where $U(\mathbf{x})$ is as defined in Eq. 5.1.

2. *Measurements* $Y(\mathbf{x}_i)$, $i = 1, 2, \dots, n$. The model-based approach assumes that the stochastic process $S(\cdot)$ is not directly observable. Instead, the available

data consist of measurements $y(\mathbf{x}_1), \dots, y(\mathbf{x}_n)$, which are realizations of random variables $Y(\mathbf{x}_1), \dots, Y(\mathbf{x}_n)$ (Diggle et al., 2003). The $Y(\mathbf{x}_i)$ follow a common distributional family (e.g. Bernoulli, Poisson, binomial or Gaussian) depending on the mechanism that generated the data, and are mutually independent conditional on the signal $S(\cdot)$. The responses have conditional means $E[Y(\mathbf{x}_i)|S(\cdot)]$ and possibly additional parameters that characterize higher order moments of the distribution (e.g. the variance τ^2 in case of the Gaussian distribution).

3. *Link function* $g(\cdot)$. The link function links the conditional means of the $Y(\mathbf{x}_i)|S(\cdot)$ to the linear predictor $S(\mathbf{x}_i)$. The choice of the link function depends on the distribution of the responses. The natural candidate link function for binomial distributed measurements is the logit, that for Poisson distributed measurements the log, and for Gaussian distributed measurements the identity.

The GLGM can now be written as:

$$g(E[Y(\mathbf{x}_i)|S(\cdot)]) = S(\mathbf{x}_i), \quad (5.3)$$

where $S(\mathbf{x}_i)$ is defined by Eq. 5.2. For brevity and notational clarity the spatial location indicator \mathbf{x}_i will be denoted as i in the remainder of this chapter.

5.2.2 The GLGM for digital mapping of soil type

This chapter focusses on the GLGM for categorical responses: of interest is spatial prediction of soil types. For spatial prediction of continuous, Gaussian distributed responses (soil attributes) a GLGM with the identity link function can be used. This is equivalent to using the universal kriging model (kriging with external drift) with variance parameters estimated by residual maximum likelihood (REML), of which Lark and Webster (2006) provide an excellent description.

Observed soil types at the sampling locations can be considered realizations of a multinomial probability distribution, which makes the multinomial logit model the natural candidate model of the data generating mechanism. However, a GLGM for multinomial distributed data cannot be fitted with existing geostatistical software. An alternative is to consider the observed soil types realizations of Bernoulli random variables. The Bernoulli distribution is a special case of the binomial distribution with the number of trials equal to 1. In other words, the Bernoulli distribution has only two outcomes (success or fail). Bernoulli-distributed data can be modelled with the binomial logit-linear GLGM. Suppose that a soil map legend contains K entries. Let \mathbf{Y}_k denote a random vector for soil type k with Bernoulli-distributed variables (Y_{k1}, \dots, Y_{kn}) for a set of sampling locations $\mathbf{x}_i, i = 1, \dots, n$. Conditional expectations of the Y_{ki} , $E[Y_{ki}|S(\cdot)]$, are binomial probabilities (i.e. probabilities of

occurrence of soil type k). The observations associated to the sampling locations, the y_{ki} , are realizations of the random variables Y_{ki} . The y_{ki} take value 1 if soil type k is observed at sampling location i and 0 otherwise. This means that there are K realizations y_{ki} at each sampling location i : one for each of the Y_k (i.e. one for each soil type distinguished on the soil map). One of these realizations has value 1 and the other $K - 1$ realizations have value 0, hereby assuming that soil types are mutually exclusive and only a single soil type can be observed at a sampling location. Each of the K soil types is modelled separately with the binomial logit-linear GLGM in which the link function is the logit. To simplify notation we denote $E[Y_{ki}|S(\cdot)]$ as π_{ki} in the definition of the logit-linear GLGM:

$$\text{logit}(\pi_{ki}) = \log\left(\frac{\pi_{ki}}{1 - \pi_{ki}}\right) = S_{ki} = \mathbf{d}_{ki}^T \boldsymbol{\beta}_k + U_{ki} . \quad (5.4)$$

Modelling each soil type separately yields K binomial probabilities at each prediction location \mathbf{x}_0 . A pragmatic approach can be then taken to obtain the multinomial probabilities by dividing each binomial probability by the sum of the individual binomial probabilities. For notational convenience subscript k is left out in the sections where parameter estimation and spatial prediction are discussed.

5.2.3 Parameter estimation

So far the GLGM for digital mapping of soil type is defined in the form of the binomial logit-linear GLGM. Next, the parameters of this model must be estimated. Once the parameters have been estimated, these can be used in the model to derive the probability distribution of the soil type at unsampled locations. The model parameters are the coefficients $\boldsymbol{\beta}$ of the trend model, the variance σ^2 of the stochastic component $U(\cdot)$ (Eq. 5.2) and the parameters of the autocorrelation function ρ (range) and the nugget-to-sill ratio. These parameters determine the conditional distribution of $Y(\cdot)$ given $S(\cdot)$.

Estimation of the GLGM parameters for non-Gaussian distributed variables is not straightforward. In case of soil type, the observations y_i of the Bernoulli-distributed variables Y_i are not direct observations of the underlying Gaussian process $S(\cdot)$ and can therefore not be directly used to estimate the parameters of the GLGM. A solution is to use Markov Chain Monte Carlo (MCMC) methods (Diggle et al., 1998; Mosegaard and Sambridge, 2002; Minasny et al., 2011). MCMC is a general-purpose technique for simulating from complex probability distributions. It constructs a Markov chain, which has the desired distribution as its equilibrium distribution. Simulation from the chain after equilibrium has been attained yields a sample from the target distribution. Here MCMC is used to simulate samples (realizations) of S_i conditional on the observations y_i . Next, the simulated S_i are used to estimate the model parameters by maximum likelihood. If this is repeated a sufficiently large

number of times the proper conditional distributions of the signal and its parameters are reconstructed. This method for parameter estimation is referred to as Monte Carlo (MC) maximum likelihood (Christensen, 2004).

For notational convenience, let us assemble all model parameters into one vector of parameters θ . We wish to estimate θ from observations y_i of the Y_i using the maximum likelihood method. Thus, we seek that value of θ for which the likelihood $L(\theta)$ is maximised.

The likelihood of the parameter vector is defined as:

$$L(\theta) = f(\mathbf{y}|\theta), \quad (5.5)$$

where f is probability distribution and \mathbf{y} is the n -element vector containing the y_i . The influence of the model parameters on the observations is passed on through the signal $S(\cdot)$. To unravel the structure of the likelihood the first step is therefore to incorporate the signal into the equation:

$$\begin{aligned} f(\mathbf{y}|\theta) &= \int f(\mathbf{y}, s|\theta) ds \\ &= \int f(\mathbf{y}|s, \theta) f(s|\theta) ds \\ &= \int f(\mathbf{y}|s) f(s|\theta) ds, \end{aligned} \quad (5.6)$$

where in the last step it was used that the \mathbf{y} are independent from θ given s . Both densities in Eq. 5.6 are known in analytical form, but the difficulty is that the n -dimensional integration (i.e. over the distribution of the S_i) is numerically prohibitive. Christensen (2004) proposes an alternative method to evaluate Eq. 5.6 (see also Diggle and Ribeiro Jr. (2007), section 5.5.1). The idea is to introduce a sensibly chosen density function $\tilde{f}(s)$, and use this to rewrite Eq. 5.6 as proportional to an expected value that can be evaluated using Monte Carlo simulation:

$$\begin{aligned} \int f(\mathbf{y}|s) f(s|\theta) ds &= \int \frac{f(\mathbf{y}|s) f(s|\theta)}{\tilde{f}(\mathbf{y}, s)} \tilde{f}(\mathbf{y}, s) ds \\ &= \int \frac{f(\mathbf{y}|s) f(s|\theta)}{f(\mathbf{y}|s) \tilde{f}(s)} \tilde{f}(\mathbf{y}, s) ds \propto \int \frac{f(\mathbf{y}|s) f(s|\theta)}{f(\mathbf{y}|s) \tilde{f}(s)} \tilde{f}(s|\mathbf{y}) ds \\ &= \int \frac{f(s|\theta)}{\tilde{f}(s)} \tilde{f}(s|\mathbf{y}) ds \\ &= \tilde{E}\left[\frac{f(S|\theta)}{\tilde{f}(S)}\right], \end{aligned} \quad (5.7)$$

where the joint density $\tilde{f}(\mathbf{y}, s)$ is defined as $\tilde{f}(\mathbf{y}, s) = f(\mathbf{y}|s) \tilde{f}(s)$. The density $\tilde{f}(s)$ may be any valid density that has positive value for all s (but preferably chosen suit-

ably, see below). The expectation in Eq. 5.7 is taken over the proposed distribution of S given \mathbf{y} ($\tilde{f}(s|\mathbf{y})$), which is proportional to $f(\mathbf{y}|s)\tilde{f}(s)$. The result of Eq. 5.7 can be approximated by repeated MCMC sampling from the proposed distribution of S given \mathbf{y} , evaluating the expression for each simulated value s_j (both probability densities in Eq. 5.7 are easily computed because both are multivariate normal so their analytical expression is known) and averaging:

$$L(\boldsymbol{\theta}) = \tilde{E}\left[\frac{f(S|\boldsymbol{\theta})}{\tilde{f}(S)}\right] \cong \frac{1}{J} \sum_{j=1}^J \frac{f(s_j|\boldsymbol{\theta})}{\tilde{f}(s_j)}, \quad (5.8)$$

where J is the number of MCMC simulations. The maximum likelihood estimates of the model parameters in vector $\boldsymbol{\theta}$ are obtained by identifying the $\boldsymbol{\theta}$ that maximizes Eq. 5.8 for the simulated MCMC samples s_1, \dots, s_J from $[S|\mathbf{y}]$. To find this $\boldsymbol{\theta}$, Eq. 5.8 is repeatedly evaluated for different values of $\boldsymbol{\theta}$, using some form of numerical optimization, until the maximum is found. Although any valid distribution $\tilde{f}(s)$ may be chosen, Christensen (2004) mentions that one may choose $\tilde{f}(s) = f(s|\boldsymbol{\theta}_0)$ for some sensible value $\boldsymbol{\theta}_0$. Trend coefficients obtained by fitting a generalized linear model (GLM) to the data, and variogram parameters obtained by fitting a variogram model to the deviance residuals of the GLM were used as initial parameter values $\boldsymbol{\theta}_0$.

5.2.4 Spatial prediction

The aim of digital soil mapping is to predict the target soil variable at unobserved locations from observations at neighbouring locations, preferably with the help of a set of environmental covariates. In case of the binomial logit-linear model this means predicting the conditional expectation of the Bernoulli variable Y at unsampled locations \mathbf{x}_0 given the neighbouring observations \mathbf{y} : $E[Y(\mathbf{x}_0)|\mathbf{y}]$. Like parameter estimation, spatial prediction with the logit-linear model is not straightforward as the Y_i are not Gaussian-distributed and therefore cannot be interpolated directly by kriging. Eq. 5.4 shows that conditional expectations of Y_i are computed from S_i . This means that if one can predict $S(\cdot)$ at \mathbf{x}_0 , the conditional expectations of Y_i can be obtained by back-transformation of the link function. Because the S_i are unobserved at the sampling locations \mathbf{x}_i , spatial prediction with the logit-linear GLGM takes several steps (Diggle et al., 2003; Diggle and Ribeiro Jr., 2007).

First, realizations of $S(\cdot)$ must be obtained at sampling locations \mathbf{x}_i . Like for parameter estimation, MCMC methods can be used to simulate samples of S_i conditional to the observations y_i from $f(s|\mathbf{y})$ with parameters $\hat{\boldsymbol{\theta}}$ as estimated by the procedure of the previous section. MCMC sampling is repeated J times, yielding J realizations of $S(\cdot)$ at each sampling location: $s_i^{(j)}$, with $j = 1, 2, \dots, J$ and $i = 1, 2, \dots, n$, which shall be denoted by n -element vector $\mathbf{s}^{(j)}$.

Second, $S(\cdot)$ is predicted at \mathbf{x}_0 by kriging a single MCMC sample $\mathbf{s}^{(j)}$:

$$\hat{S}^{(j)}(\mathbf{x}_0) = E[S(\mathbf{x}_0)|\mathbf{s}^{(j)}] = \mathbf{d}(\mathbf{x}_0)^T \boldsymbol{\beta} + \mathbf{c}^T \boldsymbol{\Sigma}^{-1} (\mathbf{s}^{(j)} - \mathbf{D}\boldsymbol{\beta}), \quad (5.9)$$

where $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix of the random effect $U(\cdot)$, \mathbf{c} is the n -element vector containing the covariances between $S(\mathbf{x}_0)$ and \mathbf{s}^j , \mathbf{D} is the $n \times p$ design matrix that associates each of the n observations with a value of each of the p covariates and $\boldsymbol{\beta}$ contains the p covariate coefficients. The prediction variance is:

$$\text{var}[S(\mathbf{x}_0)|\mathbf{s}^{(j)}] = \sigma^2 - \mathbf{c}^T \boldsymbol{\Sigma}^{-1} \mathbf{c}. \quad (5.10)$$

The prediction variance does not depend on the value of \mathbf{s}^j and is therefore similar for the MCMC samples. The kriging step yields J interpolated surfaces of $S(\cdot)$.

Third, each of these surfaces is back-transformed to the original scale by Taylor series expansion of $g^{-1}[\hat{S}^{(j)}(\mathbf{x}_0)]$ (see also Diggle and Ribeiro Jr., 2007, section 6.5) to obtain the expected value of Bernoulli variable Y at \mathbf{x}_0 given the simulated values of the signal at the sampling sites $\mathbf{s}^{(j)}$:

$$\begin{aligned} E[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}] &= g^{-1}[\hat{S}^{(j)}(\mathbf{x}_0)] \\ &\cong \frac{e(x)}{1+e(x)} + 0.5v(x) \frac{e(x)\{1-e(x)\}}{\{1+e(x)\}^3}, \end{aligned} \quad (5.11)$$

where $e(x) = \exp[\hat{S}^{(j)}(\mathbf{x}_0)]$ and $v(x) = \text{var}[S(\mathbf{x}_0)|\mathbf{s}^{(j)}]$. Note that the first term in Eq. 5.11 is the back-transform of the non-spatial binomial model. The back-transform of the prediction variance equals:

$$\text{var}[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}] \cong \left(\frac{e(x)}{1+e(x)}\right)^2 v(x) + 0.5v(x)^2 \left(\frac{e(x)\{1-e(x)\}}{\{1+e(x)\}^3}\right)^2. \quad (5.12)$$

Fourth, the minimum mean squared error predictor given the observed data, $E[Y(\mathbf{x}_0)|\mathbf{y}]$, can be approximated by averaging the J prediction surfaces:

$$E[Y(\mathbf{x}_0)|\mathbf{y}] \approx \frac{1}{J} \sum_{j=1}^J E[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}]. \quad (5.13)$$

The prediction variance $\text{var}[Y(\mathbf{x}_0)|\mathbf{y}]$ is approximated by averaging the J surfaces with prediction variances $\text{var}[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}]$ and then adding the variance of the J conditional expectations $E[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}]$:

$$\text{var}[Y(\mathbf{x}_0)|\mathbf{y}] \approx \frac{1}{J} \sum_{j=1}^J \text{var}[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}] + \frac{1}{J-1} \sum_{j=1}^J (E[Y(\mathbf{x}_0)|\mathbf{s}^{(j)}] - E[Y(\mathbf{x}_0)|\mathbf{y}])^2. \quad (5.14)$$

Finally, the K multinomial probabilities are obtained by dividing each binomial probability by the sum of the individual binomial probabilities:

$$E[Y_{ki}|\mathbf{y}_k]^* = \frac{E[Y_{ki}|\mathbf{y}_k]}{\sum_{k=1}^K E[Y_{ki}|\mathbf{y}_k]}, \quad (5.15)$$

The predicted soil type at \mathbf{x}_0 is the soil type with the largest probability. The confusion index was used as a measure of prediction uncertainty (Burrough et al., 1997):

$$CI(\mathbf{x}_0) = 1 - (E[Y_{1,k}(\mathbf{x}_0)|\mathbf{y}_k] - E[Y_{2,k}(\mathbf{x}_0)|\mathbf{y}_k]), \quad (5.16)$$

in which subscript 1 indicates the largest conditional expectation (probability) and 2 the second largest.

5.3 Case study

The GLGM is applied to map soil type in an area covering 16 750 ha in the southeast corner of the province of Drenthe, the Netherlands (latitude 52.67–52.85 N, longitude 6.85–7.09 E) (Fig 5.1). The study area is part of the man-made agricultural peat landscape that covers an extensive area in the northeast of the country. In 2009 a conventional soil survey started in this area with the aim to update the 1:50 000 national soil map for the areas with peat soils. Recent studies on the status of the peat soils in this landscape revealed major changes in soil conditions since the 1:50 000 survey in the early 1970s and a substantial areal decline of organic soils. Based on recent inventories, shallow peat soils (peat layer less than 40 cm thick) are now found in an estimated 47% of the area mapped as thick peat soils, while mineral soils now cover approximately 55% of the area mapped as shallow peat soils (van Kekem et al., 2005; de Vries et al., 2009).

5.3.1 Landscape and soils

Until the 17th century, the study area was completely covered with vast highmoor bogs. These bogs developed on late-Pleistocene, slightly undulating aeolian sand deposits. Large-scale, systematic reclamation of the area started in the early 17th century and lasted until the mid-20th century. The bogs were drained and the upper part of the peat was excavated for use as a fuel. Deeper peat layers, consisting of sedge and reed peat, were less suitable for fuel and left behind. To prepare the soil for agriculture, peat remains were covered with sand from the vast network of canals that were dug to drain the area and to transport the turf. The sand cover was levelled and slivers of subsoil peat were mixed through it. Large areas of the reclaimed agricultural peat soils were deeply cultivated during the second part of the 20th century,

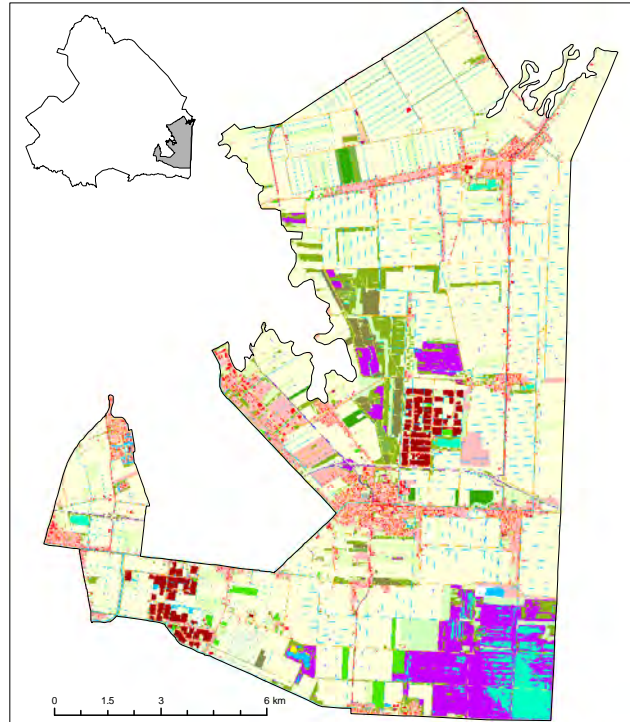


Figure 5.1: Location of the study area in the province of Drenthe with highmoor in purple, forest in green, agriculture in yellow and built-up in pink and red.

which dislocated subsoil peat and sand layers. A few parts of the study area, mainly in the southeast corner and central area, were never reclaimed and cultivated. These areas are now highmoor nature reserves (1 800 ha). Not surprisingly, as a result of the methods for reclaiming and cultivating the former highmoor swamps there is considerable short-range variation in soil conditions such as the topsoil organic matter content and the thickness of the remaining peat layer. Furthermore, spatial heterogeneity at short distances increased over time by decomposition of peat remains that is caused by deep drainage and intensive agriculture.

The existing 1:50 000 national soil map of the study area, which was completed in the 1970s and has become outdated, distinguishes four major soil types: three peat soils and one mineral soil (Fig. 5.2, left panel). Currently, a fifth soil type is found now that the peat layer has been worn away by oxidation in large parts of the study area. The 1:50 000 soil legend further subdivides these major soil types on the basis of for instance presence of gley mottles and soil texture of mineral soils, thickness of the peat layer in deep peat soils and peat type. Here, however, digital soil mapping focusses on the five major soil types. The five major soil types are (reported areas

refer to areas on the existing 1:50 000 map of the study area):

- *Reclaimed shallow peat soils (iW)*. Soils with at most 40 cm of peat within 80 cm from the surface and a man-made topsoil. Topsoil SOM content typically ranges between 5–25% and can be highly variable at short distances as a result of the reclamation process. These soils cover 3 590 ha or 22.4% of the area.
- *Reclaimed deep peat soils (iV)*. Like the reclaimed shallow peat soils but with at least 40 cm of peat within 80 cm from the surface. The peat layer can extend below auger observation depth (~ 120 – 150 cm below surface). These soils cover 9 025 ha (56.2%).
- *Raw peat soils (V)*. These are unreclaimed peat soils found in the few remaining highmoor bogs and swamps. The original peat layer is largely intact and the man-made topsoil is absent. These soils cover 3 175 ha (19.8%).
- *Podzols (Hn)*. Soils with a podzol-B horizon that originally developed in the late-Pleistocene sands before these were covered by peat in the Holocene. These soils cover 266 ha (1.6%). Peat remains can be present in the soil profile, often mixed with mineral material.
- *Hydromorphic earth soils (pZ)*. These sandy soils developed in the late-Pleistocene sands under somewhat wetter conditions than the podzols. They have a dark, humic topsoil that directly overlays the C-horizon. Gley mottles can be present in the subsoil. Although earth soils were not mapped in the study area during the 1:50 000 soil survey, these are now found at locations where the overlying peat layer has completely disappeared. Like the podzols, these soils can have peat remains in the profile.

5.3.2 Data

Soil point data were collected during a soil survey in southeast Drenthe in 2009 and 2010 and contain 4 168 observations on soil type (Fig. 5.2, right panel). At each sampling location an auger boring was made from which the soil profile was observed and classified according to the Dutch soil classification system for detailed soil surveys (ten Cate et al., 1995). The classified soils were recoded to the five major soil types described above.

Nine data layers were available from which a suite of 27 environmental covariates was derived (Table 5.1 provides an overview):

- *Digital elevation model (DEM)*. A 25-m resolution DEM, constructed from LIDAR measurements¹. From the DEM four relative elevation layers were derived by subtracting the local mean elevation within search radii of 250, 500,

¹www.ahn.nl

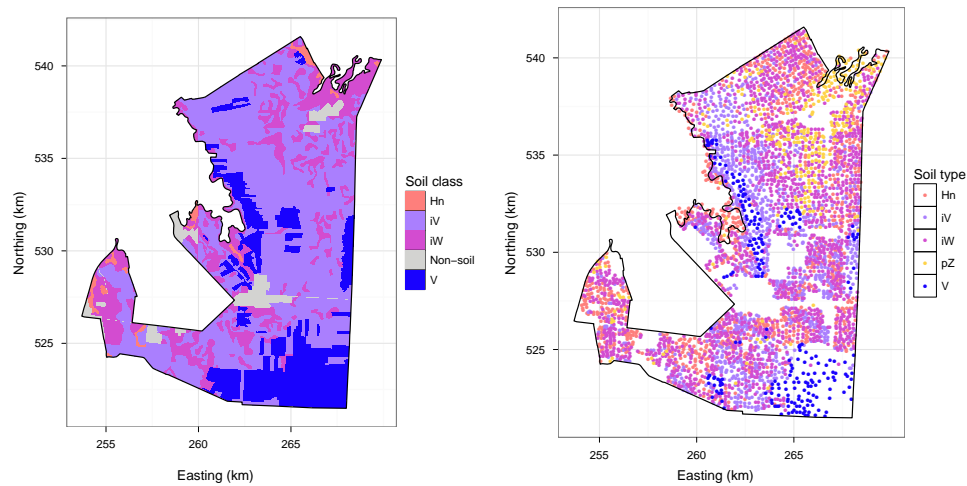


Figure 5.2: The simplified 1:50 000 national soil map of the study area (left) and the point dataset with observed soil types (right).

750, and 1000 m from the actual elevation. These layers capture local relief. One relative elevation layer (based on the 750-m search radius) was reclassified into layers with two, three and four classes to account for a possible non-linear relation between soil type and relative elevation. In addition, a layer with historic elevation was constructed by inverse distance weighted interpolation of a network of elevation measurements from the 1960s (1.2 ha^{-1}). A layer representing elevation change was computed from the actual and historic elevation layers and reclassified into two classes. Elevation change is informative because peat excavation and decomposition lower the surface.

- *Land cover maps.* This set contains five layers depicting land cover in 1900 (50-m resolution), 1940 (25 m), 1960 (25 m), 1980 (25 m) and 2003 (25 m) (Clement and Kooistra, 2003; Knol et al., 2003, 2004; Hazeu, 2005). The layer from 1900 distinguishes ten classes and was reclassified into a layer with two classes and into a layer with three classes. The layer from 2003 distinguishes 23 classes and was reclassified into four layers, each with a different combination of classes. The layers from 1900, 1940, 1960, 1980 and 2003 were combined into a map that represents reclamation period, i.e. the period when reclamation for agriculture took place. Four layers were derived from the reclamation period map, each with a different combination of reclamation periods.
- *National Soil Map* (Steur and Heijink, 1991). Polygon map at scale 1:50 000 as described in section 5.3.1. In addition to soil type, five soil variables were derived from map unit classifications: one variable representing peat type, three

variables representing peat thickness, and one variable indicating the status (degraded or not degraded) of peat soils iV and V based on the results of a quick-scan of these soils (van Kekem et al., 2005).

- *Paleogeography map* (Spek, 2004). Polygon map at scale 1:50 000 representing a reconstruction of the geography of Drenthe around 1000 AD, distinguishing 12 geographical units. The map was converted to a layer of 25-m resolution, from which two new layers were extracted: one with the former extent of fen peat, one with the former extent of highmoor.

5.3.3 Digital soil type mapping with the GLGM

Digital mapping of soil type k , $k = 1, \dots, 5$ (Fig. 5.2) was done as follows. First, a Bernoulli variable was created for each sampling location that takes 1 if soil type k is observed at that location and 0 otherwise. Second, a GLM with logit-link function was fitted and the deviance residuals at the sampling locations were calculated. Covariates for the trend component of the GLM were selected from the suite of environmental covariates using a manual step-wise approach (starting with the strongest predictor based on univariate analysis) with the Akaike Information Criterion (AIC) (Webster and McBratney, 1989) as selection criterion. The coefficients of the trend model were used as initial estimates of the parameters of the trend component of the GLGM for soil type k (β_{k0}). Third, an exponential variogram model was fitted to the deviance residuals using maximum likelihood. The parameters obtained from the variogram were used as initial estimates of the variance parameters of the stochastic component of the GLGM for soil type k (the sill σ_{k0}^2 , the range ϕ_{k0} and the nugget τ_{k0}^2). The initial model parameters were combined in parameter vector $\hat{\theta}_{k0}$. Fourth, 150 000 MCMC samples of S_k were generated at each sampling location \mathbf{x}_i given the y_{ki} and $\hat{\theta}_{k0}$. Of these, the first 25 000 were discarded as burn-in and of the remaining 125 000 every 50th was sampled, yielding 2 500 simulations of S_k . These simulated values were used to obtain $\hat{\theta}_k$, the maximum likelihood estimate of θ_k . Finally, for spatial prediction again 125 000 MCMC samples of S_k were generated at each sampling location, but this time given the y_{ki} and $\hat{\theta}_k$. Again, the first 25 000 were discarded as burn-in and of the remaining 125 000 every 50th was sampled. Kriging was performed at the nodes of a 50-m grid for each of the 2 500 simulated datasets s_k . The predictions and prediction variances were back-transformed to the original scale and then averaged over the 2 500 surfaces.

Model selection, parameter estimation and spatial prediction were repeated for each of the five soil types, resulting in five predicted binomial probabilities at each prediction location. From these probabilities the multinomial probabilities were computed by Eq. 5.15. Subsequently, the soil type with the largest predicted probability at each raster node was used to create the soil type map. In addition the confusion index was mapped. Two preliminary measures of accuracy were computed from the

Table 5.1: List of environmental covariates.

Description	Levels	Code
<i>Digital elevation model*</i>		
Relative elev, 250 m	-	REL250
Relative elev, 500 m	-	REL500
Relative elev, 750 m	-	REL750
Relative elev, 1000 m	-	REL1000
Relative elev, 2 cl	1 = <0, 2 = >0	REL2
Relative elev, 3 cl	1 = <-25, 2 = -25-25, 3 = >25	REL3
Relative elev, 4 cl	1 = <-25, 2 = -25-0, 3 = 0-25, 3=>25	REL4
<i>DEM + point measurements[†]</i>		
Elevation change, 2 cl	1 = <-25, 2 = >-25	EC2
Elevation change, 3 cl	1 = <-50, 2 = -50-25, 3 = >25	EC3
<i>Land cover maps</i>		
Land cover 1900, 2 cl	1 = agriculture, 2 = natural	LC1900.2
Land cover 1900, 3 cl	1 = grassland, 2 = cropland, 3 = natural	LC1900.3
Current land cover, 2 cl	1 = agriculture, 2 = natural	COV2
Current land cover, 2 cl	1 = highmoor, 2 = other	COV2a
Current land cover, 3 cl	1 = agriculture, 2 = forest, 3 = rangeland/highmoor	COV3
Current land cover, 3 cl	1 = agriculture, 2 = forest/rangeland, 3 = highmoor	COV3a
Reclamation period, 2 cl	1 = <1940, 2 = >1940	RECLAM2
Reclamation period, 3 cl	1 = <1940, 2 = 1940-2003, 3 = >2003	RECLAM3
Reclamation period, 3 cl	1 = <1940, 2 = 1940-1980, 3 = >1980	RECLAM3a
Reclamation period, 3 cl	1 = <1900, 2 = 1900-1940, 3 = >1940	RECLAM3b
<i>Soil map, scale 1:50 000</i>		
Soil type	1 = Hn, 2 = iV, 3 = iW, 4 = V,	SOIL
Peat type	1 = Oligotrophic, 0 = Mesotrophic	PEATTYP
Peat thickness [‡] , 2cl	1 = 0-120, 2 = >120	PEATTHK2
Peat thickness [‡] , 2cl	1 = 0-40, 2 = >40	PEATTHK2a
Peat thickness [‡] , 3cl	1 = 0-40, 2 = 40-120, 3 = >120	PEATTHK3
Peat status	1 = deformed peat soil, 0 = unchanged,	PEATSTAT
<i>Paleogeography map</i>		
Former fenpeat coverage	1 = Yes, 0 = No	FENPEAT
Former highmoor coverage	1 = Yes, 0 = No	HIGHMOOR

* Relative elevation in cm. Negative numbers indicate relatively low positions.

[†] Elevation change in cm. Negative numbers indicate subsidence.

[‡] Thickness in cm.

predicted soil map. The first is the calibration purity, which is the proportion of observation locations at which the soil map predicts the correct soil type. The second is the theoretical purity. This is the expected proportion of a map that will be correctly classified given the probability model and is equal to the spatial mean of maximum probability over the region (Brus et al., 2008).

5.3.4 Digital soil type mapping with non-spatial models

In addition to digital mapping with a spatial model, soil type was also predicted with two non-spatial models. The first model is the multinomial logit model (section 2.2.3). The MLM for a categorical variable with K outcome categories has $K - 1$ logit functions. One outcome category is chosen as the reference category. The model coefficients of the $K - 1$ logit functions are estimated by maximum likelihood with respect to the reference category (Hosmer and Lemeshow, 2000). In an MLM the logit function of the probability of occurrence of soil type k at location x_i is assumed to be a linear combination of the covariates:

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{\pi_k^*}\right) = \mathbf{D}\boldsymbol{\beta}_k, \quad (5.17)$$

where π_k is the probability of outcome category k , π_k^* is the probability of the reference category, \mathbf{D} and $\boldsymbol{\beta}_k$ are as defined in Eq. 5.9. The multinomial probability of outcome category k then equals:

$$\pi_k = \frac{\exp(\mathbf{D}\boldsymbol{\beta}_k)}{1 + \sum_k^{K-1} \exp(\mathbf{D}\boldsymbol{\beta}_k)}. \quad (5.18)$$

Covariates were selected from Table 5.1. The probabilities of the five soil types were predicted at the nodes of 50-m grid. The soil map with the largest probability was used to construct the soil map.

The second non-spatial model is the logit-linear GLM for Bernoulli-distributed data, which shall be hereafter referred to as BLM (Bernoulli Linear Model). The selected GLGM and MLM can have different sets of covariates, which confounds the effect of the spatial model component on the predicted probabilities. To assess the effect of the spatial component on the predicted probabilities, five soil type-specific BLMs were fitted in addition to the MLM. The trend models of the BLMs contained the same covariates as the associated GLGMs. The five BLMs were used to predict the probabilities of the five soil types at the nodes of a 50-m grid. These probabilities were scaled to multinomial probabilities by applying Eq. 5.15. For both non-spatial models a map of the confusion index was constructed, and the theoretical and calibration purities were calculated.

5.3.5 Validation

Sampling strategy

The soil maps were validated with independent probability sample data. The validation data were collected with a stratified simple random sampling design (Brus et al., 2011). The map units of the soil map constructed from predictions with the GLGM

formed the five strata. At each sampling location the soil profile was described and classified from an auger bore observation. A total of 125 sampling locations were selected, which were allocated to the strata roughly proportional to their surface areas. Sampling locations where permission was denied or proved otherwise impossible to sample were replaced with randomly selected locations from a reserve list. Field-work took place in March 2011.

Statistical inference

Three quality measures for the categorical soil type maps are considered: overall purity, map unit purity (user's accuracy) and class representation (producer's accuracy) (Brus et al., 2011). Overall purity is defined as the proportion of the mapped area in which the predicted soil type, which is the soil type as depicted on the map, equals the true soil type. In other words, it is the areal proportion correctly classified. To estimate the overall purity an indicator variable is created, which takes the value 1 if the observed soil type equals the predicted soil type and 0 otherwise. For each stratum the average of this indicator is computed. The overall purity is then estimated as the weighted average of the stratum purities, with weights equal to the relative sizes of the strata.

The map unit purity is the proportion of the map unit correctly classified. If the map units were used as the sampling strata, then the map unit purities area estimated by the strata means (in case of the soil map based on predictions with the spatial model). If the map units do not equal the strata, then the map unit purities must be estimated with the ratio estimator (in case of the soil map based on predictions with the non-spatial models). This estimator is used for purity estimates of so-called *domains* (sub-areas of interest) (Brus et al., 2011). Class representation of soil type k is the proportion of the area where the actual soil type k occurs that is also mapped as type k . Class representations are also estimated by the ratio estimator.

The accuracies of the soil maps can be compared by introducing variable q_{hi} defined as $y_{hi}^s - y_{hi}^n$, where y_{hi} is an indicator that takes value 1 if the predicted soil type at location i in stratum h equals the observed soil type and 0 otherwise, superscripts s and n indicate spatial model and *non*-spatial model. Variable q can have values -1, 0, and 1. The mean purity difference (MPD) of two soil maps, \bar{q} , is estimated in similar fashion as the overall purity. To test whether the estimated MPD differs significantly from 0 it was assumed that the estimated MPD follows a normal distribution.

5.3.6 Software

All statistical modelling was carried out using the software R, Version 2.13.0 on a dual-core laptop computer with 2.53 GHz processors and 3Gb RAM on Windows

32-bit platform. Parameter inference and spatial prediction with the GLGM were done with the package `geoRglm` (Christensen and Ribeiro Jr., 2002). This package implements the Langevin-Hastings algorithm for MCMC simulation and a general-purpose (quasi-Newton) optimization algorithm to identify the maximum likelihood estimates of the model parameters. The MLM was fitted using the `mlogit` package. Figures were created with package `ggplot2` (Wickham, 2009).

5.4 Results

5.4.1 Prediction models

Table 5.2 shows the selected trend models of the GLGMs and MLM, and the estimates of the variance parameters of the GLGMs. The covariates are listed in order of the strength of correlation between covariate and soil type based on univariate analysis. Not surprisingly, the 1:50 000 soil map and its derivatives (peat type and peat layer thickness) are strong predictors of current soil type. The land cover map with three classes (agriculture, forest/rangeland, highmoor) proved also to be an effective predictor. This can be explained by the fact that mineral soils *Hn* and *pZ* and reclaimed shallow peat soils *iW* almost exclusively occur under agriculture, whereas raw peat soils *V* almost exclusively occur under forest vegetation and in highmoor areas. Reclaimed deep peat soils *iV* mainly occur under agriculture but are also found under rangeland, where they are the dominant soil type. The rangeland area was used for agriculture until the mid 1990s. At that time agriculture was abandoned and natural vegetation restored, which slowed peat decomposition. Covariates related to elevation and landscape type were also influential but less important.

Roughly 15 to 40% of the residual variance of the signal is spatially structured at medium distances (practical range of the exponential variogram models varies approximately from 1000–3500 m). Spatial dependence in the residual is weakest for the reclaimed shallow peat soils, with only 17% of the residual variance spatially structured. These soils are in a pedologically intermediate position between the mineral soils and reclaimed deep peat soils, which means that both mineral soils as well as deep peat soils frequently occur in areas where shallow peat soils dominates. This short-distanced spatial heterogeneity results in a relatively large nugget variance. Areas dominated by mineral soils and deep peat soils are expected to be relatively more homogeneous than shallow peat soil dominated areas, which results in a larger part of the variance that is spatially structured.

Table 5.2: Selected trend models and estimates of the variance parameters.

Soil type	<i>n</i>	Covariates	Variance	<i>a</i> [*] (m)	Spatial dep. [†]
<i>Generalized linear geostatistical models</i>					
Hn	935	PEATTHK3+EC2+PEATTYP+COV3a+REL3	0.99	726	0.34
iV	1 023	PEATTHK2+SOIL+COV3a+EC2+REL1000+PEATSTAT	0.95	846	0.34
iW	1 670	COV3a+PEATTHK2+RECLAM3+SOIL+REL4+HIGHMOOR	1.24	348	0.17
pZ	303	PEATTYP+FENPEAT+PEATTHK3	0.47	1 112	0.39
V	237	COV3a+PEATTHK2+PEATTYP	0.06	387	0.24
<i>Multinomial logit model</i>					
-	4 168	PEATTHK3+COV2+RECLAM3+PEATTYP+REL3+PEATSTAT+EC2			

^{*} Distance parameter of the exponential model. The practical range is 3*a*.

[†] The spatial dependence is the ratio between spatially structured variance and total variance [partial sill/(nugget+partial sill)].

5.4.2 Digital soil type mapping

Fig. 5.3 shows the soil maps as predicted by the spatial and non-spatial models. The general patterns of soil distribution are quite similar: the GLGM and MLM predict similar soil types at 87% of the prediction sites, for the GLGM and BLM this is 89%. Raw peat soils are predicted in the highmoor nature reserves. All models predict reclaimed deep peat soils *iV* in former agricultural areas under natural vegetation (west-central part) and in former highmoor areas that were reclaimed for agriculture after the 1:50 000 soil survey was completed (south-central and east-central areas). The most obvious effect of including a spatial component in the model is reflected by the greater predicted area of mineral soils (*Hn*, *pZ*) in the GLGM map (+630 ha), at the expense of the shallow peat soils *iW*. Comparing the GLGM and MLM maps shows that the GLGM predicted greater areas of podzols *Hn* and deep peat soils *iV*, while the MLM predicted greater areas of thin peat soils *iW* and hydromorphic earth soils *pZ*. Spatial patterns of the CI are very similar (Fig. 5.3). The median CIs for predictions with the GLGM was 0.78, for the BLM and MLM this was 0.73. Including a spatial component in the prediction model did not reduce the uncertainty about the prevailing soil type at the prediction sites. For each model the median CI is smallest for map unit *V*, followed by *iV*, *iW*, *pZ* and *Hn*. The calibration purities are 61.9% for the GLGM; 56.5% for the MLM, and 56.7% for the BLM. The theoretical purities are 57.1% for the GLGM; and 58.3% for the MLM, and 58.2 % for the BLM.

Fig. 5.4 depicts the difference between the probabilities predicted by the GLGM and BLM for each soil type. The maps show, with an exception for soil type *V*, that the spatial component had a distinct effect on the predicted probabilities. The areas with positive differences for a specific soil type (probability with GLGM larger than with the BLM) correlate very well with the areas with frequent observations of that soil type.

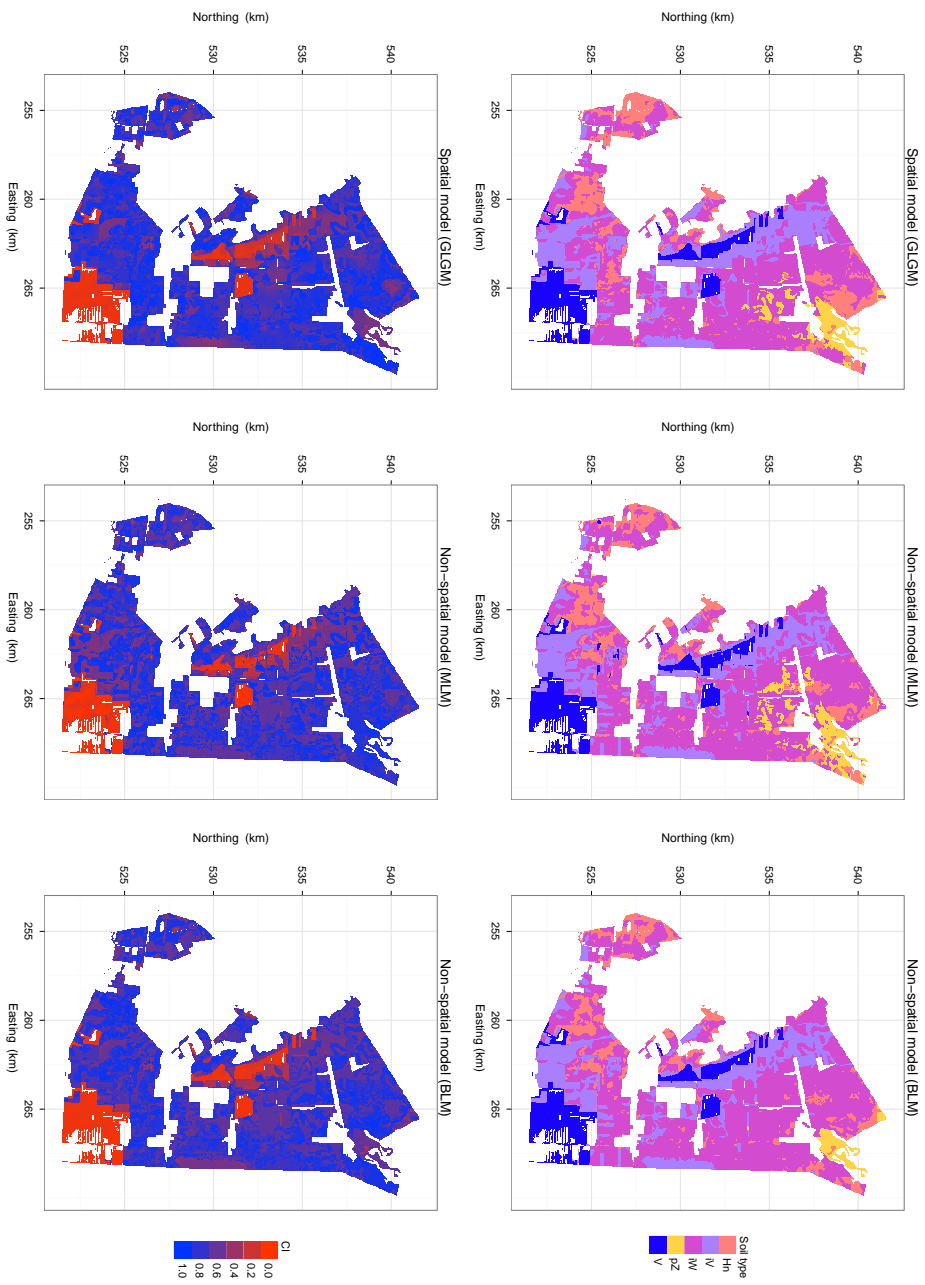


Figure 5.3: Soil type maps as predicted with the spatial and non-spatial models (top) and associated confusion index maps (bottom).

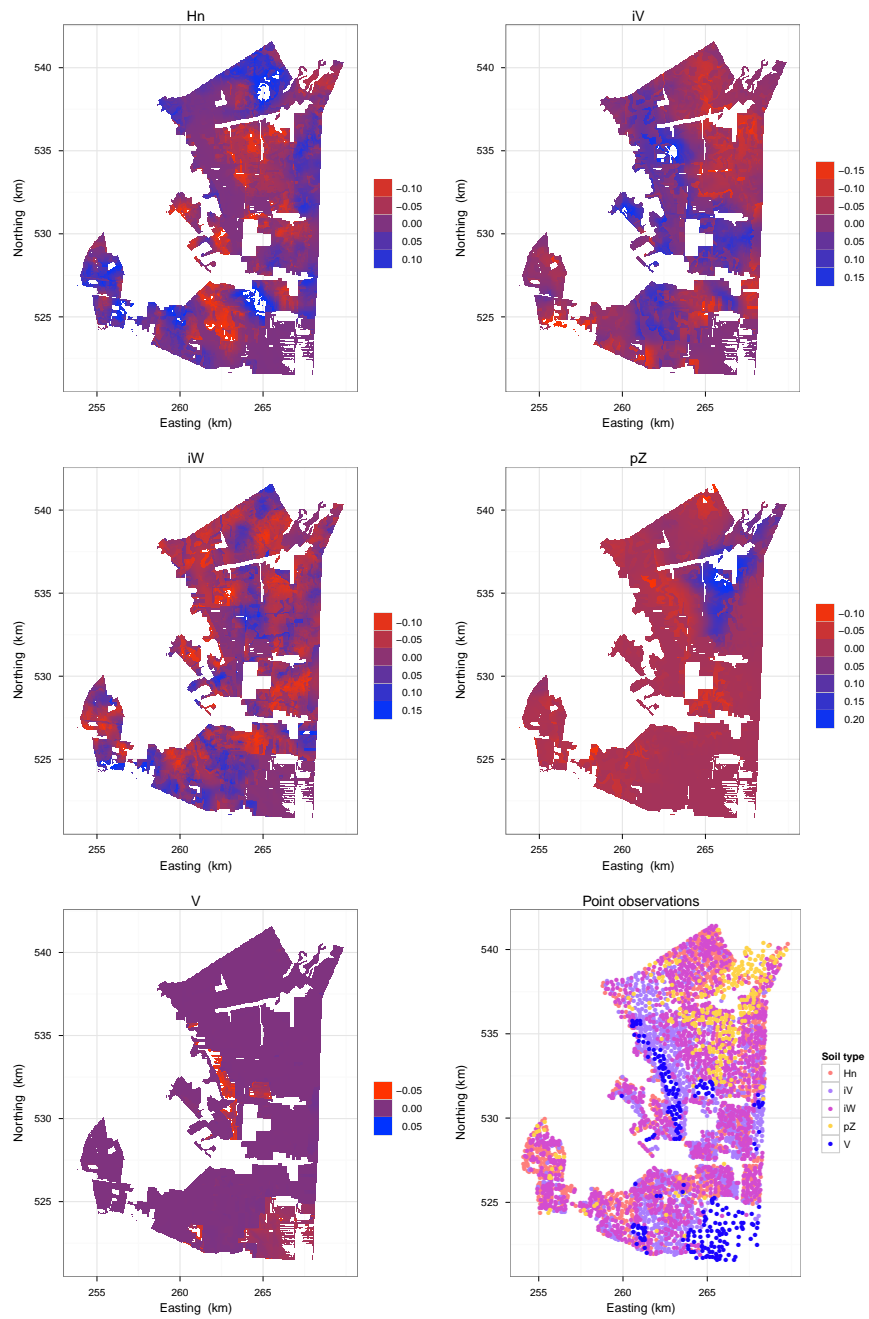


Figure 5.4: Difference in predicted probabilities by the spatial and non-spatial models. Positive numbers indicate a larger probability for the spatial model.

5.4.3 Validation

The results of validation with independent data are listed in Table 5.3. Overall purity of the soil map created with the spatial model is 54.9%, which is 3.5% smaller than the MLM map ($p = 0.105$) and 1.7% smaller than the BLM map ($p = 0.260$). The theoretical purities are somewhat larger than the overall purities, but all fall within one standard error of the estimated means and thus give a fairly good estimate of the actual map purities.

Table 5.4 shows the sample error matrices from which it is evident that confusion between soil types is largest for the reclaimed shallow peat soils *iW*. This is not surprising as these soils are in pedologically intermediate position between mineral soils and reclaimed deep peat soils. The superior overall purity of the MLM map compared to the GLGM map is mainly caused by the better representation of *iW* (for MLM 35 out of 51 observations correctly represented, with GLGM 31 out of 51 locations), which are less confused with *Hn*. Map unit purities and class representations vary greatly between soil types and between models. Raw peat soils *V* occur in very specific areas that are easily identified from land cover maps (used as covariate). The occurrence of these soils is very well predicted by all three models. Map unit *iV* had the lowest purity for all models. However, these soils are mainly confused with *iW*, suggesting that at locations where the models predict soil type *iV*, it is highly probable that a peat soil is found. Only the thickness of the peat layer is uncertain.

As a final analysis the three soil maps were generalized into three groups, (mineral soils (*Hn*, *pZ*), reclaimed peat soils (*iV*, *iW*) and raw peat soils (*V*), and then validated. The estimated overall purities are 75.1% for the GLGM map, 79.3% for the MLM

Table 5.3: Estimates of overall purity, map unit purity and class representation for the three prediction models. The standard errors of the accuracy estimates are between brackets and n is the number of validation observations.

	GLGM		MLM		BLM	
	n	purity	n	purity	n	purity
<i>Overall purity</i>	125	54.9 (4.4)	125	58.4 (4.4)	125	56.6 (4.4)
<i>Map unit purity</i>						
<i>Hn</i>	22	48.3 (10.7)	14	69.3 (11.8)	14	57.1 (13.6)
<i>iV</i>	25	40.0 (10.0)	23	43.2 (10.5)	24	41.7 (10.3)
<i>iW</i>	56	55.1 (6.7)	69	55.7 (6.3)	68	54.1 (6.1)
<i>pZ</i>	10	60.7 (17.6)	7	41.5 (16.5)	7	85.7 (13.9)
<i>V</i>	12	91.7 (8.3)	12	92.3 (7.7)	12	91.7 (8.3)
<i>Class representation</i>						
<i>Hn</i>	21	49.1 (9.8)	21	49.1 (9.8)	21	34.8 (9.2)
<i>iV</i>	24	39.6 (8.4)	24	39.1 (8.8)	24	39.6 (8.4)
<i>iW</i>	51	63.6 (5.7)	51	71.8 (5.0)	51	73.4 (4.7)
<i>pZ</i>	17	22.9 (6.6)	17	16.3 (6.3)	17	19.6 (6.5)
<i>V</i>	12	92.1 (7.3)	12	100 (0.0)	12	92.1 (7.3)

map, and 77.3% for the BLM map. The improvement in overall purity for all three models is primarily caused by pooling soil types *iV* and *iW*, illustrating the difficulty in predicting the correct peat thickness in the man-made cultivated peatlands.

Table 5.4: Sample error matrices showing the counts of predicted versus observed soil type in the validation sample. Column totals are identical for each model.

Mapped soil type	Observed soil type					Total
	Hn	iV	iW	pZ	V	
<i>GLGM</i>						
Hn	11	2	7	2	0	22
iV	0	10	12	2	1	25
iW	9	10	31	6	0	56
pZ	1	1	1	7	0	10
V	0	1	0	0	11	12
<i>MLM</i>						
Hn	11	0	3	2	0	16
iV	0	10	11	2	0	23
iW	9	12	35	8	0	64
pZ	1	1	2	5	0	9
V	0	1	0	0	12	13
<i>BLM</i>						
Hn	8	1	3	2	0	14
iV	0	10	11	2	1	24
iW	13	12	36	7	0	68
pZ	0	0	1	6	0	7
V	0	1	0	0	11	12
Total	21	24	51	17	12	

5.5 Discussion and conclusions

In this chapter the generalized linear geostatistical model for digital soil type mapping was presented and its use illustrated with a real-world case study for updating the soil map for cultivated peatlands. Predictive mapping with the GLGM is computationally intensive since modelling involves MCMC simulation and MC maximum likelihood estimation of the model parameters, but feasible on a standard laptop computer. For our case study with over 4 100 point observations and 55 000 prediction locations, total computation time for parameter estimation and prediction with a single (soil type-specific) GLGM was about 45 hours.

Validation results showed that the non-spatial MLM generated a more accurate soil map than the GLGM. Furthermore, a comparison of the GLGM map with the BLM map showed that inclusion of a spatial component in the model had considerable effect on the predicted probabilities and the soil type distribution, although it did not improve map accuracy. The fact that predictions with the spatial model were

not more accurate than those with the non-spatial models was not completely unexpected given the substantial heterogeneity in soil conditions at short distances in the Dutch cultivated peatlands, where very different soils can be found only few metres apart (section 5.3.1). The three prediction models were able to predict the major spatial patterns of soil variation but failed to capture the details. Nevertheless, this study provides valuable insight into the performance of spatial and non-spatial method for soil type mapping in the Dutch peatlands and can guide decisions on strategies for nationwide updating of the peat areas of the 1:50 000 soil map in the near future. One should be careful with generalizing the findings in this chapter to areas with more natural landscapes where soil-landscape relationships are less disturbed by human activities and soil cover is more homogenous. In such situation, spatial prediction of soil type is more likely to benefit from using a spatial model. It would be worthwhile to compare the performance of a spatial with a non-spatial model, both calibrated with real-world data, for soil type mapping in a natural landscape.

The large difference between calibration purity and actual purity is striking. This emphasizes the importance of validating soil maps with independent data, collected by probability sampling (Brus et al., 2011). The difference might be attributed simply to chance: the calibration purity falls within two standard errors of the overall purity. One other reason for the large difference we can think of is over-estimation of the spatial continuity for soils Hn , iV and pZ (which exhibit the strongest spatial autocorrelation, Table 5.2). The difference in calibration purity between the GLGM and BLM is attributed to better representation of these soil types at the observation locations by the GLGM map. Preferential sampling for data collection might have resulted in over-estimation of the strength of spatial dependence for these three soils. In the agricultural part of the study area soils Hn , iV and pZ are typically found as large and small inclusions in the dominant iW unit. Soil surveyors focus on delineating these inclusions on 1:50 000 scale, which might lead to biased selection of sampling sites towards these soils. Occasionally sampling sites were discarded because the observed soil did not fit the soil surveyor's general picture of soil variation in the field under investigation. Another reason for over-estimation of the strength of spatial dependence might just be a poor choice of the initial model parameters θ_0 . Christensen (2004) warns that the choice of θ_0 is not trivial. We took a pragmatic approach and used variance parameters estimated from the deviance residuals as initial estimates. Model parameters θ_0 are used for MCMC simulation of the underlying but unobservable signal $S(\cdot)$. The simulated values are in turn used to estimate θ by MC maximum likelihood. It is therefore expected that θ_0 will have some effect on $\hat{\theta}$ but we have no insight in the robustness of MC maximum likelihood estimation against misspecification of the model. Instead of using initial estimates obtained from the data, Diggle et al. (2002) and Ben-Ahmed et al. (2010) used a Bayesian approach.

The multinomial probability distribution was modelled via soil type-specific Bernoulli distributions. This pragmatic approach was chosen because for binomial (Bernoulli)

data the theory of GLGM has been worked out and software is available, whereas for multinomial data this is not the case. Hartzel et al. (2001) and Hedeker (2003) have shown that the multinomial logit model can be extended to incorporate non-spatial Gaussian random effects. This MLM with random effects might be further extended by incorporating additional parameters of a covariance function in the random model component and Bayesian specification of the model.

The GLGM assumes that the Y_i are mutually independent conditional on the underlying Gaussian process $S(\cdot)$. Spatial dependence in Y is only modelled through S , which implies that whereas the soil type probabilities have spatial structure, the actual realizations of Y may be quite noisy, and possibly more noisy than is realistic. For example, suppose that at two adjacent grid cells i and j the probabilities of observing soil type A are $P[(y_i) = 1] = \pi(y_i) = 0.8$ and $P[(y_j) = 1] = \pi(y_j) = 0.8$. Because of independence, the joint probability of observing soil type A at both locations is $\pi(y_{i,j}) = 0.8^2 = 0.64$. A cluster of nine grid cells with equal probabilities has a joint probability of only 0.13. However, since the same soil type often occurs in larger patches these joint probabilities appear to be too small. The GLGM cannot account for this effect because it assumes independent observations given the spatial process $S(\cdot)$. An alternative to the GLGM is the autologistic model (e.g. Augustin, 1996; Hoeting et al., 2000). In the autologistic model the data are modelled as a locally dependent Markov random field. Here the probability at a prediction location is modelled as a linear combination of observations in the neighbourhood and a set of covariates. Thus the autologistic model can represent much more contiguous, and perhaps more realistic, realizations of soil spatial variation. Autologistic models, however, can only be used for binomial-distributed data and suffer some severe drawbacks (Dormann, 2007).

Concluding, the GLGM is based on a valid theoretical framework and can be fitted with available geostatistical software, which makes it a potentially attractive model for soil type mapping. The GLGM was applied to update a soil map of the Dutch cultivated peatlands but gave no better predictions in terms of map purity than two non-spatial models. Therefore the usefulness of the GLGM for future soil map updating in the Dutch peatlands seems limited. Issues that require further investigation are the extension of GLGM to multinomial data and the validity of the assumption of conditionally independent observations in the context of soil type mapping.



Chapter 6

Efficiency comparison of conventional and digital soil mapping for updating soil maps of a cultivated peatland

This chapter compares the efficiency of digital soil mapping (DSM) methods for updating soil type and property maps with that of conventional soil mapping (CSM) methods. For digital soil type mapping the GLGM (presented in Chapter 5) was used. For soil property mapping (SOM content and peat thickness) two methods are considered for both DSM and CSM. For DSM these are the method proposed in Chapter 3 and the conventional geostatistical method. For CSM these are the representative profile descriptions (RPD) and map-unit-means (MUM) methods. In addition, the effect of mapping effort (expressed in a monetary unit per ha) on accuracy is assessed for digital soil type and property maps. For CSM the MUM method gave better results than the RPD. For DSM both methods gave similar results in terms of accuracy. Validation results further illustrated that DSM methods produced soil type and property maps that were of similar accuracy as those produced by CSM methods. Furthermore, DSM maps were produced much more efficiently than the CSM maps: costs per ha were a factor three to four smaller without compromising accuracy. This shows that for future updating of soil information, DSM can be an attractive alternative to CSM.

Based on: B. Kempen, D.J. Brus, J.J. Stoorvogel, G.B.M. Heuvelink, F. de Vries
Submitted to Soil Science of America Journal

6.1 Introduction

Digital soil mapping (DSM; McBratney et al., 2003) has become an established alternative to conventional soil mapping (CSM; Soil Survey Division Staff, 1993) during the first decade of the 21st century. The explosive increase in the use of DSM techniques has been made possible because of rapid developments in computing and information technology, increased availability of cheap and easily accessible digital environmental data and fast methodological advances in the field of pedometrics which has been supported by growing availability of (open-source) computer software for statistical computing such as R (R Development Core Team, 2008). DSM is nowadays widely applied to produce soil maps and populate soil databases across the world (e.g. Lagacherie and McBratney, 2007; Hartemink and McBratney, 2008; Sanchez et al., 2009; Boettinger et al., 2010).

In CSM soil maps are generally created using free survey. In free survey the soil surveyor employs a conceptual (mental) soil-landscape model to select observation locations at which the most useful information is likely to be obtained (Bregt, 1992b). Landscape features as seen in the field or aerial photographs, available environmental data such as a DEM and past experiences of the surveyor in similar landscapes are taken into account (Hewitt, 1993). CSM results in a soil type map and a set of soil profile descriptions. Each map unit is characterized by one or more representative soil profile of the soil types that comprise the map unit. These profiles are used for the interpretation of the soil map (Bregt and Beemster, 1989). Conventional soil maps are general-purpose soil maps: they provide information on the spatial distribution (in three dimensions) of a wide range of soil properties.

In DSM (geo)statistical models are used to relate field observations on soil type or property to readily available, spatially exhaustive environmental data. These data should represent important soil forming factors that explain the spatial variation of the target soil attribute. For example, derivatives from digital elevation models (DEM) can represent the influence of topography on soil formation while satellite imagery can represent the effects of vegetation and climate. Once a (geo)statistical model is fitted to the data, i.e. the soil-landscape relationships are quantified, then this model can be used to spatially predict the target soil attribute at unobserved locations given the observed environmental data at these locations. Contrary to conventional soil maps, maps obtained by DSM methods are typically specific-purpose soil maps: a map is created of a specific soil property at a specific depth interval. Multivariate methods can be used for DSM but these become prohibitive when a large number of properties is considered. Furthermore, complex soil forming processes might be difficult to quantify and to represent by environmental explanatory variables, while these can be more easily taken into account in conventional soil maps. Also, DSM methods for soil type are limited by the number of soil types that can be handled (Brus et al., 2008).

Conceptually DSM and CSM are very similar: both approaches use a soil-landscape model to predict soil at unobserved locations. The main difference is that in CSM the soil-landscape model is a qualitative (non-documented) model based on soil surveyor expert-knowledge, while in DSM the soil-landscape model is a quantitative, (geo)statistical model. Because of its qualitative nature, CSM is often considered as much an art as science (Hewitt, 1993). Main criticisms of CSM include irreproducibility, soil bodies being represented as discrete, homogeneous entities, and the lack of quantified measures of uncertainty (Hewitt, 1993; Goovaerts and Journel, 1995). DSM does not suffer from these shortcomings. Prediction models can be stored and run again, different models of spatial variation (discrete, mixed, continuous; Heuvelink, 1996) can be chosen and proper use of (geo)statistical methods result in predictions with quantified uncertainty. In addition, DSM is assumed to be more efficient than CSM. Here more efficient means that fewer soil observations (i.e. less field work) are required to produce a map with similar accuracy. However, this assumption has to our knowledge never been tested in the literature. Perhaps DSM requires as many or more observations than CSM to reach the same accuracy. For instance, the results in Chapter 4 showed that depth profiles modelled with DSM methods for three-dimensional mapping of the soil organic matter (SOM) content performed only marginally better than depth profiles obtained from representative profile descriptions.

Issues on accuracy of DSM products and DSM efficiency have also become increasingly important to soil survey in the Netherlands. Approximately 365 000 ha mapped as peat soils on the 1:50 000 national soil map requires updating. Roughly one third of these peatlands are cultivated. Intensive land use and deep drainage in combination with shallow peat layers has resulted in major changes in soil conditions since the 1:50 000 survey was conducted in the 1970s. A 20 000 ha area of cultivated peat soils in the province of Drenthe has been re-mapped with conventional survey methods in 2009 and 2010. However, updating the entire peat area in the Netherlands by CSM is not a viable option given the high costs of such an undertaking. Map updating in the near future will have to rely more and more on alternative, cheaper methods such as DSM. The results presented in Chapter 2 on updating the existing 1:50 000 with DSM methods were promising, but did not meet the 70% purity standard generally assumed for the 1:50 000 soil map.

In order to move DSM activities in the Netherlands forward from experimental to operational phase, insight in the current performance of DSM methodology—for updating soil type as well as soil property maps—in terms of map accuracy and mapping efficiency with respect to CSM is essential. The objective of this chapter therefore is to evaluate and compare the efficiency, in terms of mapping effort and accuracy, of digital and conventional soil mapping methods for updating soil type and property maps for a cultivated peatland in the province of Drenthe in the Netherlands.

6.2 Materials and methods

6.2.1 Study area

The study area covers 16 750 ha in the southeast corner of the province of Drenthe, the Netherlands (latitude 52.67–52.85 N, longitude 6.85–7.09 E) (Fig. 5.1). The study area is part of the man-made agricultural peat landscape that covers an extensive area in the northeast of the country. Non-soil areas cover 2 950 ha.

Landscape

Until the 17th century, the study area was completely covered with vast highmoor bogs. These bogs developed on late-Pleistocene, slightly undulating aeolian sand deposits. Large-scale, systematic reclamation of the area started in the early 17th century and lasted until the mid-20th century. The bogs were drained and the upper part of the peat was excavated for use as a fuel. Deeper peat layers, consisting of sedge and reed peat, were less suitable for fuel and left behind. To prepare the soil for agriculture, peat remains were covered with sand from the vast network of canals that were dug to drain the area and to transport the turf. The sand cover was levelled and slivers of subsoil peat were mixed through it. Large areas of the reclaimed agricultural peat soils were deeply cultivated during the second part of the 20th century, which dislocated subsoil peat and sand layers. A few parts of the study area, mainly in the southeast corner and central area, were never reclaimed and cultivated. These areas are now highmoor nature reserves (1 800 ha). Not surprisingly, as a result of the methods for reclaiming and cultivating the former highmoor swamps there is considerable short-range variation in soil conditions such as the topsoil organic matter content and the thickness of the remaining peat layer. Furthermore, spatial heterogeneity at short distances increased over time by decomposition of peat remains caused by deep drainage and intensive agricultural use.

Soils

The existing 1:50 000 national soil map of the study area, which was completed in the 1970s and has become outdated, distinguishes four major soil types: three peat soils and one mineral soil (Fig 6.1). Currently, a fifth soil type is found now that the peat layer has been worn away by oxidation in large parts of the study area. The 1:50 000 soil legend further subdivides these major soil types on the basis of for instance presence of gley mottles and soil texture of mineral soils, thickness of the peat layer in deep peat soils and peat type. Here, however, digital soil mapping focusses on the five major soil types. This is a limitation of DSM methods for soil type mapping; for CSM there is obviously no need for such generalization. The five

major soil types are (reported areas refer to areas on the existing 1:50 000 map of the study area):

- *Reclaimed shallow peat soils (iW)*. Soils with at most 40 cm of peat within 80 cm from the surface and a man-made topsoil. Topsoil SOM content typically ranges between 5–25% and can be highly variable at short distances as a result of the reclamation process. These soils cover 3 590 ha or 22.4% of the area.
- *Reclaimed deep peat soils (iV)*. Like the reclaimed shallow peat soils but with at least 40 cm of peat within 80 cm from the surface. The peat layer can extend below auger observation depth ($\sim 120 - 150$ cm below surface). These soils cover 9 025 ha (56.2%).
- *Raw peat soils (V)*. These are unreclaimed peat soils found in the few remaining highmoor bogs and swamps. The original peat layer is largely intact and the man-made topsoil is absent. These soils cover 3 175 ha (19.8%).
- *Podzols (Hn)*. Soils with a podzol-B horizon that originally developed in the late-Pleistocene sands before these were covered by peat in the Holocene. These soils cover 266 ha (1.6%). Peat remains can be present in the soil profile, often mixed with mineral material.
- *Hydromorphic earth soils (pZ)*. These sandy soils developed in the late-Pleistocene sands under somewhat wetter conditions than the podzols. They have a dark, humic topsoil that directly overlays the C-horizon. Gley mottles can be present in the subsoil. Although earth soils were not mapped in the study area during the 1:50 000 soil survey, these are now found at locations where the overlying peat layer has completely disappeared. Like the podzols, these soils can have peat remains in the profile.

6.2.2 Data

Soil point data

The point dataset was collected by the free survey method during a soil survey in 2009 and 2010 and contains 4 168 observations (Fig. 6.3, top-left panel). Two types of observations can be distinguished: soil profile observations ($n = 1\,715$) and delineation observations ($n = 2\,453$). The former are made to determine where the various soil mapping units occur. These observations locations are chosen purposefully by the soil surveyor—who considers the selected locations ‘representative’ for the agricultural field that is surveyed—and have fairly regular spatial coverage, except in nature reserves (forests, highmoor swamps) where locations are selected based on accessibility. At the selected locations auger borings were made to describe

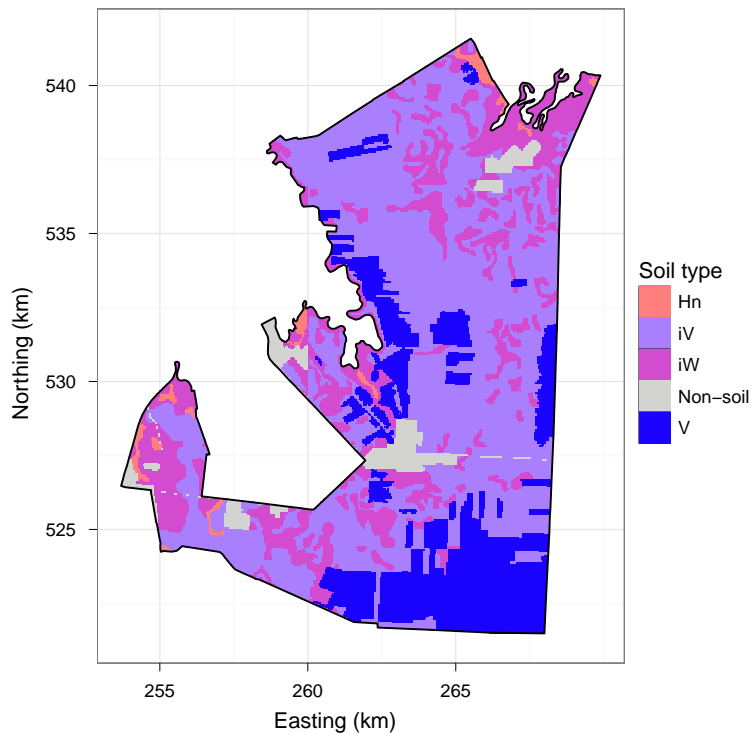


Figure 6.1: The 1:50 000 national soil map for the study area with generalized legend.

and classify soil profiles according to the Dutch soil classification system for detailed soil surveys (ten Cate et al., 1995). The SOM, sand and loam contents of the soil horizons were estimated by hand. Delineation observations are made to assist plotting of the map delineation boundaries. At the locations of these observations the soil was classified (often not in full detail) but not described. Delineation observations can therefore not be used for soil property mapping. It is worth noting here that DSM typically follows a different sampling strategy (e.g. Brus and Heuvelink, 2007a; Walvoort et al., 2010) than CSM. The sampling strategy applied in study can, therefore, be suboptimal for DSM.

Environmental covariates

Nine data layers were available from which a suite of 27 environmental covariates was derived (Table 5.1 provides an overview):

- *Digital elevation model (DEM)*. A 25-m resolution DEM, constructed from LI-

DAR measurements¹. From the DEM four relative elevation layers were derived by subtracting the local mean elevation within search radii of 250, 500, 750, and 1000 m from the actual elevation. These layers capture local relief. One relative elevation layer (based on the 750-m search radius) was reclassified into layers with two, three and four classes to account for a possible non-linear relation between soil and relative elevation. In addition, a layer with historic elevation was constructed by inverse distance weighted interpolation of a network of elevation measurements from the 1960s (1.2 ha⁻¹). A layer representing elevation change was computed from the actual and historic elevation layers and reclassified into two classes. Elevation change is informative because peat excavation and decomposition lower the surface.

- *Land cover maps*. This set contains five layers depicting land cover in 1900 (50-m resolution), 1940 (25 m), 1960 (25 m), 1980 (25 m) and 2003 (25 m) (Clement and Kooistra, 2003; Knol et al., 2003, 2004; Hazeu, 2005). The layer from 1900 distinguishes ten classes and was reclassified into a layer with two classes and into a layer with three classes. The layer from 2003 distinguishes 23 classes and was reclassified into four layers, each with a different combination of classes. The layers from 1900, 1940, 1960, 1980 and 2003 were combined into a map that represents reclamation period, i.e. the period when reclamation for agriculture took place. Four layers were derived from the reclamation period map, each with a different combination of reclamation periods.
- *National Soil Map* (Steur and Heijink, 1991). Polygon map at scale 1:50 000 as described in section 6.2.1. In addition to soil type, five soil variables were derived from map unit classifications: one variable representing peat type, three variables representing peat thickness, and one variable indicating the status (degraded or not degraded) of peat soils *iV* and *V* based on the results of a quick-scan of these soils (van Kekem et al., 2005).
- *Paleogeography map* (Spek, 2004). Polygon map at scale 1:50 000 representing a reconstruction of the geography of Drenthe around 1 000 AD, distinguishing 12 geographical units. The map was converted to a layer of 25-m resolution, from which two new layers were extracted: one with the former extent of fen peat, one with the former extent of highmoor.

6.2.3 Updating soil maps

This section briefly describes the methods used in this study to generate the updated conventional and digital soil type and property maps. For a more elaborate description of the DSM methods for soil property mapping see Lark and Webster (2006) and Chapters 3 and 4 and for soil type mapping Diggle et al. (2003); Diggle and Ribeiro Jr. (2007) and Chapter 5.

¹www.ahn.nl

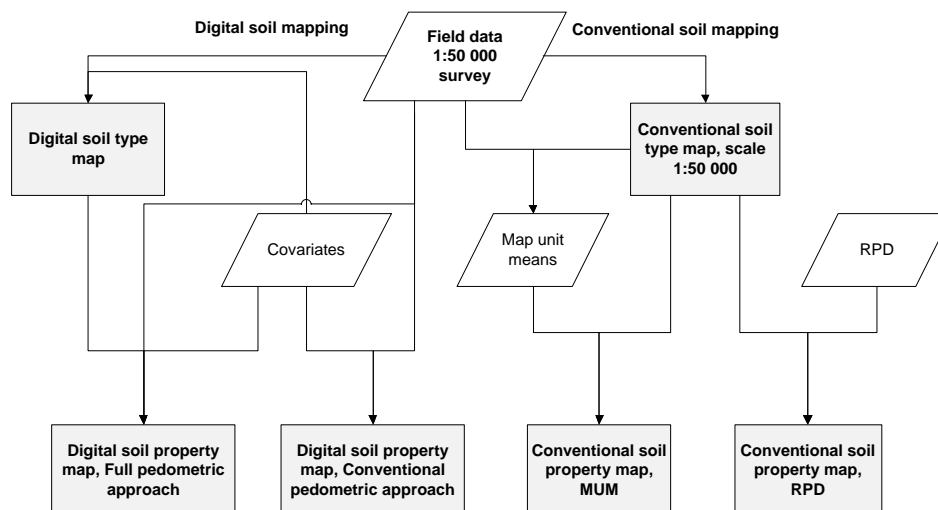


Figure 6.2: Schematic overview of the soil mapping methods.

Digital soil type mapping was done with the `geoRglm` package (Christensen and Ribeiro Jr., 2002) and digital soil property mapping with the `geoR` and `gstat` packages (Ribeiro Jr. and Diggle, 2001; Pebesma, 2004) in `R` (R Development Core Team, 2008). In addition to soil type, the soil organic matter (SOM) content (in mass%) of the topsoil (0–20 cm) and the thickness of the peat layer within 150 cm from the soil surface were mapped. Two digital and conventional methods to generate the soil property maps. For CSM these were i) representative soil profile descriptions (RPs) and ii) map unit means (MUMs). For DSM these were i) a full pedometric approach (FPA), meaning that spatial prediction involves the updated digital soil type map, and ii) the conventional pedometric approach (CPA), which is standard universal kriging. Fig. 6.2 provides a schematic overview of the mapping methods.

The FPA was applied in addition to the CPA to investigate the merits of using an updated 1:50 000 digital soil type map for spatial prediction. Soil type has proven to be a useful covariate for soil property mapping (e.g. Heuvelink and Bierkens, 1992; Brus et al., 1996; Voltz et al., 1997; Goovaerts, 2011). It might therefore be worthwhile to update the soil map first and next use the resulting map as covariate for spatial prediction, instead of using the existing, outdated soil map. An additional advantage of using a digital soil type map (that represents soil type by probability distributions) for prediction is that in that case soil type observed at the sampling locations may be used as predictor (Chapter 3). Observed soil type typically has a stronger predictive relationship with a soil property than mapped soil type because this relationship is not confounded by soil map impurities.

Soil type maps

Conventional soil mapping

The conventional soil type map was created by four experienced soil surveyors using free survey at scale 1:50 000 with over 4 000 observations of soil type. A DEM with 5-m resolution was used to assist delineation of soil boundaries. Each map unit was recoded to one of the five major soil types so that the map has the same legend as the digital soil type map for accuracy comparison. A preliminary measure of accuracy was computed from the soil map: the calibration purity. This is the proportion of observation locations at which the soil map predicts the correct soil type.

Digital soil mapping

The generalized linear geostatistical model (GLGM) was used to map soil type. The GLGM is central to the model-based geostatistical framework (Diggle et al., 1998) and can be used to model non-Gaussian distributed spatial data. The GLGM has three components (Diggle and Ribeiro Jr., 2007). The first component is the *signal process* $S(\cdot)$, which is a real-valued Gaussian spatial process with $E[S(\mathbf{x})] = m$, $\text{var}[S(\mathbf{x})] = \sigma^2$, and correlation function $\rho(\mathbf{h})$. Here m is a spatial trend $\mathbf{d}(\mathbf{x}_i)^T \boldsymbol{\beta}$, where $\mathbf{d}(\mathbf{x}_i)$ is a vector of explanatory variables at location \mathbf{x}_i and $\boldsymbol{\beta}$ is a vector of trend coefficients. The second component is the *measurement process* $Y(\cdot)$. Realizations of this process are the observed data $y(\mathbf{x}_1), \dots, y(\mathbf{x}_n)$, which are considered measurements of the signal process. The $Y(\mathbf{x}_i)$ are assumed to follow a common distributional family (e.g. Bernoulli, Poisson, binomial or Gaussian), depending on the mechanism that generated the data, and are mutually independent conditional on the signal. The responses have conditional means $E[Y(\mathbf{x}_i)|S(\cdot)]$. The third component is the *link function* $g(\cdot)$, which links the conditional mean $E[Y(\mathbf{x}_i)|S(\cdot)]$ to the linear predictor $S(\mathbf{x}_i)$. The GLGM is thus defined as:

$$g(E[Y(\mathbf{x}_i)|S(\cdot)]) = S(\mathbf{x}_i) = \mathbf{d}(\mathbf{x}_i)^T \boldsymbol{\beta} + U(\mathbf{x}_i), \quad (6.1)$$

where $U(\mathbf{x}_i)$ is a second-order stationary, Gaussian distributed, spatial process with zero mean and variance σ^2 .

Soil type data can be considered multinomial data when the number of outcome categories, the soil types, is larger than two. Theory for fitting a GLGM for multinomial-distributed data, however, has not been worked out, nor is there available software that can fit such model. Therefore a pragmatic approach was adopted in which each soil type k , $k = 1, 2, \dots, K$, was considered a Bernoulli-distributed random variable Y_k , which can be modelled with the binomial logit-linear GLGM (Chapter 5). Realizations $y_k(\mathbf{x}_i)$ of $Y_k(\mathbf{x}_i)$ take value 1 if soil type k is observed at a sampling location i , $i = 1, 2, \dots, n$, and 0 otherwise. There are five soil types in the study area. Each soil type was modelled separately with a logit-linear GLGM. Next, the soil type-specific GLGMs were used to predict five binomial probabilities at the nodes of a 50-m raster

covering the study area, which were scaled to multinomial probabilities so that they sum to one. The soil type with the largest probability at each location was used to create the soil map.

Estimation of the model parameters of the GLGM as well as spatial prediction with this model is complex. It involves repetitive use of Markov Chain Monte Carlo methods (Minasny et al., 2011) to obtain simulations of the unobserved signal process $S(\cdot)$ and Monte Carlo maximum likelihood estimation of the model parameters (Christensen, 2004). These methods are discussed in detail in section 5.2. Also for the digital soil type map the calibration purity was computed.

Soil property maps

Conventional soil mapping

Note that to derive the RPD and MUM soil property maps the updated conventional soil type map with the full 1:50 000 legend has been used.

Representative soil profile descriptions. The two soil property values were derived from the RPDs (de Vries, 1999) associated to the map units of the updated soil map. These values were then assigned to the map delineations to which they are linked. All locations within a map unit thus receive the same value.

Soil map unit means. At each soil profile observation location ($n = 1\,715$) the two soil property values were derived from the soil profile description. Next, the average for each property was computed for each map unit from the sampling locations located within the selected map unit. Like in the RPD method, all locations within a map unit receive the same value.

Digital soil mapping

Both soil properties were spatially predicted by universal kriging (Lark and Webster, 2006). This model can be defined as:

$$Z(\mathbf{x}) = \mathbf{d}(\mathbf{x})^T \boldsymbol{\beta} + \epsilon(\mathbf{x}), \quad (6.2)$$

in which $Z(\mathbf{x})$ denotes the soil property, $\mathbf{d}(\mathbf{x})$ is a vector of covariates at location \mathbf{x} , $\boldsymbol{\beta}$ is a vector of trend coefficients and ϵ is a zero-mean, second-order stationary, spatial dependent random function. Once the parameters of the universal kriging model—the trend parameters and the variance parameters of the residual—are estimated, preferably by residual maximum likelihood (REML) (Lark and Cullis, 2004), then these can be used to predict soil property Z at prediction location \mathbf{x}_0 :

$$\hat{Z}(\mathbf{x}_0) = \mathbf{d}(\mathbf{x}_0)^T \boldsymbol{\beta} + \sum_{i=1}^n \lambda_i (\mathbf{z}_i - \mathbf{d}(\mathbf{x}_i)^T \boldsymbol{\beta}), \quad (6.3)$$

in which λ_i is the kriging weight associated to sampling location \mathbf{x}_i , $i = 1, 2, \dots, n$, and $z(\mathbf{x}_i)$ is the observed soil property at \mathbf{x}_i . Two different methods were used to calibrate the universal kriging model.

Full pedometric approach. In the FPA the actual (observed) soil type at the sampling locations is used as a covariate in the universal kriging model. This implies that at prediction locations actual soil type must be used to predict the soil property. The actual soil type is unknown at unsampled locations but can be represented by a probability model (Chapter 3), i.e. the digital soil type map. This method requires soil type-specific predictions of the soil properties at each prediction location which are averaged, using the probabilities as weights, to obtain the final prediction. For further details on model selection and spatial prediction see Chapter 4, in which a similar procedure as here was followed. Both properties were predicted for 25-m blocks centered at the nodes of a 50-m raster that covers the study area.

Conventional pedometric approach. The CPA is the standard application of universal kriging for soil spatial prediction. The same procedure for covariate selection and parameter estimation as for the FPA was followed. Both soil properties were predicted for 25-m blocks centered at the nodes of a 50-m raster by applying Eq. 6.3.

6.2.4 Mapping efficiency: effort

Mapping efficiency is here defined as the accuracy of a soil map that can be obtained given a certain amount of effort invested in the mapping process. Thus method A is more efficient than method B when method A produces a soil map of similar accuracy as method B but with less effort or produces a map of larger accuracy with similar effort. In this study mapping effort was expressed in observations ha^{-1} as well as € ha^{-1} . Note that the latter is not directly proportional to the number of observations used to create a soil map.

The aim of the mapping efficiency assessment was two-fold: i) to compare the efficiencies of CSM and DSM, ii) to assess the effect of different mapping efforts on the accuracy of the digital soil maps. For the second aim four point datasets with different sampling intensities were used. For each of these the cost of mapping in € ha^{-1} was calculated. Unfortunately, the effect of different mapping efforts on the accuracy of soil maps could only be assessed for DSM and not for CSM, since conventional soil type maps based on different sampling intensities were not produced during the 2009/2010 soil survey.

This section explains how the mapping effort of the various soil mapping methods was computed, while in the next section focusses quantifying map accuracy. In this chapter 'days' refers to person-days.

Conventional soil mapping

Sampling intensity. The soil type map is based on 4 168 field observations, which equals an sampling intensity of 0.3 ha^{-1} . Soil property maps are generated from point data with the same density since this requires the soil type map.

Costs. Only the direct costs for production of a 1:50 000 soil map and a dataset with soil profile descriptions were considered. These are all variable costs, depending on the mapped hectarage. Fixed costs such as write-off of materials, server costs for hosting the soil information system and reporting were not taken into account. The costs of a conventional soil map can be divided into four cost components, all of which are a function of area:

- *Fieldwork preparation:* $3\,500 \text{ ha day}^{-1}$. This includes preparation of field maps, elevation maps and landowner maps.
- *Fieldwork:* 75 ha day^{-1} . Fieldwork consists of describing and classifying soil profiles at observation locations from auger bores, collecting plotting observations and preliminary drawing of soil boundaries on the field maps. After consulting experienced soil surveyors it was estimated that an experienced surveyor can map 75 ha per day at scale 1:50 000.
- *Operational costs:* 125 € day^{-1} . These cover operational expenses such as car fuel.
- *GIS work:* 500 ha day^{-1} . Digitizing the soil boundaries from the field maps and entering them into a GIS. Also a soil attribute table is prepared that stores the attributes associated with the soil map polygons. The result is a GIS-layer of the soil map that can be uploaded to the soil information system.
- *Soil profile dataset compilation:* $1\,000 \text{ profile descriptions day}^{-1}$. Processing of the soil profile descriptions. The descriptions are read from a field computer, checked for errors and then converted into a data format that can be uploaded to the soil information system and used for other soil data applications.

The total cost per component was estimated from the study area size and the cost per person-day. From these the total cost of a conventional soil survey was calculated and expressed as the mapping effort in € ha^{-1} . The cost of the soil property maps equals the cost of the soil type map plus the cost of one extra person-day to apply the RPD and MUM methods.

Digital soil mapping

Sampling intensity. Four digital soil type maps were produced using four point datasets with different sampling intensities (Fig. 6.3). Soil property maps were pro-

duced using three point datasets (datasets 2, 3 and 4; listed below) because for these datasets field observations on the soil properties were available.

1. *Complete dataset* ($n = 4\,168$). Sampling intensity is 0.30 observations ha^{-1} . The soil type map created with this dataset reflects how well DSM performs when exactly the same data are used as for the conventional soil type map.
2. *Soil profile observations only* ($n = 1\,715$). Sampling intensity is 0.12 observations ha^{-1} .
3. *Two-thirds of the soil profile observations* ($n = 1\,146$). Sampling intensity is 0.08 observations ha^{-1} . The observation locations were selected from dataset 2 by spatial coverage sampling (Walvoort et al., 2010) so that the selected observation locations cover the study area as uniformly as possible.
4. *One-third of the soil profile observations* ($n = 572$). Sampling intensity is 0.04 observations ha^{-1} . Subset selection as for dataset 3 but with lower density.

Costs. The cost of a digital soil mapping can be divided into five variable cost components (V ; costs are a function of area or method) and three fixed cost components (F):

- *Fieldwork preparation: 3 500 ha day⁻¹, (V).* Preparation of field and landowner maps.
- *Environmental data collection: 5 days, (F).* Collection and processing of environmental covariates.
- *Data collection: n observations day⁻¹, (V).* Fieldwork effort for DSM is measured in the number of soil observations that can be collected in one day and depends on sampling intensity. Lower densities mean that observation locations are more widely spaced. This means that more time is required to reach the locations since distances have to be covered by foot. Two experienced soil surveyors provided estimates on how many soil profile descriptions from auger bores can be collected in one day for the sampling intensities of datasets 2-4.
- *Operational costs: 125 € day⁻¹, (V).* See CSM.
- *Soil profile dataset compilation: 1 000 profile descriptions day⁻¹, (V).* See CSM.
- *Digitizing delineation observations: 600 observations day⁻¹, (V).* This cost component only applies to dataset 1 as this is the only dataset that includes the delineation observations.
- *Preparation point dataset: 1 day, (F).* Preprocessing of the dataset with soil profile descriptions (e.g. deriving the soil properties of interest from the profile descriptions).
- *Geostatistical modelling: n days, (F).* It was assumed that computer scripts for spatial interpolation of soil type and soil property data were available. This is a likely scenario in the near future (Heuvelink et al., 2010). Only the time was

counted that it would take an experienced pedometrician to tailor the computer scripts to a specific case study, run the computer scripts, inspect the results and adjust the scripts if deemed necessary. For soil type mapping it was estimated that this requires 5 days as each of the five soil types has to be modelled separately. Soil property mapping with the FPA was estimated to take 7 days: 5 days to map soil type and 2 days to prepare and run soil-type specific universal kriging models. Soil property mapping with the CPA was estimated to take 1 day.

Like for CSM, the total cost per component was estimated from the study area size and the cost per person-day. From these the total cost of each digital soil mapping was calculated as well as the mapping effort in € ha⁻¹. Table 6.1 shows which of the soil maps produced in this study are considered for assessment of mapping efficiency and Table 6.2 of the cost components of the various soil mapping methods.

Table 6.1: Overview of the soil maps produced in this study.

Dataset	Mapping effort (obs. ha ⁻¹)	Conventional soil mapping			Digital soil mapping		
		Soil type	Soil properties		Soil type	Soil properties	
			RPD	MUM		FPA	CPA
1	0.30	X	X	X			
2	0.12				X	X	X
3	0.08				X	X	X
4	0.04				X	X	X

Table 6.2: Overview of the main cost components of each of the soil mapping methods. ST refers to soil type and SP to soil property.

Cost component	Sampling intensity							
	0.30 ha ⁻¹		0.12 ha ⁻¹		0.08 ha ⁻¹		0.04 ha ⁻¹	
	ST	SP	ST	SP	ST	SP	ST	SP
<i>Conventional soil mapping</i>								
Fieldwork preparation (ha day ⁻¹)	3500	3500						
Soil mapping (ha day ⁻¹)	75	75						
Operational costs fieldwork (€ day ⁻¹)	125	125						
Soil profile dataset compilation (obs. day ⁻¹)	1000	1000						
Mapping in GIS (ha day ⁻¹)	500	500						
Application RPD/MUM methods (days)		1						
<i>Digital soil mapping</i>								
Fieldwork preparation (ha day ⁻¹)	3500		3500	3500	3500	3500	3500	3500
Environmental data collection (days)	5		5	5	5	5	5	5
Data collection (obs. day ⁻¹)	26		16	16	14	14	12	12
Operational costs fieldwork (€ day ⁻¹)	125		125	125	125	125	125	125
Soil profile dataset compilation (obs. day ⁻¹)	1000		1000	1000	1000	1000	1000	1000
Digitizing plotting observations (obs. day ⁻¹)	600							
Preparation point dataset (days)	1		1	1	1	1	1	1
Model application (days)	5		5	7/1*	5	7/1	5	7/1

* For the two methods of digital soil property mapping: FPA/CPA

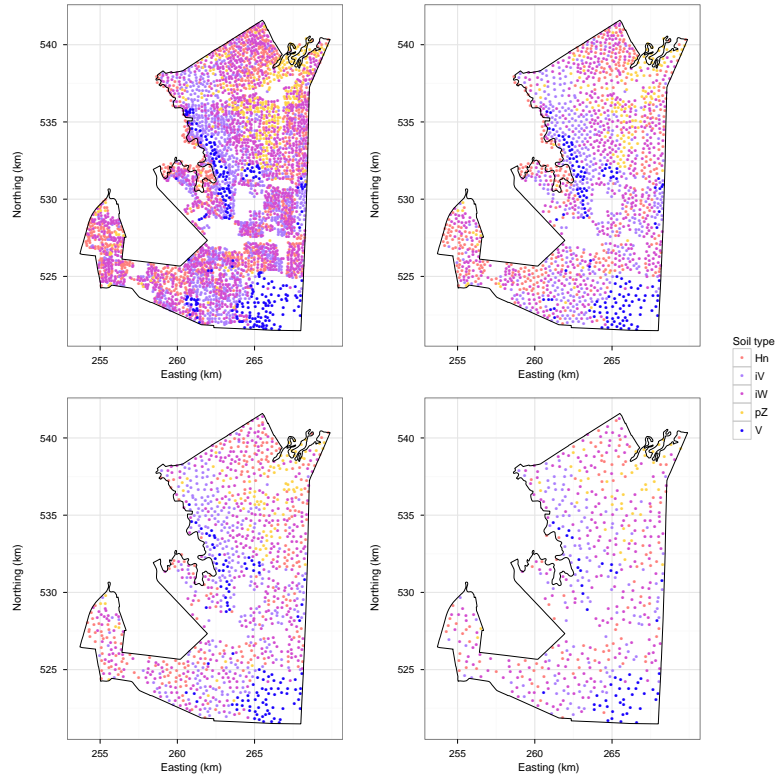


Figure 6.3: Point datasets with different sampling intensities: complete dataset (top-left, $n = 4168$), soil profile observations only (top-right, $n = 1715$), two-thirds of the soil profile observations (bottom-left, $n = 1146$), one-third of the soil profile observations (bottom-right, $n = 572$).

6.2.5 Mapping efficiency: accuracy

Sampling strategy

The soil maps were validated with independent probability sample data. These data were collected with a stratified simple random sampling design (Brus et al., 2011). The mapping units of the digital soil type map created with dataset 1 were used as strata. Within each stratum, sampling locations were selected in three steps. First, in each stratum n 50-m pixels of the prediction grid were randomly selected with replacement. Second, within each selected pixel one of the quadrants was randomly selected. Third, the selected quadrant, or 25-m block, was sub-sampled by simple random sampling. Validation data was collected at five point locations within each block.

The digital soil type maps are on point support and thus should be validated with observations on point support. The basic sampling unit for soil type, therefore, is a point location. The soil profile was described and classified from an auger bore observation at the first of the five selected locations in each selected block. Soil property maps are on 25-m block support. The sampling units for validation should have the same support. For SOM content an aliquot of the 0–20-cm soil layer was taken with a gouge auger at each of the five selected locations in each selected block. The five aliquots were bulked into a composite sample. The SOM content of this sample is used as an estimate of the block-mean. The thickness of the peat layer was determined from an auger bore observation at each of the five sampling locations. The five observed thicknesses were averaged to obtain an unbiased estimate of the mean peat thickness of the block.

A total of 125 sampling units were selected, which were allocated to the strata roughly proportional to their surface areas. Sampling units where permission was denied or proved otherwise impossible to sample were replaced with randomly selected locations from a reserve list. Fieldwork took place in March 2011. The composite soil samples were oven-dried at 105° C for at least 12 hours, and then sieved and crushed. The SOM content (in mass%) of a dry sample was determined with the weight loss-on-ignition method. The samples were combusted at 550° C for at least 3 hours. SOM content was determined from the weight difference before and after combustion.

Map quality measures

Measures of map accuracy for soil type and property maps are discussed only briefly here. For a more elaborate review, including the estimation of these measures and their variances see Brus et al. (2011) and Chapters 2 and 4.

Soil type maps. One accuracy measure was considered for the soil type maps: the overall purity. Overall purity is defined as the proportion of the mapped area in which the predicted soil type, which is the soil type as depicted on the map, equals the true soil type. In other words, it is the areal proportion correctly classified. To estimate the overall purity an indicator variable is created, which takes the value 1 if the observed soil type equals predicted soil type and 0 otherwise. For each stratum the average of this indicator is computed. The overall purity is then estimated as the weighted average of the stratum purities, with weights equal to the relative sizes of the strata.

Soil property maps. Two accuracy measures were considered for the soil property maps: the mean error (ME), which is a measure for prediction bias and the (root) mean squared error ((R)MSE), which is a measure for prediction accuracy. Estimates

of these parameters are based on the prediction error, which is the difference between predicted and true value at a validation location. The ME is estimated as the weighted average of the strata means of the error, with weights equal to the relative sizes of the strata. The MSE is estimated as the weighted average of the strata means of the squared error. The RMSE is calculated as the square root of the estimated MSE. Sampling the 25-m block at a limited number of locations induces sampling error since taking a composite sample from another set of point locations would result in a different estimate of the block mean. This results in a biased estimate of the MSE (but not of the ME) (Brus et al., 2011). The sub-sampling error was quantified by sub-sampling 25 randomly chosen blocks twice. From the 25 duplo samples the block sub-sampling variances were calculated and from these the mean sub-sampling variance. The raw estimate of the MSE estimate was corrected by subtracting the mean sampling variance.

Map comparison. To compare the accuracies of the pedometric soil maps with those of the conventional soil maps, variable q_i is defined. For the categorical soil type maps q_i is calculated as $y_i^d - y_i^c$, where y_i is an indicator that takes value 1 or 0 at validation location i , superscript d indicate digital soil map and c conventional soil map. The mean purity difference (MPD) of two soil maps, \hat{q} , is estimated in similar fashion as the overall purity. To test whether the estimated MPD differs significantly from zero it was assumed that the estimated MPD follows a normal distribution. For the soil property maps, q_i was calculated as the difference between the squared errors: $(e_i^d)^2 - (e_i^c)^2$.

6.3 Results and discussion

6.3.1 Conventional soil mapping

Soil type map

Fig. 6.4 shows the soil type map. Extensive changes in soil are evident when this map is compared with the simplified 1:50 000 soil map made in the 1970s (Fig. 6.1), illustrating the necessity to update the current map for the cultivated peatlands. The updated and old soil maps correspond for only 31% of the area. The calibration purity is 83.3%.

Soil property maps

The soil property maps are shown in Fig. 6.5. The SOM RPD map predicts generally larger SOM contents than the MUM map. The MUMs of the reclaimed peat soils are roughly 5% smaller than the SOM contents derived from RPDs whereas the MUMs

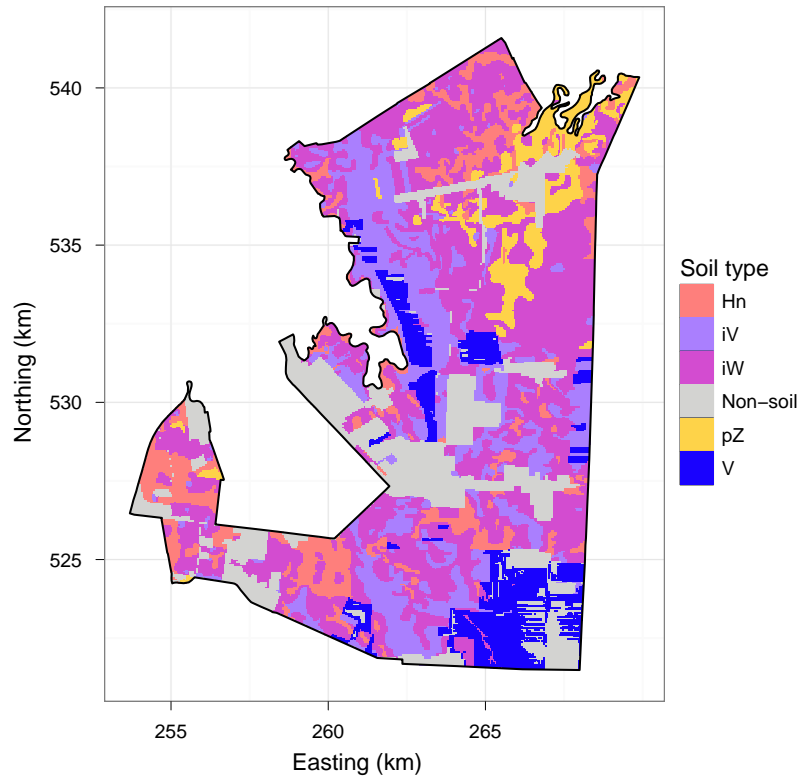


Figure 6.4: Soil type map created with conventional soil mapping.

of the raw peat soils are 10-25% larger. Differences between predicted peat layer thickness are small. The MUM method predicts somewhat less thick peat layers (5–10 cm) for the reclaimed shallow peat soils. Also, the MUM method predicts very shallow peat layers (3–7 cm) in the mineral map units whereas the RPD method predicts absence of peat. These very shallow layers are remains of much thicker peat layers that were once present.

6.3.2 Digital soil mapping

This section presents and discusses the digital soil type and property maps created from all available field data. That is dataset 1 for soil type mapping ($n = 4\,168$) and dataset 2 for soil property mapping ($n = 1\,715$). The effects of sampling intensity on geostatistical modelling and spatial prediction are discussed in next section.

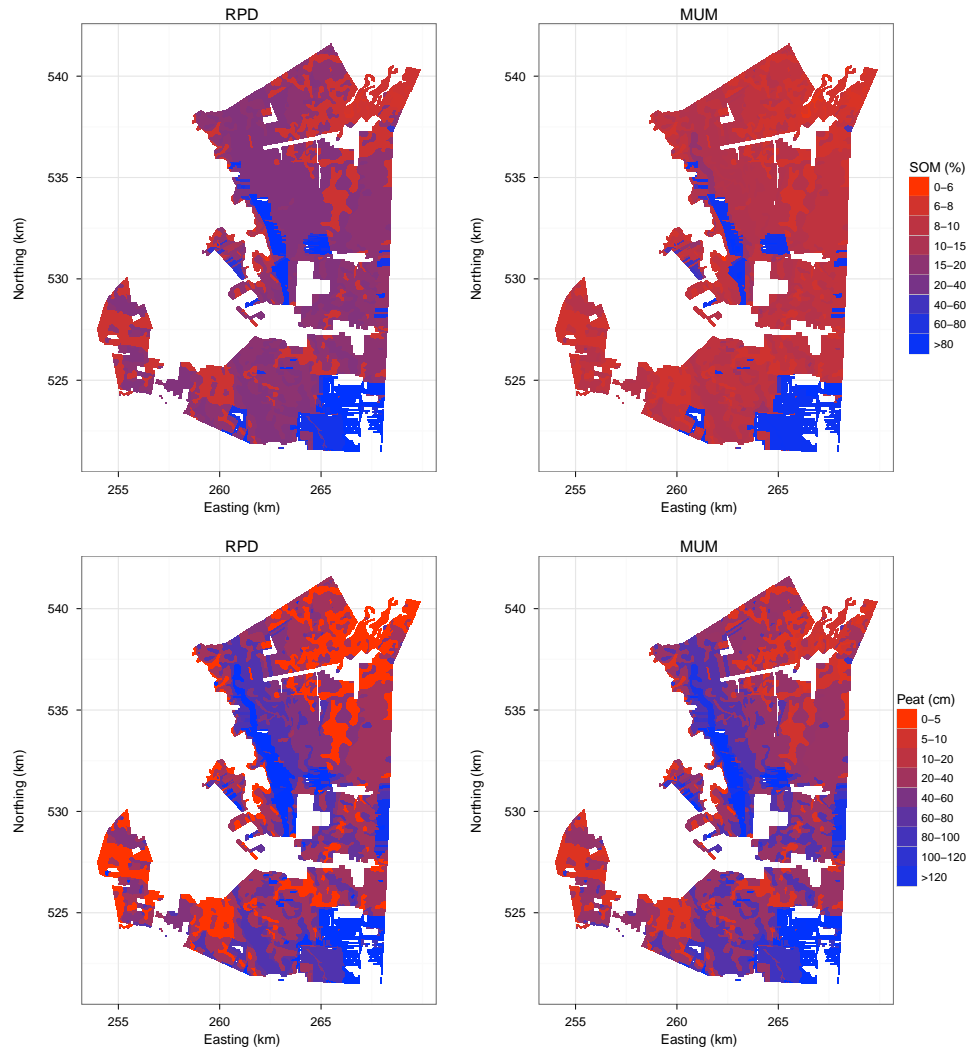


Figure 6.5: Maps depicting the organic matter content of the 0–20-cm soil layer and the thickness of the peat layer as predicted by conventional soil methods: representative profile descriptions (RPD) and soil map unit means (MUM).

Soil type map

Table 6.3 provides the selected models and the estimates of the variance parameters for digital soil type mapping with the GLGM. The environmental covariates are listed in order of the strength of correlation between covariate and soil type based on univariate analysis. Not surprisingly, the 1:50 000 soil map and its derivatives (peat

Table 6.3: Selected trend models and estimates of the variance parameters for digital soil type mapping.

Soil type	n	Covariates	Variance	a^a (m)	Spatial dep. ^b
<i>0.30 obs. ha⁻¹, n = 4 168</i>					
Hn	935	PEATTHK3+EC2+PEATTYP+COV3a+REL3	0.99	726	0.34
iV	1 023	PEATTHK2+SOIL+COV3a+EC2+REL1000+PEATSTAT	0.95	846	0.34
iW	1 670	COV3a+PEATTHK2+RECLAM3+SOIL+REL4+HIGHMOOR	1.24	348	0.17
pZ	303	PEATTYP+FENPEAT+PEATTHK3	0.47	1 112	0.39
V	237	COV3a+PEATTHK2+PEATTYP	0.06	387	0.24
<i>0.12 obs. ha⁻¹, n = 1 715</i>					
Hn	414	PEATTHK3+EC2+PEATTYP+COV2+REL2	0.84	1 015	0.26
iV	394	PEATTHK3+COV3a+PEATSTAT	0.78	850	0.34
iW	632	PEATTHK3+COV3a+PEATCOL+REL2	1.23	262	0.54
pZ	131	PEATTYP+FENPEAT+PEATTHK3+COV2	0.47	1 254	0.45
V	144	COV3a+PEATTYP	0.09	339	0.33
<i>0.08 obs. ha⁻¹, n = 1 146</i>					
Hn	246	PEATTHK3+PEATTYP+COV2+REL2	0.71	850	0.32
iV	272	PEATTHK3+COV3a+PEATSTAT	0.76	569	0.22
iW	432	PEATTHK3+COV3a+REL2	1.30	540	0.24
pZ	89	PEATTYP+FENPEAT+PEATTHK3	0.41	640	0.39
V	107	COV3a+PEATTYP	0.07	266	0.30
<i>0.04 obs. ha⁻¹, n = 572</i>					
Hn	105	PEATTHK2+PEATTYP+REL2	0.91	633	0.30
iV	134	PEATTHK3+COV3a+PEATSTAT	0.88	597	0.56
iW	235	PEATTHK3+COV3a+REL2	1.05	300	0.65
pZ	46	PEATTYP+FENPEAT+PEATTHK2	0.46	401	0.33
V	52	COV2a	0.08	557	0.24

^a Distance parameter of the exponential model. The practical range is $3a$.

^b The spatial dependence is the ratio between spatially structured variance and total variance [partial sill/(nugget+partial sill)].

type and peat layer thickness) are strong predictors of current soil type. Also the land cover map with three classes (agriculture, forest/rangeland, highmoor) proved to be an effective predictor. This can be explained by the fact that mineral soils *Hn* and *pZ* and thin agricultural peat soils almost exclusively occur under agriculture whereas raw peat soils almost exclusively occur in (forested) highmoor areas. Deep agricultural peat soils (*iV*) mainly occur under agriculture but are also found under rangeland, where they are the dominant soil type. The rangeland area was under agriculture until the mid-1990s. At that time agriculture was abandoned and natural vegetation restored, which slowed peat decomposition. Covariates related to elevation and landscape type were also significant but less important. Roughly 15 to 40% of the residual variance is spatially structured at medium distances (practical range of the exponential variogram models varies approximately from 1000–3500 m). The weak to moderate spatial autocorrelation highlights the short-distanced heterogeneity in soil conditions in the cultivated peatlands.

Fig. 6.6 (top-left) shows the soil type map generated with dataset 1. The calibration

purity of the digital soil type map is 61.9%. The general spatial pattern of soil distribution is similar to that of the conventional soil map. The digital and conventional maps correspond for 70% of the area. Mineral soils have smaller area on the digital map than on the conventional map (2 500 ha versus 3 000 ha). The maps have the smallest correspondence for map unit *pZ*: only 38% of the area depicted on the conventional soil map is mapped as *pZ* on the digital soil map. For the podzol unit *Hn* this is 55%. Map unit *iW* has larger area on the digital map (6 900 ha versus 6 350 ha).

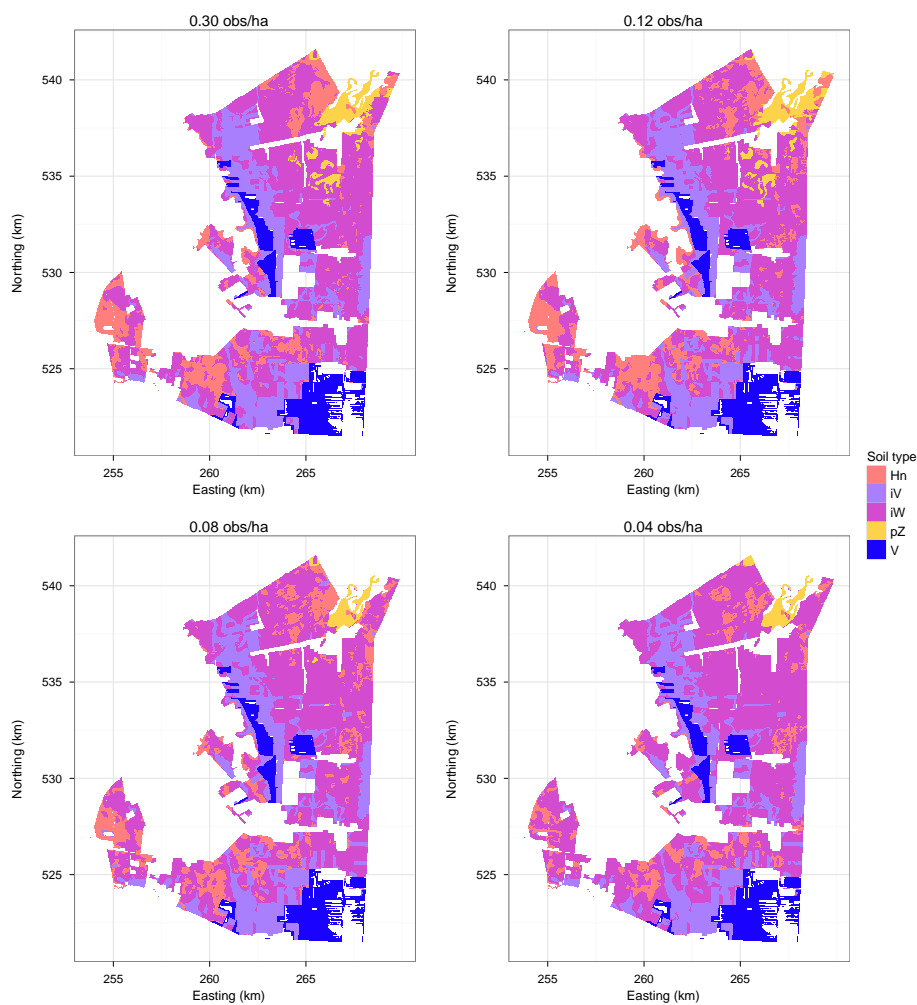


Figure 6.6: Soil type maps created with digital soil mapping using four different sampling intensities.

Soil property maps

Table 6.4 lists the selected models and the REML estimates of the variance parameters for digital soil property mapping with the full pedometric approach (FPA) and conventional pedometric approach (CPA). The observed SOM contents were log-transformed to reduce positive skewness of the residuals (FPA: 1.92, CPA: 3.33). The predictions and their variances were back-transformed for visualization and validation. The peat thickness residuals were mildly skewed only (FPA: 0.31, CPA: 0.66) and were left on the original scale. As expected, observed soil type was for both properties the most important predictor in the FPA. The FPA uses observed soil type as covariate in the trend model. Mapped soil type was offered as covariate in the CPA but was not selected for the most parsimonious model. Instead ‘peat thickness class’ and ‘land cover’ were selected for SOM content (both indicative for soil type, Table 6.3), whereas soil map-derived covariates ‘peat thickness class’ with three outcome levels and ‘peat status’ were selected for peat layer thickness. The FPA trend models fitted the observed data (much) better than the CPA trend models, as judged from the fraction of explained variance. Observed soil type proved a strong predictor for peat layer thickness: the residual variance in the FPA is less than half of that in the CPA. This is not surprising because peat layer thickness is the most important diagnostic soil property of the five soil types considered in this study. For both soil properties and pedometric methods, between 35 and 40% of the residual variance is spatially structured at practical ranges between 2 000 and 3 000 meters.

Table 6.4: Selected prediction models for digital soil property mapping.

Obs. density (ha ⁻¹)	<i>n</i>	Covariates	<i>R</i> ^{2*}	Variance	<i>a</i> [†] (m)	Spatial dep. [‡]
<i>SOM content - full pedometric approach</i>						
0.12	1 715	SOIL.OBS+RECLAM3	0.73	0.155	731	0.39
0.08	1 146	SOIL.OBS+RECLAM3	0.74	0.159	999	0.33
0.04	572	SOIL.OBS+RECLAM3	0.78	0.172	943	0.38
<i>SOM content - conventional pedometric approach</i>						
0.12	1 715	PEATTHK2+COV3a+RECLAM3	0.63	0.218	718	0.37
0.08	1 146	PEATTHK2+COV3a+RECLAM3	0.66	0.206	831	0.33
0.04	572	PEATTHK2+COV3a+RECLAM3	0.66	0.215	851	0.36
<i>Peat layer thickness - full pedometric approach</i>						
0.12	1 715	SOIL.OBS+COV3a+PEATTHK2	0.80	374	769	0.36
0.08	1 146	SOIL.OBS+PEATTHK2	0.80	399	569	0.38
0.04	572	SOIL.OBS+PEATTHK2	0.79	411	401	0.65
<i>Peat layer thickness - conventional pedometric approach</i>						
0.12	1 715	PEATTHK3+COV3a+PEATSTAT	0.55	890	1 060	0.40
0.08	1 146	PEATTHK3+COV3a+PEATSTAT	0.56	882	1 117	0.33
0.04	572	PEATTHK3+COV3a+PEATSTAT	0.55	884	522	0.41

* The *R*² of the selected trend model.

† Distance parameter of the exponential model. The practical range is 3*a*.

‡ Spatial dependence: the ratio between spatially structured variance and total variance.

Fig. 6.7 (top- and bottom-left) shows the maps of topsoil organic matter content created with the two pedometric approaches. The maps are very similar with only some minor difference along the eastern border and slightly larger SOM contents in the north area in the FPA map compared with the CPA map. Large SOM contents are predicted in the highmoor areas with raw peat soils. Predicted SOM contents in the agricultural areas are typically smaller than 10%, with somewhat larger SOM contents for the reclaimed peat soils than for the mineral soils. An explanation for the similarity might be that topsoil organic matter contents of the agricultural soils (Hn , pZ , iV , iW) are very similar. As a result the FPA estimate is largely independent of the estimated probabilities.

Fig. 6.8 (top- and bottom-left) shows the maps of predicted peat layer thickness. Again the general spatial patterns are very similar, with deep peat layers along the north-western and east-central borders and the southeast corner. These areas coincide with the areas where deep peat soils are found. At a more detailed level, however, the two digital maps show distinct differences. The spatial pattern in the FPA map is controlled by the predicted probability distributions and mark either areas of large confusion between soil types or small confusion, resulting in a less smooth predicted surface compared with the CPA map. In general, the predicted surfaces of the pedometric maps are much smoother than those of the conventional maps because of kriging and probability weighted averaging (in case of FPA). Also predictions by pedometric methods are location-specific and follow the spatial pattern of the field data.

6.3.3 Mapping efficiency: effort

Sampling intensity

Table 6.3 shows the soil type-specific trend models and variance parameters for each of the four point datasets for digital soil type mapping. The trend models are very similar. Covariates derived from the current 1:50 000 soil map remain the most important predictors together with land cover and, to a lesser extent, elevation-related covariates. Only some of the less important covariates are dropped from the models compared with the trend model for the full dataset. They lost their significance as a result of a decreasing number of calibration observations. Estimates of the variance parameters are reasonably stable for the four sampling intensities. Reducing sampling intensity did not appear to have a very pronounced effect on the information content of the data. Given the great similarities between the geostatistical models, it is not surprising that the soil type maps created by these models are also very alike, as shown in Fig. 6.6. All maps show a similar general pattern of soil spatial variation, with raw peat soils (V) in the southeast and west-central regions, deep agricultural peat soils (iV) stretching along the western border and in the south-central part, min-

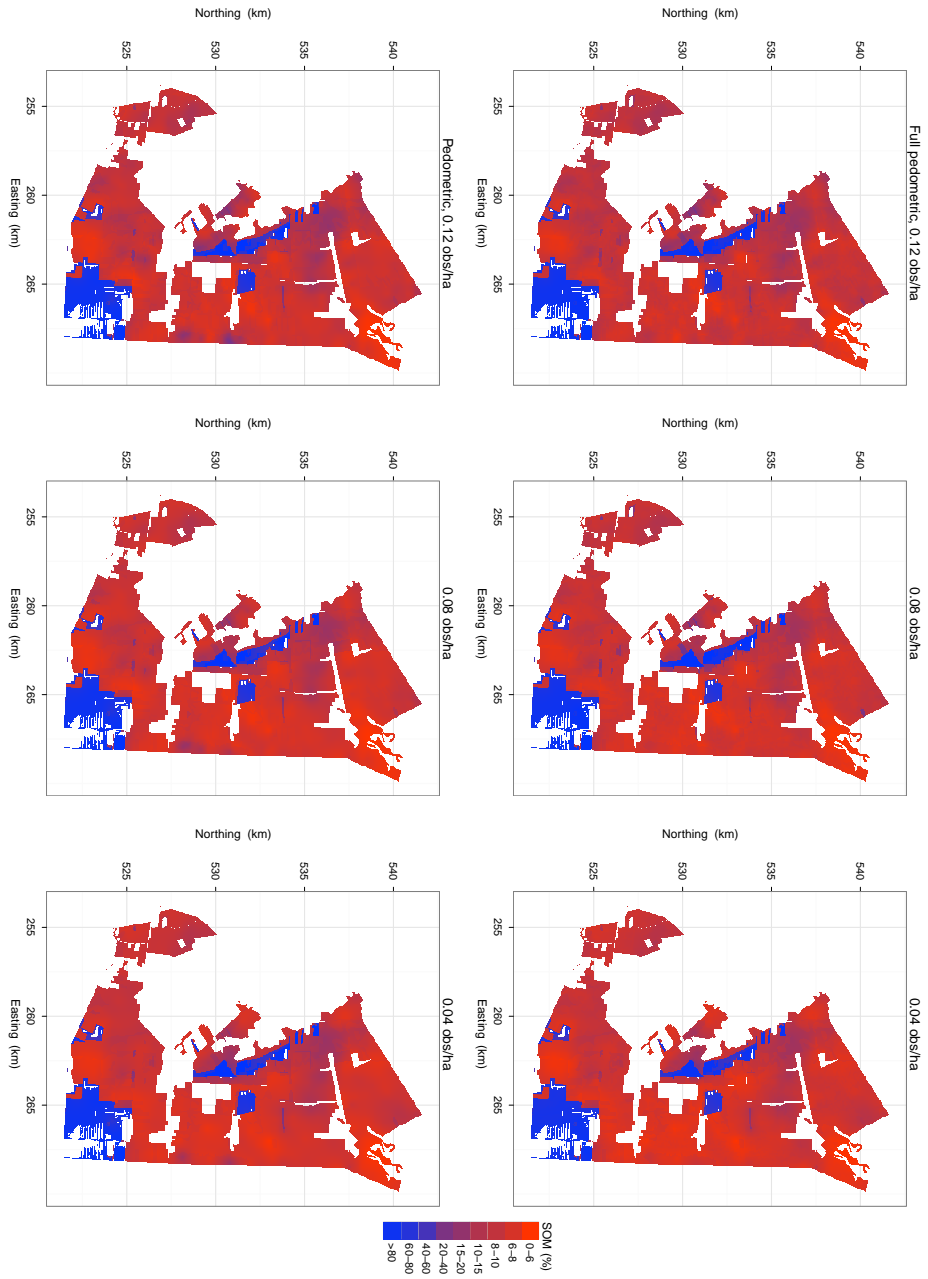


Figure 6.7: Predicted SOM content of the 0–20-cm soil layer for three different sampling intensities. The figures at the top depict the maps created with the full pedometric approach, those at the bottom with the conventional pedometric approach.

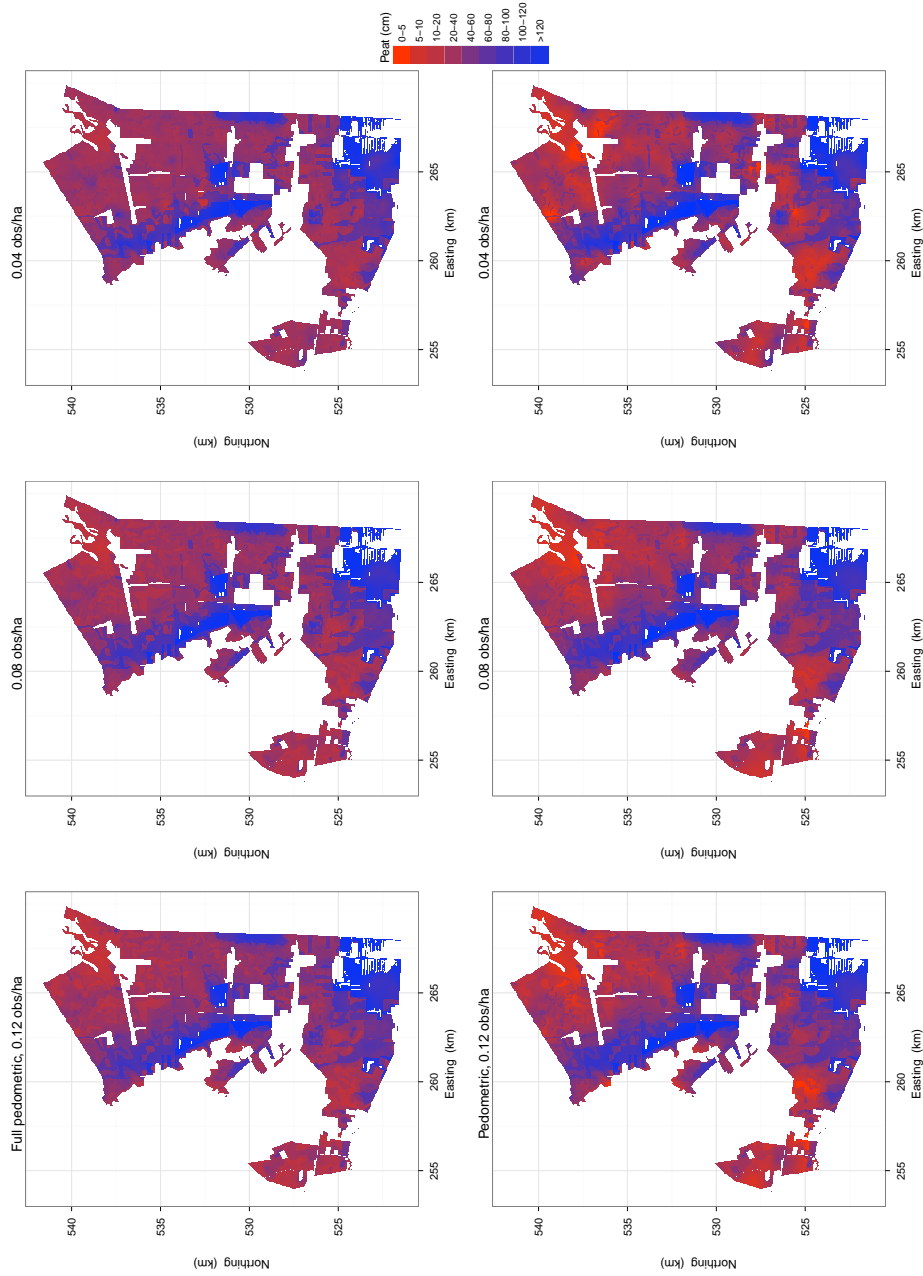


Figure 6.8: Predicted peat thickness for three different sampling intensities. The figures at the top depict the maps created with the full pedometric approach, those at the bottom with the conventional pedometric approach.

eral soils (*pZ* and *Hn*) dominating the northeast corner as well as some larger patches in the southwest. The maps mainly differ in the extent and distribution of the mineral soils. As the observations density decreases the mineral soils seem to become less well represented, especially on the maps based on densities 0.08 ha^{-1} and 0.04 ha^{-1} .

Sampling intensity hardly affected the geostatistical models for the soil properties (Table 6.4). The trend models are identical, except for the peat thickness model in the FPA with density 0.12 ha^{-1} , and the variance parameters have similar magnitude. The ranges are fairly constant and the spatial dependence parameter varies between 0.35–0.40. Only the variance parameters of the peat thickness models vary considerably between the sampling intensities. As peat layer thickness is a spatially much more heterogeneous property in the study area than SOM, it is not surprising that this effect first becomes apparent for this property.

Costs

Table 6.5 shows the cost of each component for the different mapping methods and sampling intensities. Total cost of the conventional soil type map is 156 k€ or 11.3 € ha^{-1} , which is 1.17 times or 1.6 € ha^{-1} more expensive than the digital soil map. As in CSM (e.g. Bie and Beckett, 1971; Bregt, 1992a; Finke, 2000), fieldwork is the largest cost component in DSM. Collection of field data makes up 86% of the total costs for conventional soil type mapping and between 75% and 88% of the total costs for digital soil type mapping, depending on sampling intensity. A reduction of the number of field days has therefore great effect on the total cost of soil mapping. The costs of digital soil type mapping decreased by a factor 2.8 from 9.7 € ha^{-1} for a density of 0.30 observations ha^{-1} to 3.4 € ha^{-1} for a density of 0.04 observations ha^{-1} . Compared with conventional soil type mapping, digital soil type mapping at a density of 0.04 observations ha^{-1} decreases total costs by a factor 3.3 for the study area.

It is worth noting that because of the fixed costs for geostatistical modelling, the cost-savings DSM compared to CSM increase with the size of the survey area. For example if the size of the area increases with a factor 22 (roughly the size of the area with peat soils in the Netherlands that requires updating) then the total cost of CSM would increase by a factor 22 as all cost components are a function of area, whereas in our example the total cost of DSM would increase by a factor 16 for a density of 0.04 observations ha^{-1} making DSM 4.5 times cheaper than as CSM (3.5 M€ vs 0.77 M€). However, if the size of the area decreases with a factor 10 then the cost of DSM for a density of 0.04 observations ha^{-1} would decrease by roughly a factor 5, making it more expensive than CSM. The question that now remains is how reducing mapping costs by reducing the number of sampling locations affects map accuracy.

Table 6.5: Costs in k€ of the soil mapping components. ST refers to soil type and SP to soil property.

Cost component	Sampling intensity							
	0.30 obs ha ⁻¹		0.12 obs ha ⁻¹		0.08 obs ha ⁻¹		0.04 obs ha ⁻¹	
	ST	SP	ST	SP	ST	SP	ST	SP
CONVENTIONAL SOIL MAPPING								
Fieldwork preparation	2.8	2.8						
Soil mapping	112	112						
Operational costs fieldwork	23	23						
Soil profile dataset compilation	1.5	1.5						
Mapping in GIS	17	17						
Application RPD/MUM methods		0.8						
Total Costs (k€)	156.3	157.1						
Costs (€ ha⁻¹)	11.3	11.4						
DIGITAL SOIL MAPPING								
Fieldwork preparation	2.8		2.8		2.8		2.8	
Environmental data collection	4		4		4		4	
Data collection	97.3		65.1		49.9		29.2	
Operational costs fieldwork	20		13.4		10.3		6	
Digitizing plotting observations	3							
Soil profile dataset compilation	1.5		1.5		0.8		0.4	
Preparation point dataset	0.8		0.8		0.8		0.8	
Geostatistical modelling	3.8		3.8		5.3/0.8		3.8	
Total Costs (k€)	133.2		91.4		72.4		47	
Costs (€ ha⁻¹)	9.7		6.6		5.2		3.4	

* Costs for FPA/CPA.

6.3.4 Mapping efficiency: effects of method and effort on accuracy

Soil type mapping

Table 6.6 contains the validation results of the soil type maps. Overall purity of the conventional map is 55.7% while that of the digital map is slightly lower (54.9%). The difference is not significant ($p = 0.430$). The overall purity of the original 1:50 000 soil map (Fig. 6.1) was estimated as only 27.4% and confirms the necessity to update. The large difference between the calibration (83.3%) and overall purity of the conventional map is striking. It is an effect of the preferential selection of observation locations during survey. The soil surveyors selected locations they judged representative and avoided for example field edges and small terrain features such as depressions. Occasionally sampling locations were discarded because the observed soil did not fit the soil surveyor's general picture of soil variation in the field under investigation. This shows the importance of validation with independent data collected by probability sampling (Brus et al., 2011).

The equal performance of the two mapping methods for soil type is rather surprising as beforehand was expected that CSM would outperform DSM given the 22% difference in calibration purity. Both methods mapped the general patterns of soil distribution but failed to capture the details. The reason for that is the complexity of soil spatial variation. Soil conditions can be very heterogeneous, even at short distances, as a result of reclamation and cultivation methods (see section 6.2.1). The landscape and soils are man-made and as a result soil-landscape relations are often disturbed or even absent. This makes it very challenging to map soils on the basis of point ob-

servations and landscape features, even with CSM at scale 1:50 000, and makes large impurities in the map units inevitable. For this reason I think that an overall purity between 55 and 60% is an adequate result, despite the fact that a purity of 70% is the generally accepted standard for the 1:50 000 soil map in the Netherlands. It will be hard to improve on these results at the scale level of the survey.

Reducing mapping effort in three steps from 9.7 € ha⁻¹ to 3.4 € ha⁻¹ did not have a negative effect on the accuracy of the digital soil type maps produced. On the contrary, the overall purities of the maps created with reduced sampling intensities were larger than the purity of the full dataset map: 6.7% ($p = 0.005$) for the 0.12 ha⁻¹ map, 3.6% ($p = 0.100$) for the 0.08 ha⁻¹ map, and 0.9% ($p = 0.381$) for the 0.04 ha⁻¹ map. Since the four prediction models had different trend models, autocorrelation structures and calibration data, it is hard to pinpoint the cause of these purity differences. More important is that a soil type map could be produced with DSM at 30% of the costs of CSM but with similar accuracy.

Table 6.6: Validation results for the soil type maps. The standard errors of the estimated purities are between brackets.

Sampling intensity (ha ⁻¹)	Conventional soil mapping		Digital soil mapping	
	Costs (€ ha ⁻¹)	Purity	Costs (€ ha ⁻¹)	Purity
0.30	11.3	55.7 (4.5)	9.7	54.9 (4.4)
0.12			6.6	61.6 (4.3)
0.08			5.2	58.5 (4.4)
0.04			3.4	55.8 (4.4)

Soil property mapping

The mean sub-sampling variance of SOM was 8.3%² and that of the peat layer thickness 27.3 cm². The estimated MSEs were corrected for this sub-sampling variance. Table 6.7 shows the validation results for the conventional and digital soil property maps.

Bias. The estimated global mean errors show that predictions of SOM with the conventional MUM-method and with both DSM methods are severely biased ($p < 0.000$). The source of the bias are the (hand-estimated) SOM contents in the point dataset, which were used to calibrate the geostatistical models and to estimate the map unit means. Measured SOM contents at the 125 validation locations were on average 3.7% larger than the hand-estimates at the same locations, which almost exactly equals the bias. As the field observations on SOM at the validation locations were consistent with those at the calibration locations, it was assumed that the calibration data were equally biased. Because the bias component makes up 10–15% of

the MSE reported for the DSM and CSM-MUM, it was decided to report not only the MSE in Table 6.7 but also the variance component of the MSE: the variance of the prediction error (VPE). The VPE is calculated by subtracting the squared ME from the MSE. This measure allows a more objective evaluation of the accuracy of the different prediction methods and of the effect of mapping effort on accuracy as the bias is not caused by the methods. Note that the reported RMSEs are computed from the MSEs. Furthermore, the MSEs are used for statistical comparison of the maps. There was no evidence for bias in the predicted peat thicknesses.

Conventional Soil Mapping. The conventional MUM method outperformed the RPD method for both SOM ($p = 0.120$) and peat ($p = 0.054$) based on MSE. For SOM the difference in prediction performance was mainly attributed to map unit *V* (MSE difference: 145, $p = 0.178$). MUM also outperformed RPD for the other map units but absolute MSE differences were small and not significant for map units *iV* ($p = 0.374$) and *iW* ($p = 0.219$). For peat thickness MUM outperformed RPD for all map units (each with $p < 0.060$), except map unit *V*. Predictions of SOM and peat thickness using RPDs and the old 1:50 000 soil map were also validated. Predictions were seriously biased (ME-SOM: 7.3%, ME-PEAT: 38 cm) and RMSEs were twice as large as those obtained with predictions with RPDs and the updated soil map (RMSE-SOM: 22%, RMSE-PEAT: 53 cm).

Digital Soil Mapping. For SOM the FPA results in smaller MSE than CPA for each of the three densities. Only for density 0.08 ha^{-1} the difference was found to be significant ($p = 0.048$). For peat thickness CPA outperformed FPA in terms of MSE but the differences were weakly significant only, with $p \approx 0.200$ for each case. These results show that soil property prediction with FPA did not benefit from the (much) better fitting trend models compared with CPA (Table 6.4). The merits of FPA to soil property mapping strongly depend on how well actual soil type is represented by the location-specific probability distributions. If prediction uncertainty of soil type is large, i.e. predicted probabilities are marginally different from each other, then the benefit of soil type-specific predictions of soil properties with a geostatistical model calibrated on observed soil type followed by weighted-averaging with soil type-specific probabilities as weights is lost. In the study area the uncertainty about the actual soil type is fairly large, especially in the agricultural areas dominated by mineral soils and shallow peat soils (Chapter 5), which might explain why FPA did not outperform CPA in all cases. Although the FPA did not result in more accurate maps of peat thickness, the method has some additional advantages compared with CPA such as a supposedly more realistic representation of the prediction error variance. Moreover, FPA allows to quantify the proportion of the prediction error variance that arises from uncertainty about the actual soil type at a prediction location (Chapter 3). It is questionable, however, whether these advantages are worth the extra investment of approximately 0.30 € ha^{-1} .

For SOM mapped with FPA, the effect of mapping effort on accuracy is not unambiguous. Reducing sampling intensity from 0.12 ha⁻¹ to 0.04 ha⁻¹ did not significantly affect the MSE ($p = 0.256$). The accuracy of the map based on the dataset with 0.08 observations ha⁻¹ is larger than the 0.12 ha⁻¹ map ($p = 0.083$) and the 0.04 ha⁻¹ map ($p = 0.096$). Closer examination of the validation data, however, reveals that the MSE difference between the 0.08 ha⁻¹ map and the other two maps can be attributed to a single validation location with an extreme error for the 0.12 ha⁻¹ and 0.04 ha⁻¹ maps. For SOM-CPA the 0.12 ha⁻¹ and 0.08 ha⁻¹ maps have similar accuracy, while the accuracy of the 0.04 ha⁻¹ map is smaller than that of the other maps ($p = 0.054$). For peat-FPA the accuracy steadily decreases as mapping effort decreases. The MSE of the 0.04 ha⁻¹ map is 16% smaller than that of the 0.12 ha⁻¹ ($p = 0.013$) and the RMSE is 8% smaller. The accuracy of the peat-CPA maps show a similar trend with $p = 0.054$ for the MSE difference between the 0.12 ha⁻¹ and 0.04 ha⁻¹ maps. Based on these results peat thickness mapping seems more affected by reduction of mapping effort than SOM mapping. This is not surprising since the spatial variation in peat thickness is much larger than the variation in SOM content, as illustrated by the validation sample. For the agricultural soils *Hn*, *iV*, *iW* and *pZ* the average SOM contents are reasonable similar: 10.0%, 13.5%, 12.4% and 9.8%. For unreclaimed peat soil *V* this is 81%. For peat thickness the soil-type averages are much more variable: 6 cm for *Hn*, 68 cm for *iV*, 27 cm for *iW*, 8 cm for *pZ* and 116 cm for *V*. In general a spatially more variable soil property requires more observations for geostatistical modelling than a less variable property to achieve a similar level of accuracy.

So far the effects of DSM method and mapping effort on map accuracy were evaluated on basis of statistical significance. I will now shortly evaluate the results based on relevance for practice. For SOM the difference in RMSE between the 0.12 ha⁻¹ and 0.04 ha⁻¹ maps is less than 0.5%, which is hardly relevant considering the average SOM content in the study area is 18.8% (estimated from the validation sample). The same is true for peat thickness. Here the difference in RMSE between the 0.12 ha⁻¹ and 0.04 ha⁻¹ maps only is 1.7 cm, while the average peat thickness is 39 cm. Doubling the mapping effort in terms of costs ha⁻¹ results in a small, statistically significant but not relevant increase in map accuracy and is certainly not worth the extra effort.

Conventional versus Digital Soil Mapping. I first compare conventional soil property maps with the digital maps created with the largest mapping effort. I focus here on the digital maps created with CPA because this method was more efficient than FPA. SOM predictions by the RPD approach were slightly better than those obtained by CPA, although the difference in MSE was not significant ($p = 0.341$). The MSE of the MUM map is less than half of that of the digital soil map. Despite being quite large, the difference in MSE is only moderately strong significant ($p = 0.113$). This is the result of the large variance of the squared prediction errors for DSM caused by two extremes. These extreme errors are located in anthropogenic artefacts resulting from

the agricultural reclamation of the former highmoor. One location is located on top of a small remnant of the former highmoor in the middle of agricultural fields. The other location is located on a small stretch of reclaimed land in the large highmoor nature reserve in the southeast of the study area. At these atypical locations soil conditions strongly differ from those in the immediate surroundings. If these artefacts are not captured by environmental covariates in DSM, then prediction errors can be very large. In CSM anthropogenic artefacts, such as two meter-high peat ridges in flat agricultural land, can be easily recognized by the soil surveyor and drawn on a soil map. Peat thickness predictions by DSM were substantially better than those by CSM. The MSE obtained by DSM was 24% lower than the MSE obtained by the MUM method ($p = 0.105$) and 32% lower than that of RPD method ($p = 0.033$). Location-specific predictions of peat thickness clearly are more accurate than predictions by a constant mean or representative value of a map unit. This can be explained substantial short-range spatial variation of peat thickness: the five map units of the conventional soil map explain only 51% of the total variance of peat thickness in the validation dataset (for SOM this is 69%). This type of variation is better modelled by a continuously varying surface (e.g. as obtained by kriging) than by a discrete model of spatial variation.

Reducing the effort in terms of costs ha^{-1} by 50% increases the MSE for peat thickness, but the MSE is still smaller than of those conventional methods, although the difference with the MUM method was not significant ($p = 0.304$). For SOM, reducing costs by 50% did not have effect on the performance of DSM compared with CSM. These results show that DSM is much more efficient in generating up-to-date soil property maps than CSM. DSM generates maps that are as good (SOM) or slightly better (peat thickness) than the conventional maps at less than one-third of the costs of CSM. For the study area this means that digital soil property mapping is 112 k€ cheaper than conventional soil property mapping.

The results presented here compare well with findings in studies from the 1980s and early 1990s (Van Kuilenburg et al., 1982; Bregt and Beemster, 1989; Leenhardt et al., 1994; Brus et al., 1996). In these studies relatively simple pedometric (kriging) methods—no use is made of environmental covariates—were compared with conventional methods. All these authors found no significant differences in map accuracy between predictions by (ordinary) kriging and predictions based on (large-scale, 1:10 000-1:50 000) conventional soil maps. This study also shows that a large-scale soil map can be as good a model of spatial variation than the geostatistical model. However, conventional soil maps come at a cost. If existing soil maps are in need of updating, such as in the Dutch cultivated peatlands, then DSM is a much more efficient choice than CSM.

Mean versus Median Squared Error. If the (squared) error distribution is (strongly) skewed then the mean might be a misleading measure of the 'average' error. For

such distribution the median error is a more robust statistic of the ‘average’ error. The squared error distributions of SOM and peat thickness are both strongly skewed, mainly resulting from reclamation artefacts in the landscape, as explained above. These artefacts are part of study area and thus validation locations from these artefacts should not be excluded from the validation sample. Another source of skew is the presence of two distinct sub-populations of errors with very different means and variances. One sub-population comprises the agricultural areas and the other the highmoor areas. The median squared errors are reported in Table 6.7. Striking are the large differences with the MSEs. Separate validation statistics were computed for the two sub-populations of errors (data not shown). Separate analysis for the two sub-populations, however, does not result in noteworthy differences with the conclusions regarding the effect of methods and effort on mapping efficiency presented earlier. The same is true for analysis on basis of the median squared error instead of the MSE.

Table 6.7: Validation results for the soil property maps with estimates of the mean error (ME), mean squared error (MSE), variance component of the MSE (VPE), Median is the median of the squared errors, root mean squared error (RMSE) and the coefficient of determination R^2 of fitting measured against predicted values. The standard errors of the estimates of the ME and MSE are between brackets.

	Sampling intensity							
	0.30 ha ⁻¹		0.12 ha ⁻¹		0.08 ha ⁻¹		0.04 ha ⁻¹	
	SOM	Peat	SOM	Peat	SOM	Peat	SOM	Peat
CONVENTIONAL SOIL MAPPING - RPD								
ME	-0.72 (0.91)	1.7 (2.5)						
MSE	109 (37.3)	697 (127)						
VPE	108	694						
Median	18.5	196						
RMSE	10.8	26.4						
R^2	0.76	0.62						
Costs (€ ha ⁻¹)		11.3						
CONVENTIONAL SOIL MAPPING - MUM								
ME	-3.15 (0.73)	2.5 (2.3)						
MSE	66.5 (15.6)	624 (119)						
VPE	56.6	618						
Median	21.2	131						
RMSE	7.5	24.9						
R^2	0.86	0.64						
Costs (€ ha ⁻¹)		11.3						
DIGITAL SOIL MAPPING - FPA								
ME			-3.78 (1.03)	2.16 (2.09)	-3.75 (0.95)	2.52 (2.15)	-4.68 (1.12)	2.46 (2.25)
MSE			134 (56)	505 (102)	125 (51)	534 (113)	142 (56)	585 (121)
VPE			119	500	111	527	120	579
Median			16.8	149	19.4	199	19.4	219
RMSE			11.6	22.5	11.7	23.1	11.9	24.2
R^2			0.75	0.68	0.77	0.66	0.73	0.63
Costs (€ ha ⁻¹)				6.7		5.4		3.5
DIGITAL SOIL MAPPING - CPA								
ME			-3.67 (0.99)	0.92 (2.03)	-3.68 (0.99)	0.73 (2.09)	-3.49 (1.02)	0.71 (2.18)
MSE			138 (58)	470 (86)	138 (57)	500 (98)	147 (63)	543 (104)
VPE			124	469	124	499	135	543
Median			22.1	121	23.0	159	19.4	168
RMSE			11.7	21.7	11.7	22.4	12.1	23.3
R^2			0.76	0.70	0.76	0.68	0.75	0.66
Costs (€ ha ⁻¹)				6.4		5.0		3.2

6.3.5 Relevance

The Dutch Ministry of Economics, Agriculture and Innovation has commissioned an update of the 1:50 000 soil map for the areas with peat soils. This area covers 365 000 ha. Resources available for the nationwide update are not sufficient to carry out the soil survey with conventional methods. More efficient methods, such as digital soil mapping, will have to be used or will have to be integrated with conventional soil mapping to accomplish the update with the given resources. Outside the Netherlands, integration of DSM methods within conventional soil survey is already becoming common practice (Bui, 2007; Howell et al., 2008; Moore et al., 2010). In the Netherlands however, DSM is in an exploratory phase but it has to move forward now to the operational phase. Insight in efficiency of DSM for updating soil type as well as soil property maps is essential to plan update strategies in the near future.

The presented results apply to peatlands that are under intensive cultivation and one must be careful with generalizing the results to other landscapes and other areas in the world. Yet, this study provided valuable insights for future (digital) soil mapping in general. It put DSM to the test and showed that, at least for updating, state-of-the-art DSM methods do not necessarily produce more accurate maps than CSM. Still, DSM is attractive for updating because it is much more efficient than CSM. Finally, cultivated peatlands are of global interest and are the focus of a great variety of research projects such as carbon stock inventory and monitoring (Beilman et al., 2008; Grønlund et al., 2008; Leifeld et al., 2011), carbon sequestration potential of soils (Freibauer et al., 2004) and estimation of greenhouse gas emissions and balances (Maljanen et al., 2010; Berglund and Berglund, 2010, 2011). Such studies require accurate and up-to-date information on soil conditions and their spatial distribution. It is up to the (pedometric) soil mapping community to provide these data.

6.4 Conclusions

The study presented in this chapter evaluated and compared the mapping efficiency of conventional and digital soil mapping for updating soil maps for a cultivated peatland in the Netherlands. Validation of the created soil maps with independent probability sample data generally showed little difference in prediction performance between the two methods. Differences in accuracy that were found to be statistically significant were often not relevant for practice.

For soil property mapping with conventional methods the ‘map unit means’ method gave better results than the ‘representative profile method’. For digital soil mapping the conventional pedometric approach (classical universal kriging) performed as well as the full pedometric approach (soil type-specific prediction of soil properties followed by probability-weighted averaging using probabilities derived from a

pedometric soil type map) for SOM content and better than the full pedometric approach for peat thickness. Updating the soil type map first and then using this map for spatial prediction of soil properties did not seem worth the extra mapping effort.

For SOM mapping CSM performed as well as DSM, while DSM performed slightly better for peat thickness. Although differences in accuracy were small, the DSM maps were produced much more efficiently. Costs per ha were a factor three to four smaller than for CSM. This shows that DSM exploits the available information much more efficient than CSM. The digital soil type map had similar accuracy as the conventional soil type map but was produced at less than one-third of its costs. These results were somewhat surprising and counter-intuitive. It was anticipated that for soil type mapping CSM would outperform DSM but that for soil property mapping DSM would produce better results given the assumed superiority of DSM in terms of accuracy and representation of spatial variation. Based on these results I conclude that it is challenging in a data-rich country such as the Netherlands where detailed and high-quality (yet outdated) soil information is available, to improve on maps made with conventional methods by using state-of-art methods for soil mapping. Nevertheless, DSM is still an attractive alternative to CSM for updating soil maps because of its much greater mapping efficiency.



Chapter 7

Discussion and conclusions

7.1 Introduction

In the Netherlands there is a growing demand for quantitative, up-to-date and reliable soil information to support policy-making for a wide variety of environmental and agro-economic issues. The general-purpose 1:50 000 soil map with nationwide coverage, originally designed for soil suitability analysis and central to the Dutch soil information system *BIS*, is the main source of soil information to date. Yet, this map increasingly fails to provide the type of information desired by the current generation soil data users. First, the national soil map is becoming outdated in areas with peat soils and so are soil property maps that are typically derived from this map. Second, most recent soil data application require (raster) maps with quantitative information on the spatial distribution of soil properties, preferably with quantified accuracy and at various spatial scale levels. In 2009 the Dutch national government commissioned a six-year research programme, *BIS2014*, with the aim to update and upgrade soil information in *BIS*. Digital soil mapping (DSM) might be very useful to resolve current shortcomings of soil spatial information in the Netherlands. However, the merits of using DSM methods for updating soil spatial information as well as the efficiency of these methods compared with conventional methods has not been investigated and evaluated yet. This thesis aimed to address these issues and assessed whether DSM can be a viable alternative to conventional soil mapping (CSM) for updating soil information in the Netherlands.

This chapter summarizes and discusses the main research findings of this thesis and their implications for the future of DSM for Dutch soil mapping in section 7.2. Section 7.3 takes a somewhat broader perspective and discusses implications of operational DSM for soil information systems, with special attention to *BIS*. Finally, the main conclusions are presented in section 7.4.

7.2 Updating soil information with digital soil mapping

It is estimated that 60% of the area mapped as peat soils on the national 1:50 000 soil map (or 365 000 ha) requires updating. The dire need for updating is exemplified by the low purity (27%) of the generalized 1:50 000 map of the study area in Chapters 5 and 6, and the substantial bias and large prediction errors in soil property maps derived from this map (section 6.3.4). This thesis therefore focussed on the development, exploration and evaluation of several DSM methods for updating soil type and soil property maps in the Netherlands. Subsection 7.2.1 summarizes the main research findings for each of the sub-objectives defined in section 1.4.1 and answers the research questions on a general level. In the three ensuing subsections, several aspects related to operational DSM that merit further attention are discussed. These are relevant environmental covariates for updating, the use of legacy data, and the

validation of soil maps. Finally, some further thought is given on operationalizing DSM for updating the 1:50 000 soil map based on the findings in this thesis (subsection 7.2.5).

7.2.1 Methodological contributions and research findings

Updating the 1:50 000 soil type map

Chapters 2 and 5 explore the use of respectively simple and state-of-the-art DSM methods for updating the 1:50 000 map. In Chapter 2, a framework is presented for soil type mapping with multinomial logit model (MLM). In this framework special attention is given to the important role of pedological knowledge in the process of selecting a DSM model. MLMs were used to re-map the soil within ten map units of the generalized 1:50 000 map. No additional fieldwork was done for data collection. Only legacy soil point data obtained from *BIS* were used to investigate the utility of these data for updating. Soil mapping not only focussed on the areas with peat soils but also on the mineral soils as it was hypothesised that DSM with high-resolution covariates and a large point dataset could also be used to increase the purity of the mineral map units by mapping out impurities. The latter can be regarded as upgrading of the soil map through disaggregation of the map units. The validation results (Table 2.7) showed that improvement in map purity (6%) compared with the existing generalized 1:50 000 map was modest and mainly attributed to better representation of soil distribution within the peat map units of this map. However, map unit purities and class representations for the four peat soils depicted on the updated map were still small, particularly those of the shallow peat soils. Disaggregation of mineral map units to increase detail did not seem to be a worthwhile effort based on the Drenthe case study. A reason for this can be the high level of detail already present in the existing 1:50 000 map. The median size of the map delineations in Drenthe is 21 ha, which is only 3.5 times larger than the minimum delineation size based on cartographic principles. The map units themselves reflect regional and local effects of topography and parent material. Disaggregating map units would therefore apply soil mapping on field-scale level, given the small size of the map delineations. At this scale level not only short-range variation in (micro)topography and parent material affect soil distribution but also anthropogenic factors such as levelling of microrelief, raising the soil surface to increase its bearing capacity, topsoil removal and deep cultivation, the effects of which cannot be properly captured by the biophysical covariates that were used for modelling.

The MLM ignores spatial dependence in the data, and might therefore be suboptimal for soil type mapping. Chapter 5 addresses this issue and investigates if a soil map generated by a spatial model is more accurate than one generated by a non-spatial model. This chapter introduces the generalized linear geostatistical model

(GLGM) and shows how this model can be used for soil type mapping. The GLGM is central to the methodological framework of model-based geostatistics (Diggle and Ribeiro Jr., 2007), which can be considered state-of-the-art in DSM (Lark et al., 2006). A pragmatic approach was adopted in which the five soil types in the study area were modelled separately with binomial logit-linear GLGMs. This has the advantage that for each soil type a different set of covariates can be used. In the MLM one set of covariates is used, which may result in sub-optimal models for the individual logit functions. Prediction with soil type-specific GLGMs resulted in five binomial probabilities at each prediction location, which were scaled to multinomial probabilities by dividing the binomial probabilities by their sum. Validation showed that for the cultivated peatlands use of the GLGM did not result in more accurate predictions than use of the non-spatial MLM. The implications of these results in the context of operational digital soil type mapping within the *BIS2014* project are further discussed in section 7.2.5.

Digital soil type mapping methods predict probability distributions of soil type at the nodes of a prediction grid. A soil map can then be derived from these distributions as well as measures that quantify the uncertainty associated with the predictions such as the Shannon entropy (Chapter 2) and the confusion index (Chapter 5). The former has the advantage that it takes the entire probability distribution into account, while the latter has a more intuitive interpretation. Furthermore, up-to-date digital soil type maps that represent soil type by probability distributions can be used to generate soil property maps (Chapters 3 and 4).

Updating soil property maps

It is beyond doubt that general-purpose soil type maps provide valuable information on the spatial distribution of soil properties and that these maps should be used, if available, for geostatistical modelling of soil properties (e.g. Goovaerts and Journel, 1995; Brus et al., 1996; Liu et al., 2006). First updating a soil type map with DSM and then using this map for soil properties mapping might therefore be advantageous. In addition, digital soil type maps represent soil type by probability distributions and this offers new possibilities for spatial prediction of soil properties that fully exploit the information provided by these maps. Availability of a full probability distribution of soil types at sampling locations means that instead of mapped soil type, observed soil type can be used as covariate. This has the advantage that the relationship between soil property and soil type is not confounded by impurities in the map units, which might result in more accurate predictions of the target property.

In Chapter 3, a model for spatial prediction of quantitative soil properties is developed that exploits the information provided by an updated digital soil type map. The model was applied to map the topsoil organic matter content, using the soil type map created in Chapter 2. Validation of the results showed that there was no

strong evidence for better prediction performance of the proposed model over the conventional kriging method. Similar results were found in Chapter 6; this time for two different soil properties and for trend models that included additional covariates to soil type. These results indicate that success of the method depends on how well the probability-based soil type map represents the actual soil type. If there is large uncertainty about the prevailing soil type at prediction locations then the advantage of prediction based on a strong, informative relationship between soil type and property (as was shown in Chapters 3 and 6) is undone by probability-weighted averaging of the soil type-specific predictions with marginally different probabilities. Despite little improvement in prediction performance, the model developed in Chapter 3 has some advantages related to quantification of prediction accuracy compared to the conventional kriging method that might favour its use (section 3.5).

Current need for quantitative, up-to-date information on the spatial distribution of soil properties is not only restricted to two dimensions. Data on variation of soil properties with depth is becoming increasingly important in agro-environmental modelling and soil resource assessment and monitoring (e.g. de Groot et al., 2005; Kros et al., 2011). An important objective of this thesis therefore was to develop a method for modelling the three-dimensional spatial distribution of soil properties. Chapter 4 presents such method and illustrates its use for the soil organic matter (SOM) content mapping. It makes use of the methodology developed in Chapter 3 and a digital soil type map (for the illustrative case study the map developed in Chapter 2 was used). The presented approach combines pedological knowledge with geostatistical modelling, which is an important difference with previous studies on the topic of three-dimensional soil mapping (Minasny et al., 2006; Malone et al., 2009; Mishra et al., 2009). The method predicts not a single predicted depth function at each prediction site but a probability distribution of soil-type specific depth functions. Modelling probability distributions of soil-type specific depth functions is conceptually appealing. The methodology is closely related to the conventional approach for three-dimensional modelling of soil variation. As in the conventional approach, a soil type map forms the starting point. For each of the soil types depicted on the map a model for the soil property depth profile is defined. In the conventional approach discontinuous (stepped) functions derived from representative soil profile descriptions are used as depth model. In the presented pedometric approach these functions are refined for each soil type individually based on characteristics of the depth profile of the soil property of interest. This refinement likely better represents the true depth profiles. Additional advantages compared to the conventional approach are described in section 4.4. However, despite these advantages SOM stocks predicted with conventional depth functions derived from representative profile descriptions were of similar accuracy as predictions by more sophisticated pedometric depth functions.

The proposed methodology for three-dimensional modelling of soil spatial variation is data-demanding but this is not likely to hamper application in other areas in the

Netherlands given the data-rich environment. Application might rather be limited by the number of soil types, or better, by the number of functional depth profiles that are required to model the vertical variation of a specific soil property—in Chapter 4, five functional depth profiles were defined to model the vertical variation of SOM in ten soil types (Table 4.3). Applying the presented method for soil mapping in an area with a great number of functional profiles might become tedious.

One aspect of three-dimensional mapping of depth functions that has not been considered in this thesis is the uncertainty associated with the predicted depth functions. Recently, Malone et al. (2011) proposed to use prediction intervals to quantify the accuracy of predicted continuous depth functions. Perhaps more interesting than the accuracy of the predicted functions itself, is to quantify the effect of the uncertainty associated with the predicted function parameters on properties derived from the functions (e.g. the SOM stock). This can be done using Monte Carlo stochastic simulation of the depth function parameters. Another issue regarding digital soil property mapping that has not been addressed in this thesis but that is worth further research, is DSM of soil property dynamics. Soil point data in *BIS* are mainly used for generating ‘static’ soil property maps, i.e. the spatial distribution of soil properties when the data were collected. Environmental policy questions, however, increasingly focus soil dynamics, i.e. on quantifying and mapping temporal change of soil. For instance change in SOM content (Reijneveld et al., 2009; Hanegraaf et al., 2009), peat thickness, phosphate saturation degree, or soil subsidence rate (Hoogland et al., 2011). Future soil mapping activities will likely shift more and more from ‘static’ soil mapping to soil monitoring and mapping soil change. This requires methods for spatio-temporal DSM, of which Webster and Heuvelink (2006), Heuvelink and Griffith (2010) and Minasny et al. (2011) provide recent examples.

Accuracy and efficiency of digital and conventional soil mapping for updating

DSM methods are nowadays widely applied to produce soil maps (e.g. Boettinger et al., 2010). Despite these efforts little is known about the efficiency of DSM methods in terms of cost per ha mapped, the relationship between costs and map accuracy and its relation to the efficiency of CSM methods. McBratney et al. (2003) and more recently Grunwald (2009) identified the economics of DSM as a relevant topic for further research since DSM is gradually becoming operational. Issues on accuracy of DSM products and DSM efficiency become also increasingly important to soil survey in the Netherlands. Whether DSM can be an alternative to CSM for updating is partly determined by the efficiency of DSM compared to that of CSM, which was investigated in Chapter 6.

Accuracy assessment of soil maps is an important topic in this thesis. Soil maps are of little value if their accuracy is unknown. Validation of soil maps should therefore be an integral part of any DSM procedure. Nevertheless, Grunwald (2009) found in

a DSM review study that more than one third of ninety recent DSM studies did use any performance test. Information on map accuracy is also imperative for evaluation and comparison of different soil mapping methods, which is key to each chapter in this thesis. Each of the digital soil maps created in this thesis was validated with independent data. The importance of using independent data for validation was illustrated by large differences between calibration purity and actual purity in Chapters 2 and 5. Comparing predictions with observations at calibration sites generally over-estimates the actual purity. Also the theoretical purity does not always provide a good measure of actual purity as is shown in Chapter 2 and for example by Brus et al. (2008). The validation datasets used in this thesis were all collected by a design-based sampling strategy involving probability sampling and design-based estimation of accuracy measures, which is preferred over other methods because estimators of map quality and associated standard errors are unbiased and model-free (Brus et al., 2011). An important aspect when designing the sampling strategy is the support of the validation observations; these should be equal to the support of the predictions (Chapter 6). Map quality measures used for soil type maps were overall purity, map unit purity and class representation. Those for soil property maps were the mean error (ME), which is a measure for bias; and the (root) mean squared error ((R)MSE), which is a measure for accuracy.

Chapter 6 investigates the efficiency of DSM with respect to CSM for updating soil type and property maps. The digital soil type map created for a small part of the cultivated peatlands (Chapter 5) had similar accuracy as the conventional soil map (Table 6.6) but mapping costs per ha were a factor 1.2 lower (Table 6.5). Further analysis on the effect of mapping effort on accuracy of digital soil maps revealed that a soil type map produced with roughly one-seventh of the observations used in conventional mapping, had similar accuracy as the conventional map, while costs per ha were a factor 3.3 lower. This shows that DSM much more efficiently exploits the information in the data than CSM. The purities of the digital soil type maps in Chapters 5 and 6 roughly varied between 55% and 60%. These results may seem modest at first, but comparison with the purity of the updated conventional map (55%) shows that these purities are roughly what can be realistically expected in the cultivated peatlands given the invested effort. The highly heterogeneous nature of soil spatial variation in this landscape and disturbed relationships between soil and landscape features complicates soil mapping. In this respect, map purities reported in Chapters 5 and 6 are quite adequate. Furthermore, the validation results in Chapters 2, 5 and 6 showed that the prediction models were able to capture the general patterns of soil spatial variation. The models differentiated peat and mineral soils very well but had difficulty in predicting the type of peat soil, which is determined by thickness of the peat layer and topsoil lithology.

Evaluation of accuracy and efficiency for soil property mapping gave results similar to those for soil type mapping. The difference in accuracy between the two methods is small for both SOM content and peat thickness. Differences that were

found, were generally not relevant for practical purposes. Updating soil property maps with DSM, however, was roughly three times more efficient than with CSM. The results from Chapter 6 also showed that 1:50 000 soil map is a very adequate model for predicting the spatial distribution of soil properties—assuming the map is up-to-date—which can be attributed to the high level of detail in the map in combination with the generally strong association between (basic) soil properties and soil type. The predictive capability of the map is exemplified by an example from the study area in Chapter 6. Here the updated conventional soil map explains 90% of the variation in SOM content in the validation dataset and 73% of the variation in peat thickness.

Concluding, validation results in Chapters 4 and 6 show that it is very challenging in a data-rich country such as the Netherlands, where detailed and high-quality soil information are available, to improve on predictions made with conventional methods by using state-of-art DSM methods. The fact that soils and landscapes in the Netherlands are intensively managed (disturbed) contributes to this challenge. Nevertheless, the results of Chapter 6 indicate that DSM can be a cost-effective alternative to conventional methods for updating without compromising map accuracy. This makes it an attractive alternative to CSM. Mapping with DSM methods is three times more efficient as mapping with CSM methods. These findings are encouraging for future update activities. It should be kept in mind, however, that these findings apply to the cultivated peatlands. One should be careful with generalizing the results to other soilscapes and even to other types peatlands, such as for example the fen meadow landscape in the western and northern part of the Netherlands.

7.2.2 Covariates for digital soil mapping

Throughout this thesis a great variety of (biophysical) environmental covariates were used as predictors in the DSM models. The strongest covariates for DSM proved to be, not surprisingly, covariates derived from the existing 1:50 000 soil map. Despite the fact that the map is outdated, it still is an extremely valuable source of ancillary information for DSM. The availability of this map sets DSM in the Netherlands apart from most DSM in other areas around the world, where often only very coarse, small-scale soil maps are available (e.g. Bui and Moran, 2003; Henderson et al., 2005), if at all. The results and conclusions presented in this thesis should therefore be interpreted in the context of DSM where a large-scale general-purpose soil type map is available. The importance of land cover history for refining soil information from the 1:50 000 soil map was shown by Sonneveld et al. (2002) and Schulp and Veldkamp (2008). Also in this thesis land cover (history) covariates proved to be valuable predictors. Relevant, but less important covariates were derived from ground-water class maps, landform maps and a DEM. In a relatively flat country such as the Netherlands, DEM derivatives are generally of less importance for spatial prediction than elsewhere (e.g. McKenzie and Ryan, 1999; Chaplot et al., 2001; Florinsky et al.,

2002). This is another important difference between DSM in the Netherlands and other countries.

Strong human impact on soil formation and spatial distribution complicates DSM in the Netherlands in general and in the cultivated peatlands in particular, as discussed in Chapters 5 and 6. Currently the only available information on mechanical disturbance of the original soil profile is stored in the attribute table of the soil map. This attribute was used as a covariate in one of the prediction models in Chapter 4, but is neither up-to-date nor complete. An ongoing project aims to collect all available spatial data on soil disturbances from the various agencies that hold these data. Once collected, the data will be compiled into a GIS-layer and stored in *BIS*. This layer might be of great value for DSM in the human-dominated landscapes of the Netherlands.

7.2.3 Legacy soil data for digital soil mapping

Legacy soil data are existing data, stored in soil databases. These data comprise soil maps but also soil point data. The latter were often collected during conventional soil surveys, typically without a clear sampling design, and are of different quality and reliability (Carré et al., 2007; Finke, 2011). The Netherlands is a data-rich country with a wealth of legacy soil data: over 300 000 soil profile descriptions and a nationwide 1:50 000 soil map are stored in *BIS*. The relevance of the soil map for DSM was highlighted in previous sections; point data contain a wealth of information as well. These data, however, are not perfect: point data can be outdated and unevenly distributed over geographical space, locations recorded before the introduction of hand-held GPS devices may have large positional errors, profile descriptions may have missing data, recorded soil property values have different degrees of accuracy as a result of different measurement methods. Legacy point data were used for DSM in Chapters 2, 3 and 4 without paying much attention to its inherent imperfections. Point data were only screened for age using an arbitrary threshold. However, quality assessment of legacy data to ensure proper use of these data for DSM becomes ever more important now that DSM becomes operational for soil mapping across the globe, with many countries having extensive soil resource databases filled with legacy data (Rossiter, 2004). This requires development of procedures for efficient updating of outdated point information, for combining data from different sources and with different uncertainties and for assessing the effect of positional error on the accuracy of the predictions.

Most profile descriptions of peat soils in *BIS* are one or more decades old. The current peat layer thickness is likely to be smaller than the recorded thickness. Outdated profile descriptions should not be used uncritically for DSM. These should either be discarded from the legacy dataset (like in this thesis) or updated. Updating is preferred over discarding because it makes the most out of existing data, but is only

attractive when updating is more efficient than collecting new data. Updating can be done by revisiting a selected set of sampling locations. New soil profile descriptions from these locations can then be used to derive an empirical model of the rate of decline of peat thickness. This model can then be used to update additional outdated point observations. Issues that must be dealt with are censored observations (the initial thickness of the peat layer is unknown) and absence of peat at the time of revisiting (the decline rate cannot be estimated). Updating point observations introduces additional uncertainty in the recorded soil properties which should be accounted for when using the updated observations for spatial interpolation. Grimm and Behrens (2010) and Nelson et al. (2011) propose methods to study the effect of positional error on inferred relationships between soil and covariates and its contribution to prediction uncertainty, while Walvoort et al. (2011) are currently developing a method for combining legacy soil data from different sources and with different uncertainties for soil mapping.

7.2.4 Validation of soil maps

The overall purity of soil type maps is a general and commonly used accuracy measure. It signifies the proportion of the map in which the mapped soil type equals the actual soil type. It is, however, a very strict assessment in which every error in the validation set is given equal weight: confusing two mineral soils receives the same penalty as confusing a mineral soil with a peat soil, which are taxonomically farther apart. It can therefore be questioned if the overall purity is the most sensible accuracy measure. Furthermore, soil type itself is almost never of direct interest in agro-environmental research. Users want interpreted information: soil type maps are (still) used to derive other data layers such as soil suitability maps or output of dynamic models that use the soil type map as input. Validation should therefore ideally assess the effect of a wrong classification on these end products. For example, confusing soil type *Hn* with *pZ* or *iW* with *iV* in the soil maps in Fig. 5.3 does not affect the accuracy of a topsoil organic matter map derived from these maps since these soils have topsoils with similar properties. The '*functional purity*' of the map for topsoil organic matter mapping is 77.3% (section 5.4.3; its overall purity is 54.9%). Functional purities are more informative and perhaps more relevant to practitioners such as modellers and policy-makers than the overall purity that is traditionally reported. I therefore recommend, as long as soil type maps are created and used, to report functional purities for soil properties most relevant for policy making, in addition to the overall purity.

The ME and (R)MSE are commonly reported and generally accepted accuracy measures for quantitative soil maps. However, these are summary measures and in some cases evaluation of map accuracy only on these measures might fall short. For example, validation of the soil property maps in Chapter 6 shows the limitations of using the (R)MSE as a single measure of accuracy when the error distribution is (strongly)

skewed. In such case the median squared error is a more robust statistic of the 'average' squared error. It is therefore recommended to express the quality of quantitative soil maps by the cumulative distribution function (CDF) of the (squared) error (Brus et al., 2011). The CDF is more informative than a single statistic derived from this function. From the CDF several parameters can be computed such (e.g. the mean) or a percentile (e.g. the median).

7.2.5 Operational DSM for updating soil type maps

One of the objectives of the *BIS2014* programme is a nationwide update of the 1:50 000 soil map for the areas with peat soils. The update project started in 2009 in Drenthe using the CSM approach. By 2011 roughly 22 000 ha was completed (of which Fig. 6.4 shows 16 800 ha). In spring 2011 the national government, who funds the programme, decided that progress was too slow and mapping too expensive. They aimed at a nationwide update to be completed by 2014, with the first part—the northern fen peat landscape—to be delivered early 2012. Resources made available are not sufficient to carry out the soil survey with conventional methods. More efficient (DSM) methods shall have to be used or integrated with conventional methods to accomplish the update for the given budget and within the given time frame. The studies in Chapters 2, 5 and 6 were of an explorative nature and aimed to gain insight in the efficiency of different DSM methods in terms of accuracy and costs for updating the 1:50 000 soil map. The experiences obtained from this thesis can guide further development of DSM methods for updating and the development of a methodological framework for integrating DSM with conventional methods for large-scale (1:10 000) mapping. Such maps are still produced on regular basis for rural land restructuring projects. Based on the findings in this thesis the following observations can be made on the utility of DSM for updating the 1:50 000 soil type map.

The MLM is preferred to be used for predictive mapping of soil type for two reasons. First, the MLM is a much simpler model and the benefit of GLGM for map updating in the peatlands was not proven. Second, use of the GLGM for soil mapping is still in experimental stage with several methodological issues left to be addressed (section 5.5). It is premature to employ the GLGM for nationwide updating.

In Chapter 2, two serious drawbacks of MLM were identified that might hamper its use for operational soil type mapping. The first drawback is that certain structures in the data can cause numerical problems when fitting the model in presence of categorical covariates (see section 2.4.1). This limits the number of soil types that can be modelled, and hampers the use of potentially strong covariates, as was frequently encountered when fitting the MLMs in Chapter 2. The second drawback concerns the number of soil types that can be handled practically—apart from the fact that a large number of them can cause numerical problems. Nationwide, the 1:50 000

map legend discerns 57 peat map units. Added to this number are several types of mineral soils that are now found in the peatlands at locations where peat has disappeared. Not all these soils will occur in one peat landscape but even for a relatively small area the number of map units can potentially be large. For example, the updated conventional soil map of the cultivated peatlands for the study area in Chapters 5 and 6 already has 25 map units (18 peat and 7 mineral). The first drawback of MLM can be solved by clustering soil types and/or factor levels of the covariates, or by discarding covariates from the model. This can limit, however, utility and predictive capability of the MLM when relevant soil types cannot be distinguished by the model or when relevant covariates cannot be used. The second drawback can be solved by focussing on predictive mapping of generalized soil types that are differentiated on the basis of key diagnostic soil properties that require updating, such as thickness of the peat layer. In peatland areas prediction of three soil types might then suffice: mineral soils (absence of peat), shallow peat soils and deep peat soils. The full 1:50 000 legend can then be reconstructed by refining the predicted soil types with information obtained from the original soil map such as texture class, topsoil lithology and peat type, under the assumption that these diagnostic properties have not changed over time.

An entirely different approach to updating soil type maps through DSM is updating by mapping key diagnostic (continuous) soil properties. The soil types of the 1:50 000 legend are defined by a set of measurable soil properties. Some of these properties are static and existing soil maps are satisfactory, while others are dynamic and existing maps have become outdated. The idea is to map only those diagnostic properties that are subject to change, such as the thickness of the peat layer, and next construct a soil type map from the soil property maps. Spatial prediction of continuous soil properties is more straightforward than that of categorical properties and does not share many of its drawbacks. Furthermore, Chapter 6 shows that mapping soil properties in the (cultivated) peatlands gave fairly reasonable results. I will now provide some thoughts on how map updating via DSM of diagnostic soil properties could be implemented.

Key diagnostic properties to classify soils in the peatlands are starting depth and thickness of the peat layer. In some peat landscapes, such as the fen meadow landscape in the western part of the Netherlands, topsoil organic matter content can be added to the list, but here I will only consider starting depth and thickness of the peat layer. After fitting a geostatistical model for these two properties, Monte Carlo stochastic simulation can be used to generate a sufficiently large number of realizations of these two properties from recent legacy point data in *BIS* and a suite of environmental covariates. From each simulated combination of starting depth and thickness of the peat layer, the main soil type—mineral soil, shallow peat soil, deep peat soil—can be determined through a set of decision rules based on the soil classification system. The series of simulations thus yields as many simulated soil types at each prediction location from which a probability distribution can be derived. The

soil type according to the 1:50 000 legend can then be reconstructed by augmenting the predictions with information on static properties obtained from the original soil map. Recent soil legacy point data from *BIS* can be used for mapping key diagnostic properties. In areas with sparse or outdated point data or in areas with large prediction uncertainty based on some preliminary mapping, additional data can be collected. For this use can be made of techniques to optimize sampling patterns such as proposed by Brus and Heuvelink (2007a) and Walvoort et al. (2010). Fig. 7.1 gives a schematic representation of the methodological framework for updating soil type maps via mapping key diagnostic soil properties. I am aware that this framework is not complete and that there are many details that need to be worked out before it can be made operational. This framework should therefore be considered as a guideline for further research on this topic.

It should be noted that the proposed framework deviates from the general approach followed in this thesis: soil property mapping using an updated soil type map as covariate. Nevertheless, I think that the framework merits further exploration and evaluation, which should include a comparison with the approach presented in this thesis.

7.3 Implications of operational digital soil mapping for soil information systems

Soil survey in the Netherlands is at a crossroads. The 1:50 000 soil map was completed almost twenty years ago. Since that time, methods for collecting and processing soil information have not changed much. Soil survey kept its qualitative character until to date. This is now inevitably going to change. Soil information stored in *BIS* must be updated, which requires re-mapping 365 000 ha of peatlands. Limited funds and capacity makes updating of such an extensive area with conventional methods unfeasible. In addition, a strongly increasing demand for soil property maps with quantified accuracy for environmental risk and uncertainty analyses requires development and formalization of systematic methodologies that can generate such soil information from data stored in *BIS*. This makes that DSM is moving forward to the operational phase, not only in the Netherlands but across the world (e.g. Sanchez et al., 2009; Boettinger et al., 2010).

Operationalizing DSM does not only influence the way soil maps are generated. It has implications for each step in the soil survey process (Bregt, 1992b) and requires rethinking the role of soil information systems (SIS). *BIS*, like many other contemporary SIS (Rossiter, 2004), is based on 20th-century concepts of soil survey and is not designed to adequately provide soil information necessary to address current environmental issues. This means that effective and successful operational DSM not

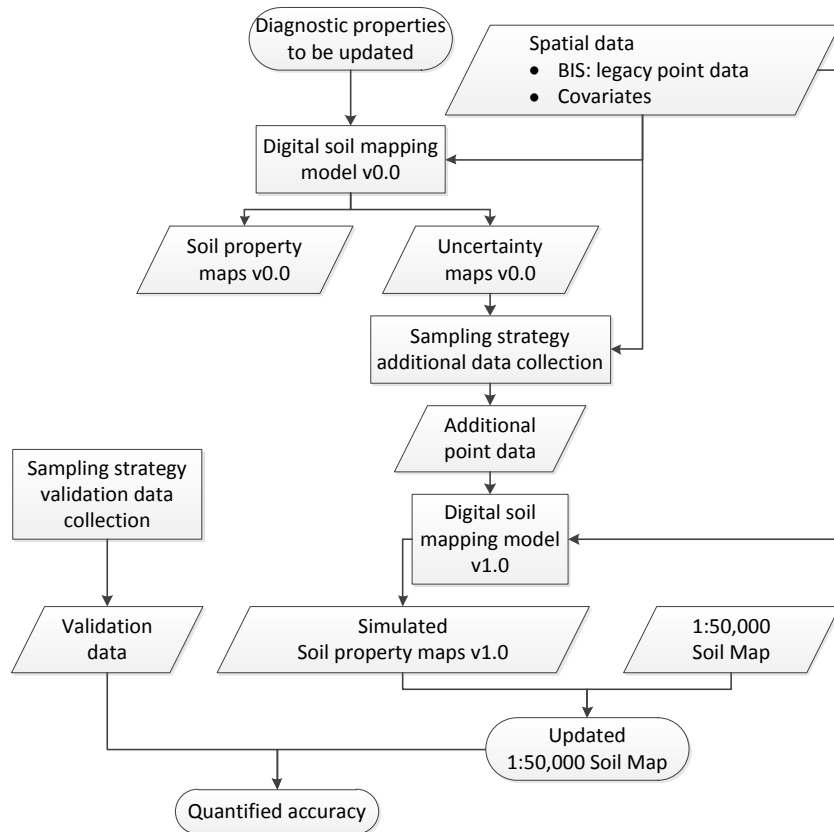


Figure 7.1: Proposed framework for updating the 1:50 000 soil map via mapping of key diagnostic soil properties

only requires a set of prediction models but also requires new methods for collecting, storing, processing, visualizing and disseminating soil information. In other words, a next-generation *BIS* is required based on a new paradigm of quantitative soil survey that better fits the new role of soil in today's environmental research. In such system, the (point) data and the methods to generate soil information from the data are central instead of soil classification and soil type maps. Furthermore, DSM introduces new mapping concepts unfamiliar to many soil data users. This poses additional challenges to successful adoption of DSM by practitioners. In this section I will give these issues some further thought and give my view on how DSM can be used and further developed to aid the development and implementation of the

next-generation *BIS*. Although this section takes a somewhat Dutch perspective, the issues discussed here apply to *SIS* in general.

Collection of soil data

Conventional methods for collecting soil data (section 1.2.1) are generally sub-optimal for DSM. In CSM sampling sites are selected on the basis of representativeness for a specific field or map unit. DSM makes other demands on sampling designs. Sampling locations should ideally cover covariate feature space as well as geographical space (e.g. Hengl et al., 2003; Minasny and McBratney, 2006a; Brus and Heuvelink, 2007a). Ideally, the locations of legacy point data should be taken into account when designing sampling strategies for DSM (e.g. Carré et al., 2007; Walvoort et al., 2010). In addition, spatial sampling can be optimized for estimation of the spatial dependence structure, required for geostatistical modelling (e.g. Lark, 2002; Zhu and Stein, 2006; Diggle and Ribeiro Jr., 2007). Demand for quantitative information on map accuracy requires design of efficient sampling strategies for validation (Brus et al., 2011), which typically differ from sampling strategies for calibration of DSM models. This means that besides DSM models, a toolbox is needed with efficient sampling designs and protocols for collection of field data for calibration of DSM models on the one hand, and validation of DSM products on the other hand. Development of efficient sampling strategies does not only apply to sampling in space but also to sampling in time to quantify and map soil property dynamics. This requires design of soil monitoring networks and development of methodologies to infer spatio-temporal trends from monitoring data (e.g. Brus and de Gruijter, 2011).

Efficient methods for data collection also include remote and proximal sensing techniques. Soil sensing can be a cost-effective method to populate soil databases with quantitative data on soil properties that are otherwise expensive to measure. It gained wide-spread popularity the last decade (e.g. Ben-Dor et al., 2002; Gomez et al., 2008; Lagacherie et al., 2008; Viscarra Rossel and Chen, 2011). Yet, remote and proximal sensing for soil data collection and mapping is still in its infancy in the Netherlands. van der Klooster et al. (2011) provide the first example from the Netherlands with some promising results. Remote sensing of soil in temperate areas is hampered by presence of vegetation. New methods for filtering the influence of vegetation from the spectral signal (Bartholomeus et al., 2011) might improve applicability of remote sensing in these areas. In-situ sensing technology (Viscarra Rossel and Walter, 2004; Viscarra Rossel et al., 2009) may facilitate more efficient and more accurate data collection in the field.

Operational DSM does not mean that the soil surveyor becomes redundant. Experienced, qualified field pedologists remain indispensable in the DSM-age, although their role in the soil survey process will change. The focus will be more on data collection—filling *BIS* with high-quality, up-to-date field observations, collected ac-

ording a statistical sampling design—and less on mapping in the field. Soil surveyors should also be trained in DSM so that they can effectively employ their extensive knowledge on soil forming processes and soil spatial variation to develop and improve DSM models (e.g. Lark et al., 2007; Orton et al., 2011), and to critically analyze the maps that these models generate.

Storage and processing of soil data

Current soil information systems (SIS), including *BIS*, store information about the soil as point observations and maps. DSM allows new, more efficient ways of storing soil information. A next-generation SIS stores soil information as pedometric models instead of maps. Maps are generated at the moment a request is submitted by a user. A SIS that stores models instead of maps has several important advantages compared to a conventional SIS (Heuvelink et al., 2010). First, it allows on the fly production of soil maps with quantified accuracy, customized to the requirements submitted by the user. Second, it enables easy updating of existing maps when new data come available. Third, it automatically documents how a map was made, including which point data were used and which covariates. Fourth, storing models instead of maps saves storage capacity. Currently a prototype SIS that stores models instead of maps is being developed for the Netherlands (Brus and Heuvelink, 2007b; Brus et al., 2009; Heuvelink et al., 2010). The procedure comprises six steps: (i) importing point observations from *BIS*, (ii) processing of profile descriptions (derive soil properties for user-defined depth interval), (iii) exploratory data analysis, (iv) building of a model of spatial variation, (v) geostatistical prediction or simulation and (vi) exporting resulting map(s). This prototype system will be further developed in the near future.

Visualization and dissemination of soil information

Open-source software for statistical computing such as R (R Development Core Team, 2008) and online mapping applications such as Google Earth offer new opportunities to access SIS and visualize digital soil survey data. The Australian soil information system ASRIS¹ already allows a large set of data layers to be viewed through Google Earth. A more advanced web-mapping application was recently developed in the US. SoilWeb² is an online soil survey browser with Google Earth and Google Maps interfaces (Beaudette and O’Geen, 2009). Google Earth is also used in the *Global-SoilMap.net* project to visualize and disseminate soil spatial information (Hengl et al., 2011). Compared to these examples of state-of-the-art techniques for visualization and dissemination of soil databases, *BIS* is a rather archaic system. As important

¹www.asris.csiro.au

²casoilresource.lawr.ucdavis.edu/map

as the development of sophisticated DSM toolboxes are proper and efficient tools to deliver the data to end-users. This is not only true for new DSM products but also for existing soil data. Google Earth and Google Maps are web-mapping applications many people are familiar with, which make them excellent platforms to circulate and deliver soil data in digital format to the public. The possibilities that these applications offer should be fully exploited to transform *BIS* into a modern, 21st-century *SIS*.

DSM and the next-generation *SIS* do not only generate soil maps, but also maps with the associated prediction accuracy. Research on visualization methods for soil data should therefore also focus on methods for presentation and communication of accuracy information (e.g. Hengl et al., 2004; Hengl and Toomanian, 2006). Hereby it is important that this information is presented in a way that is comprehensible to end-users and that shows its practical relevance, for example through data-worth analysis (Knotters et al., 2010b).

Adoption of new mapping concepts

Conventional soil maps are based on the polygon data model and are made at a specific scale. In the digital-era 'map scale' loses its meaning: in a GIS maps can be depicted at any scale. Furthermore, the polygon data model is a somewhat outdated model for representing soil spatial information. DSM introduces new mapping concepts. Digital soil maps are typically (but not necessarily) raster-based maps with a given resolution, extent and support (McBratney, 1998). Extent is the areal expanse over which predictions are made. Resolution is the pixel size with which the digital soil map is displayed with predictions made at the nodes. Support is the volume or area on which observations and predictions are made (Webster and Oliver, 2007). Digital soil maps are typically on point support, where the value of the pixel represents the value at the centre point within that cell (e.g. Chapters 3 and 4), or on block support, where the value of the pixel represents the average value within that cell (Chapter 6). Extent, resolution and support are independent concepts that can be chosen in any combination. Note that extent, resolution and support also apply to the depth and time dimensions of soil mapping.

Many soil data practitioners, especially in the non-modelling community, are used to working with conventional soil maps. They are not familiar with new mapping concepts that are introduced by the 'new-generation' digital soil maps and might therefore have difficulty in comprehending and accepting these. In addition, the advantages of DSM compared to CSM, particularly its efficiency and capability to provide estimates of prediction accuracy, and their practical relevance for agro-environmental policy making are often insufficiently clear to practitioners. Herein lies perhaps the greatest (and most important) challenge for the (digital) soil mapping community in the near future: to take away potential scepticism among prac-

tioners and to convince them of the merits of DSM. Widespread adoption of operational DSM as a credible and valuable alternative to CSM and the development of next-generation SIS is only possible with broad support among practitioners from various disciplines. Convincing them requires careful communication, for example through organizing user symposia, through publications in professional journals (e.g. Mol et al., 2005; Knotters et al., 2010a) and with practical examples on how DSM products can support and benefit the decision-making process (e.g. Knotters et al., 2010b). In addition, DSM should become an integral part of soil science curricula taught at higher and academic education institutions, as the current and coming generation of students are the soil scientists and practitioners of the future.

7.4 Conclusions

Digital soil mapping has come a long way since the introduction of geostatistics in soil science in the early 1980s. The 1990s saw fast-paced methodological developments in pedometrics, which was assisted by the rapid increase in computation power and information technology. These developments helped DSM mature in the first decade of the 21st century. Currently, DSM is moving towards the operational phase in which maps are produced on a routine and verifiable basis for regions, countries and continents. Operationalizing DSM requires rethinking the function of soil information systems and their role in today's environmental-centered research approach. Traditionally these systems store only soil data, but operational DSM requires that their functionality is extended with toolboxes for collecting and visualizing soil data and for transforming soil data to soil information.

In the Netherlands, lack of structural funding has delayed maintenance and extension of *BIS* and slowed strategic research on soil survey methods for nearly one decade. Recently, pedometric research regained momentum, initiated by the *BIS2014* research programme. This programme demands a thorough update of soil data stored in *BIS* and is a first step towards operational DSM and the next-generation *BIS*. This thesis contributes to the development of methods for digital soil type and property mapping in general and to the necessary evaluation of these methods for updating soil information in *BIS*. Summarizing, based on the work presented in this thesis it can be concluded that:

- Updating the soil map of the peatlands with digital soil type mapping proves to be challenging (Chapters 2 and 5). The multinomial logistic model as well as the GLGM were able to capture the general patterns of soil distribution but not the details. Challenges are not only methodological but also practical. Soil conditions in peatlands are highly heterogeneous at short distances as a result of reclamation and intensive soil cultivation practices, the effects of which cannot be adequately captured by available environmental covariates.

- Use of a spatial model for soil type mapping, in form of the GLGM, did not improve prediction performance in the peatlands compared to a non-spatial model (Chapter 5).
- Soil maps that represent soil type by probability distributions offer opportunities for the development of DSM methods for two- and three-dimensional soil property mapping that fully exploit this information (Chapters 3 and 4). These methods have several operational advantages compared to conventional DSM methods, although these did not result in improved predictions in the case studies of this thesis (Chapters 3 and 6).
- Validation with independent data indicates that updating soil type and property maps with DSM does not result in maps of larger accuracy than updating with conventional methods. Nevertheless, DSM is attractive for updating because it is much more efficient than CSM in terms of costs in relation to accuracy. Fieldwork effort can be greatly reduced compared to CSM, without compromising accuracy (Chapter 6).
- Effective and successful operational DSM for updating and upgrading soil information in *BIS* cannot be solely achieved by developing a toolbox with state-of-the-art methods for soil mapping. It also requires new strategies for collecting, storing, processing, visualizing and disseminating soil information, i.e. it requires a next-generation soil information system based on a new paradigm of quantitative soil survey (Chapter 7).

In this chapter several suggestions were made for follow-up research: procedures for efficient use of legacy data; use of remote and proximal sensing methods for soil data collection and mapping; design of efficient sampling strategies for spatial data collection; extension of the soil information system toolbox with more advanced (dynamic) geostatistical models; development of state-of-the-art tools for visualizing and disseminating soil data. Additionally, further attention should be given to methods for updating, both on a fundamental level, such as the development of a spatial model for multinomial data (section 5.5), and a more applied level, such as updating via mapping key diagnostic properties (section 7.2.5 and Fig 7.1). In addition to scientific challenges there are also practical hurdles that need to be overcome, such as convincing practitioners of the merits of DSM that is necessary for its widespread adoption and operational success.

Soils are back on the global political agenda, fuelled by increased awareness of the important role of soils in sustainable development. Also on national level there is renewed appreciation for the soil, which has resulted in new demands for accurate, up-to-date soil information. With this thesis I hope to have made a significant contribution towards a 21st-century soil information system based on a new paradigm of quantitative soil survey.



References

- Augustin, N.H. Muggleston M.A. Buckland, S. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33:339–347.
- Bailey, N., Clements, T., Lee, J. T., and Thompson, S. (2003). Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach. *Journal of Environmental Management*, 67(4):395–407.
- Bartholomeus, H., Kooistra, L., Stevens, A., van Leeuwen, M., van Wesemael, B., Ben-Dor, E., and Tychon, B. (2011). Soil organic carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. *International Journal of Applied Earth Observation and Geoinformation*, 13(1):81–88.
- Beaudette, D. and O'Geen, A. (2009). Soil-Web: An online soil survey for California, Arizona, and Nevada. *Computers & Geosciences*, 35(10):2119–2128.
- Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., and Goldschmitt, M. (2005). Digital soil mapping using artificial neural networks. *Journal of Plant Nutrition and Soil Science*, 168(1):21–33.
- Beilman, D., Vitt, D., Bhatti, J., and Forest, S. (2008). Peat carbon stocks in the southern Mackenzie River Basin: uncertainties revealed in a high-resolution case study. *Global Change Biology*, 14(6):1221–1232.
- Ben-Ahmed, K., Bouratbine, A., and El-Aroui, M.-A. (2010). Generalized linear spatial models in epidemiology: A case study of zoonotic cutaneous leishmaniasis in Tunisia. *Journal of Applied Statistics*, 37(1):159–170.
- Ben-Dor, E., Patkin, K., Banin, A., and Karnieli, A. (2002). Mapping of several soil properties using DAIS-7915 hyperspectral scanner data - a case study over clayey soils in Israel. *International Journal of Remote Sensing*, 23(6):1043–1062.
- Bendel, R. B. and Afifi, A. A. (1977). Comparison of stopping rules in forward stepwise regression. *Journal of the American Statistical Association*, 72:46–53.
- Berglund, O. and Berglund, K. (2010). Distribution and cultivation intensity of agricultural peat and gyttja soils in Sweden and estimation of greenhouse gas emissions from cultivated peat soils. *Geoderma*, 154(3-4):173–180.
- Berglund, O. and Berglund, K. (2011). Influence of water table level and soil properties on emissions of greenhouse gases from cultivated peat soil. *Soil Biology and Biochemistry*, 43(5):923 – 931.
- Bie, S. W. and Beckett, P. H. T. (1971). Quality control in soil survey II. The costs of soil survey. *European Journal of Soil Science*, 22(4):453–465.
- Bierkens, M. F. P. and Burrough, P. A. (1993). The indicator approach to categorical soil data. I. Theory. *European Journal of Soil Science*, 44(2):361–368.
- Bishop, T. F. A., McBratney, A. B., and Laslett, G. M. (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91(1-2):27–45.
- Blume, H. P. and Leinweber, P. (2004). Plaggen soils: Landscape history, properties, and classification. *Journal of Plant Nutrition and Soil Science*, 167(3):319–327.
- Boettinger, J., Howell, D., Moore, A., Hartemink, A., and Kienast-Brown, S., editors (2010). *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*, volume 2 of *Progress in Soil Science*. Springer, New York.
- Bouma, J. (2010). Implications of the knowledge paradox for soil science. *Advances in Agronomy*, 106:143–171.
- Bouma, J. and Droogers, P. (2007). Translating soil science into environmental policy: A case study on implementing the EU soil protection strategy in The Netherlands. *Environmental Science and Policy*, 10(5):454–463.
- Bregt, A. K. (1992a). Optimum observation density for mapping acid sulphate soils in Conoco, Indonesia: accuracy and costs. In Bregt, A. K., editor, *Processing of soil survey data (PhD Thesis)*. Wageningen University.
- Bregt, A. K. (1992b). *Processing of soil survey data*. Phd thesis, Wageningen Agricultural University.
- Bregt, A. K. and Beemster, J. G. R. (1989). Accuracy in predicting moisture deficits and changes in yield from soil maps. *Geoderma*, 43(4):301–310.
- Bregt, A. K., Bouma, J., and Jellinek, M. (1987). Comparison of thematic maps derived from a soil map and from kriging of point data. *Geoderma*, 39(4):281–291.
- Bregt, A. K., Gesink, H. J., and Alkasuma (1992). Mapping the conditional probability of soil variables. *Geoderma*, 53(1-2):15–29.
- Brus, D., Spätjens, L. E. E. M., and de Gruijter, J. J. (1999). A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. *Geoderma*, 89(1-2):129–148.

- Brus, D. J. (1994). Improving design-based estimation of spatial means by soil map stratification. A case study of phosphate saturation. *Geoderma*, 62(1-3):233–246.
- Brus, D. J. (2000). Using regression models in design-based estimation of spatial means of soil properties. *European Journal of Soil Science*, 51(1):159–172.
- Brus, D. J., Bogaert, P., and Heuvelink, G. B. M. (2008). Bayesian maximum entropy prediction of soil categories using a traditional soil map as soft information. *European Journal of Soil Science*, 59(2):166–177.
- Brus, D. J. and de Gruijter, J. J. (1993). Design-based versus model-based estimates of spatial means: theory and application in environmental soil science. *Environmetrics*, 4(2):123–152.
- Brus, D. J. and de Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma*, 80(1-2):1–44.
- Brus, D. J. and de Gruijter, J. J. (2011). Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data. *Geoderma*, 164(3-4):172–180.
- Brus, D. J., de Gruijter, J. J., and Breeuwsma, A. (1992). Strategies for updating soil survey information: a case study to estimate phosphate sorption characteristics. *European Journal of Soil Science*, 43(3):567–581.
- Brus, D. J., de Gruijter, J. J., Marsman, B. A., Visschers, R., Bregt, A. K., Breeuwsma, A., and Bouma, J. (1996). The performance of spatial interpolation methods and choropleth maps to estimate properties at points: A soil survey case study. *Environmetrics*, 7(1):1–16.
- Brus, D. J., de Gruijter, J. J., Walvoort, D. J. J., de Vries, F., Bronswijk, J. J. B., Romkens, P. F. A. M., and de Vries, W. (2002). Mapping the probability of exceeding critical thresholds for cadmium concentrations in soils in the Netherlands. *Journal of Environmental Quality*, 31(6):1875–1884.
- Brus, D. J. and Heuvelink, G. B. M. (2007a). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138(1-2):86–95.
- Brus, D. J. and Heuvelink, G. B. M. (2007b). Towards a soil information system with quantified accuracy; three approaches for stochastic simulation of soil maps. Technical report, Statutory Research Tasks Unit for Nature and the Environment, WOt-rapport 58.
- Brus, D. J. and Jansen, M. J. W. (2004). Uncertainty and sensitivity analysis of spatial predictions of heavy metals in wheat. *Journal of Environmental Quality*, 33(3):882–890.
- Brus, D. J., Kempen, B., and Heuvelink, G. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62(3):394–407.
- Brus, D. J., Vasat, R., Heuvelink, G. B. M., Knotters, M., de Vries, F., and Walvoort, D. J. J. (2009). Towards a soil information system with quantified accuracy; a prototype for mapping continuous soil properties. Technical report, Statutory Research Tasks Unit for Nature and the Environment, WOt-werkdocument 197.
- Bui, E. (2007). A review of digital soil mapping in Australia. In Lagacherie, P., McBratney, A. B., and Voltz, M., editors, *Digital Soil Mapping - An introductory perspective*, volume 31 of *Developments in Soil Science*, pages 25–37. Elsevier, Amsterdam.
- Bui, E. N., Henderson, B. L., and Viergever, K. (2006). Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling*, 191(3-4):431–446.
- Bui, E. N. and Moran, C. J. (2003). A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling Basin of Australia. *Geoderma*, 111(1-2):21–44.
- Burgess, T. M. and Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties. I The semi-variogram and punctual kriging. *European Journal of Soil Science*, 31(2):315–331.
- Buringh, P., Steur, G., and Vink, A. (1962). Some techniques and methods for soil survey in the Netherlands. *Netherlands Journal of Agricultural Sciences*, 10(2):157–172.
- Burrough, P. A., Bouma, J., and Yates, S. R. (1994). The state of the art in pedometrics. *Geoderma*, 62(1-3):311–326.
- Burrough, P. A., van Gaans, P. F. M., and Hootsmans, R. (1997). Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma*, 77(2-4):115–135.
- Buurman, P. and Sevink, J., editors (1995). *Van bodemkaart tot informatiesysteem*. Wageningen Press, Wageningen.
- Campling, P., Gobin, A., and Feyen, J. (2002). Logistic modeling to spatially predict the probability of soil drainage classes. *Soil Sci Soc Am J*, 66(4):1390–1401.
- Carré, F., McBratney, A. B., and Minasny, B. (2007). Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. *Geoderma*, 141(1-2):1–14.
- Chaplot, V., Bernoux, M., Walter, C., Curmi, P., and Herpin, U. (2001). Soil carbon storage prediction in temperate hydromorphic soils using a morphologic index and digital elevation model. *Soil Science*, 166(1):48–60.
- Chen, Z. and Kuo, L. (2001). A note on the estimation of the multinomial logit model with random effects. *The American Statistician*, 55(2):89–95.
- Christensen, O. F. (2004). Monte carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, 13(3):702–718.
- Christensen, O. F. and Ribeiro Jr., P. (2002). `geoRglm`: A package for generalized linear spatial models. *R-News* 2:26–28. Available at: <http://cran.r-project.org/doc/rnews>.
- Clement, J. and Kooistra, L. (2003). Eerste bosstatistiek digitaal; opbouw van een historisch basisbestand. Technical Report 744, Alterra.
- Commission of the European Communities (2006). Proposal for a Directive of the European Parliament and of the Council establishing a framework for the protection of soil. Technical Report COM-232, Office for Official Publications of the European Communities, Brussels.

- Cook, S. E., Corner, R. J., Grealish, G., Gessler, P. E., and Chartres, C. J. (1996). A rule-based system to map soil properties. *Soil Sci Soc Am J*, 60(6):1893–1900.
- Cosandey, A.-C., Guenat, C., Bouzelboudjen, M., Maître, V., and Bovier, R. (2003). The modelling of soil-process functional units based on three-dimensional soil horizon cartography, with an example of denitrification in a riparian zone. *Geoderma*, 112(1-2):111–129.
- de Bakker, H. and Schelling, J. (1966). *A system of soil classification for the Netherlands: The higher levels*. Pudoc, Wageningen.
- de Bakker, H. and Schelling, J. (1989). *System of soil classification for the Netherlands: The higher levels*. Winand Staring Centre, Wageningen, Second, adjusted edition, edited by D.J. Brus en C. van Wallenburg edition.
- de Groot, W. J. M., Visschers, R., Kiestra, E., Kuikman, P. J., and Nabuurs, G. J. (2005). Nationaal systeem voor de rapportage van voorraad en veranderingen in bodem-C in relatie tot landgebruik en landgebruikveranderingen in Nederland aan de UNFCCC. Technical Report 1035-3, Alterra.
- de Gruijter, J. J., Brus, D. J., Bierkens, M. F. P., and Knotters, M. (2006). *Sampling for natural resource monitoring*. Springer.
- de Gruijter, J. J. and Marsman, B. A. (1985). Transect sampling for reliable information on mapping units. In Nielson, D. and Bouma, J., editors, *Soil spatial variability: proceedings of a workshop of the ISSS and SSSA*, pages 150–163, Las Vegas.
- de Gruijter, J. J. and ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, 22(4):407–415.
- de Vries, F. (1999). Karakterisering van Nederlandse gronden naar fysisch-chemische kenmerken. Technical Report 125, DLO-Staring Centrum.
- de Vries, F. and Brouwer, F. (2006). De bodem van Drenthe in beeld. Technical Report 1381, Alterra.
- de Vries, F., Lesschen, J. P., van den Akker, J. J. H., Petrescu, A. M. R., van Huissteden, J., and van den Wyngaert, I. (2009). Bodemgerelateerde emissie van broeikasgassen in Drenthe; de huidige situatie. Technical Report 1859, Alterra.
- de Vries, F., Mol, G., Hack-ten Broeke, M. J. D., Heuvelink, G. B. M., and Brouwer, F. (2008). Het Bodemkundig Informatie Systeem van Alterra. Technical Report 1709, Alterra.
- de Vries, F., Stoffelsen, G. H., and van der Werff, M. M. (2010). Validatie bodemkaart van de veengebieden in Noord-Holland. Technical Report 2085, Alterra.
- Debella-Gilo, M. and Etzelmüller, B. (2009). Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. *CATENA*, 77(1):8–18.
- Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M. (2002). Childhood malaria in the Gambia: A case-study in model-based geostatistics. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 51(4):493–506.
- Diggle, P., Ribeiro Jr., P., and Christensen, O. F. (2003). An introduction to model-based geostatistics. In Møller, J., editor, *Spatial Statistics and Computational methods*, volume 173 of *Lecture Notes in Statistics*. Springer, New York.
- Diggle, P. J. and Ribeiro Jr., P. J. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer, New York.
- Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 47(3):299–325.
- Dijkerman, J. C. (1974). Pedology as a science: The role of data, models and theories in the study of natural soil systems. *Geoderma*, 11(2):73–93.
- Domburg, P., de Gruijter, J. J., and Brus, D. J. (1994). A structured approach to designing soil survey schemes with prediction of sampling error from variograms. *Geoderma*, 62(1-3):151–164.
- D'Or, D. and Bogaert, P. (2004). Spatial prediction of categorical variables with the Bayesian Maximum Entropy approach: The Ooypolder case study. *European Journal of Soil Science*, 55(4):763–775.
- Dormann, C. (2007). Assessing the validity of autologistic regression. *Ecological Modelling*, 207(2-4):234–242.
- European Commission (2006). Soil protection - the long story behind the strategy. Technical report, Office for Official Publications of the European Communities, Luxembourg.
- Finke, P. (1993). Field scale variability of soil structure and its impact on crop growth and nitrate leaching in the analysis of fertilizing scenarios. *Geoderma*, 60(1-4):89–107.
- Finke, P. A. (2000). Updating the (1:50,000) Dutch groundwater table class map by statistical methods: an analysis of quality versus cost. *Geoderma*, 97(3-4):329–350.
- Finke, P. A. (2011). On digital soil assessment with models and the Pedometrics agenda. *Geoderma*, In Press, Corrected Proof:–.
- Finke, P. A., Bouma, J., and Stein, A. (1992). Measuring field variability of disturbed soils for simulation purposes. *Soil Sci Soc Am J*, 56(1):187–192.
- Finke, P. A., Brus, D. J., Bierkens, M. F. P., Hoogland, T., Knotters, M., and de Vries, F. (2004). Mapping groundwater dynamics using multiple sources of exhaustive high resolution data. *Geoderma*, 123(1-2):23–39.
- Finke, P. A., Groot Obbink, D. J., Rosing, H., and de Vries, F. (1996). Actualisatie Gt-kaarten 1 : 50 000 Drents deel kaartbladen 16 Oost (Steenwijk) en 17 West (Emmen). Technical Report 439, DLO-Staring Centrum.
- Florinsky, I. V., Eilers, R. G., Manning, G. R., and Fuller, L. G. (2002). Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software*, 17(3):295–311.
- Footy, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201.
- Freibauer, A., Rounsevell, M. D. A., Smith, P., and Verhagen, J. (2004). Carbon sequestration in the agricultural soils of Europe. *Geoderma*, 122(1):1 – 23.
- Gomez, C., Viscarra Rossel, R. A., and McBratney, A. B. (2008). Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma*, 146(3-4):403–411.
- Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.
- Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma*, 103(1-2):3–26.

- Goovaerts, P. (2011). A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties. *European Journal of Soil Science*, 62(3):371–380.
- Goovaerts, P. and Journel, A. G. (1995). Integrating soil map information in modelling the spatial variation of continuous soil properties. *European Journal of Soil Science*, 46(3):397–414.
- Gorham, E. (1991). Northern peatlands: role in the carbon cycle and probable responses to climatic warming. *Ecological Applications*, 1(2):182–195.
- Grimm, R. and Behrens, T. (2010). Uncertainty analysis of sample locations within digital soil mapping approaches. *Geoderma*, 155(3-4):154–163.
- Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using random forests analysis. *Geoderma*, 146(1-2):102–113.
- Grinand, C., Arrouays, D., Laroche, B., and Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143(1-2):180–190.
- Grønlund, A., Hauge, A., Hovde, A., and Rasse, D. P. (2008). Carbon loss estimates from cultivated peat soils in Norway: A comparison of three methods. *Nutrient Cycling in Agroecosystems*, 81(2):157–167.
- Grunwald, S., editor (2006). *Environmental soil-landscape modeling: geographic information technologies and pedometrics*. Taylor & Francis, Boca Raton.
- Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, 152(3-4):195–207.
- Grunwald, S. (2010). Current state of digital soil mapping and what is next. In Boettinger, J., Howell, D., Moore, A., Hartemink, A., and Kienast-Brown, S., editors, *Digital Soil Mapping - Bridging Research, Environmental Application, and Operation*, number 2 in Progress in Soil Science, pages 3–12. Springer, New York.
- Hack-ten Broeke, M. J. D., Schut, A. G. T., and Bouma, J. (1999). Effects on nitrate leaching and yield potential of implementing newly developed sustainable land use systems for dairy farming on sandy soils in the Netherlands. *Geoderma*, 91(3-4):217–235.
- Hanegraaf, M. C., Hoffland, E., Kuikman, P. J., and Brussaard, L. (2009). Trends in soil organic matter contents in Dutch grasslands and maize fields on sandy soils. *European Journal of Soil Science*, 60(2):213–222.
- Hartemink, A., McBratney, A., and Mendonça-Santos, M., editors (2008). *Digital Soil Mapping with Limited Data*. Springer, New York.
- Hartemink, A. E. (2008). Soils are back on the global agenda. Guest Editorial. *Soil Use and Management*, 24(4):327–330.
- Hartemink, A. E., Hempel, J., Lagacherie, P., McBratney, A., McKenzie, N., MacMillan, R. A., Minasny, B., Montanarella, L., Mendonça Santos, M. L., Sanchez, P., Walsh, M., and Zhang, G.-L. (2010). GlobalSoilMap.net—A new digital soil map of the world. In Boettinger, J., Howell, D., Moore, A., Hartemink, A., and Kienast-Brown, S., editors, *Digital Soil Mapping - Bridging Research, Environmental Application, and Operation*, number 2 in Progress in Soil Science, pages 423–427. Springer, New York.
- Hartemink, A. E. and McBratney, A. (2008). A soil science renaissance. *Geoderma*, 148(2):123–129.
- Hartman, L. W. (2006). Bayesian modelling of spatial data using Markov random fields, with application to elemental composition of forest soil. *Mathematical Geology*, 38(2):113–133.
- Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, 1(2):81–102.
- Hazeu, G. W. (2005). Landelijk Grondgebruiksbestand Nederland (LGN5). Vervaardiging, nauwkeurigheid en gebruik. Technical Report 1213, Alterra.
- Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22(9):1433–1446.
- Henderson, B. L., Bui, E. N., Moran, C. J., and Simon, D. A. P. (2005). Australia-wide predictions of soil properties using decision trees. *Geoderma*, 124(3-4):383–398.
- Hengl, T. (2003). *Pedometric mapping: bridging gaps between conventional and pedometric approaches*. Phd thesis, Wageningen University / ITC.
- Hengl, T., Heuvelink, G. B. M., and Rossiter, D. G. (2007a). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315.
- Hengl, T., Heuvelink, G. B. M., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2):75–93.
- Hengl, T., MacMillan, R. A., Walsh, M., and Reuter, H. I. (2011). *Global Soil Information Facilities: a methodological framework for open soil information*. ISRIC—World Soil Information. *Under Review*.
- Hengl, T., Rossiter, D., and Stein, A. (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research*, 41:1403–1422.
- Hengl, T. and Toomanian, N. (2006). Maps are not what they seem: representing uncertainty in soil property maps. *Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, (eds. M. Caetano and M. Painho), pp. 805–813. Lisbon, 5–7 July 2006.
- Hengl, T., Toomanian, N., Reuter, H. I., and Malakouti, M. J. (2007b). Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, 140(4):417–427.
- Heuvelink, G. (2007). Soil spatial prediction with Markov random fields. *Pedometron*, 23:16–20.
- Heuvelink, G., Brus, D., de Vries, F., Vašát, R., Walvoort, D., Kempen, B., and Knotters, M. (2010). Implications of digital soil mapping for soil information systems. *Proceedings of the 4th Global Workshop on Digital Soil Mapping*, (eds. R. Napoli and R. Francaviglia), pp. 67. Rome, 24–26 May 2010.
- Heuvelink, G. B. M. (1996). Identification of field attribute error under different models of spatial variation. *International Journal of Geographical Information Science*, 10(8):921–935.

- Heuvelink, G. B. M. and Bierkens, M. F. P. (1992). Combining soil maps with interpolations from point observations to predict quantitative soil properties. *Geoderma*, 55(1-2):1–15.
- Heuvelink, G. B. M. and Griffith, D. A. (2010). Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42(2):161–179.
- Heuvelink, G. B. M. and Huisman, J. A. (2000). Choosing between abrupt and gradual spatial variation? In Mowrer, H. T. and Congalton, R. G., editors, *Quantifying spatial uncertainty in natural resources: theory and applications for GIS and remote sensing*, pages 111–117. Ann Arbor Press.
- Heuvelink, G. B. M. and Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma*, 100(3-4):269–301.
- Hewitt, A. E. (1993). Predictive modelling in soil survey. *Soils and Fertilizers*, 56:305–314.
- Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics*, 5(1):102–114.
- Hoogland, T., van den Akker, J. J. H., and Brus, D. J. (2011). Modeling the subsidence of peat soils in the Dutch coastal area. *Geoderma*, In Press, Corrected Proof:–.
- Hosmer, D. and Lemeshow, S. (2000). *Applied Logistic Regression (2nd Edition)*. John Wiley & Sons, New York.
- Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.
- Hou, Q., Young, L. J., Brandle, J. R., and Schoeneberger, M. M. (2011). A spatial model approach for assessing windbreak growth and carbon stocks. *Journal of Environmental Quality*, 40(3):842–852.
- Howell, D., Kim, Y. G., and Haydu-Houdeshell, C. A. (2008). Development and application of digital soil mapping within traditional soil survey: What will it grow into? In Hartemink, A., McBratney, A., and Mendonça-Santos, M., editors, *Digital Soil Mapping with Limited Data*, pages 43–51. Springer, New York.
- Jenny, H. (1941). *Factors of soil formation, A system of quantitative pedology*. McGraw-Hill, New York.
- Jones, R. J. A., Hiederer, R., Rusco, E., and Montanarella, L. (2005). Estimating organic carbon in the soils of Europe for policy support. *European Journal of Soil Science*, 56(5):655–671.
- Kasatkasem, T., A. M. K. V. P. K. (2005). Super-resolution land cover mapping using a markov random field based approach. *Remote Sensing of Environment*, 96(3-4):302–314.
- Knol, W., Kramer, H., Dorland, G., and Gijsbertse, H. (2003). Historisch Grondgebruik Nederland: tijdreeksen grondgebruik Noord-Holland van 1950 tot 1980. Technical Report 751, Alterra.
- Knol, W., Kramer, H., and Gijsbertse, H. (2004). Historisch Grondgebruik Nederland: een landelijke reconstructie van het grondgebruik rond 1900. Technical Report 573, Alterra.
- Knotters, M., Brus, D., Heuvelink, G., Kempen, B., de Vries, F., and Walvoort, D. (2010a). Vaste grond onder de voeten? Geactualiseerd Bodemkundig Informatie Systeem informeert over onzekerheid. *Bodem*, 20(5):22–25.
- Knotters, M., Brus, D. J., and Oude Voshaar, J. H. (1995). A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, 67(3-4):227–246.
- Knotters, M., Vroon, H., van Kekem, A., and Hoogland, T. (2010b). Deciding on the detail of soil survey in estimating crop yield reduction due to groundwater withdrawal. In Devillers, R. and Goodchild, H., editors, *Spatial Data Quality: From Process to Decisions*, pages 117–125. CRC Press.
- Koomen, A. and Maas, G. (2004). Geomorfologische Kaart Nederland (GKN); Achtergronddocument bij het landsdekkende digitale bestand. Technical Report 1039, Alterra.
- Kros, J., Frumau, K., Hensen, A., and de Vries, W. (2011). Integrated analysis of the effects of agricultural management on nitrogen fluxes at landscape scale. *Environmental Pollution*, In Press, Corrected Proof:–.
- Lagacherie, P., Baret, F., Feret, J.-B., Netto, J. M., and Robbez-Masson, J. M. (2008). Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. *Remote Sensing of Environment*, 112(3):825–835.
- Lagacherie, P., Legros, J. P., and Burfough, P. A. (1995). A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. *Geoderma*, 65(3-4):283–301.
- Lagacherie, P. and McBratney, A. B. (2007). Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In Lagacherie, P., McBratney, A. B., and Voltz, M., editors, *Digital Soil Mapping - An introductory perspective*, volume 31 of *Developments in Soil Science*, pages 3–22. Elsevier, Amsterdam.
- Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, 304:1623–1627.
- Lark, R. and Cullis, B. (2004). Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, 55(4):799–813.
- Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, 51(4):717–728.
- Lark, R. M. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1-2):49–80.
- Lark, R. M., Bishop, T. F. A., and Webster, R. (2007). Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. *Geoderma*, 138(1-2):65–78.
- Lark, R. M., Cullis, B. R., and Welham, S. J. (2006). On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor E-BLUP with REML. *European Journal of Soil Science*, 57(6):787–799.
- Lark, R. M. and Webster, R. (2006). Geostatistical mapping of geomorphic variables in the presence of trend. *Earth Surface Processes and Landforms*, 31(7):862–874.
- Lee, K. I. and Koval, J. J. (1997). Determination of the best significance level in forward stepwise logistic regression.

References

- Communications in Statistics Part B: Simulation and Computation*, 26(2):559–575.
- Leenhardt, D., Voltz, M., Bornand, M., and Webster, R. (1994). Evaluating soil maps for prediction of soil water properties. *European Journal of Soil Science*, 45(3):293–301.
- Leeters, E., Bregt, A. K., Schoumans, O. F., and Stolp, J. (1990). Verder met het bodemkundig informatiesysteem. Interne Mededeling 88, Staring Centrum.
- Leifeld, J., Müller, M., and Fuhrer, J. (2011). Peatland subsidence and carbon loss from drained temperate fens. *Soil Use and Management*, 27(2):170–176.
- Li, W. and Zhang, C. (2007). A random-path markov chain algorithm for simulating categorical soil variables from random point samples. *Soil Sci Soc Am J*, 71(3):656–668.
- Liu, T. L., Juang, K. W., and Lee, D. Y. (2006). Interpolating soil properties using kriging combined with categorical information of soil maps. *Soil Sci Soc Am J*, 70(4):1200–1209.
- Makken, H. and de Vries, F. (1989). Bodem en grondwater opnieuw op de kaart; revisie van de Bodemkaart van Nederland 1:50 000, kaartblad 12 Oost en 17 Oost. Technical Report 36, Staring Centrum.
- Maljanen, M., Sigurdsson, B., Guðmundsson, J., Óskarsson, H., Huttunen, J., and Martikainen, P. (2010). Greenhouse gas balances of managed peatlands in the Nordic countries present knowledge and gaps. *Biogeosciences*, 7(9):2711–2738.
- Malone, B., McBratney, A., and Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma*, 160(3–4):614–626.
- Malone, B. P., McBratney, A. B., Minasny, B., and Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154(1–2):138–152.
- Marchant, B. P. and Lark, R. M. (2004). Estimating variogram uncertainty. *Mathematical Geology*, 36(8):867–898.
- Marsman, B. A. and de Gruijter, J. J. (1986). Quality of soil maps; a comparison of survey methods in a sandy area. Technical Report No. 15, Soil Survey Institute.
- May, R., Van Dijk, J., Wabakken, P., Swenson, J. E., Linnell, J. D. C., Zimmermann, B., Odden, J., Pedersen, H. C., Andersen, R., and Landa, A. (2008). Habitat differentiation within the large-carnivore community of Norway's multiple-use landscapes. *Journal of Applied Ecology*, 45(5):1382–1391.
- McBratney, A., Webster, R., McLaren, R., and Spiers, R. (1982). Regional variation of extractable copper and cobalt in the topsoil of south-east scotland. *Agronomie*, 2(10):969–982.
- McBratney, A. B. (1998). Some considerations on methods for spatially aggregating and disaggregating soil information. *Nutrient Cycling in Agroecosystems*, 50(1–3):51–62.
- McBratney, A. B., Bishop, T. F. A., and Teliatnikov, I. S. (2000). Two soil profile reconstruction techniques. *Geoderma*, 97(3–4):209–221.
- McBratney, A. B. and Gruijter, J. J. (1992). A continuum approach to soil classification by modified fuzzy k-means with extragrades. *European Journal of Soil Science*, 43(1):159–175.
- McBratney, A. B., Mendonça Santos, M. L., and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1–2):3–52.
- McBratney, A. B. and Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *European Journal of Soil Science*, 37(4):617–639.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London.
- McKenzie, N. J. and Austin, M. P. (1993). A quantitative australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma*, 57(4):329–355.
- McKenzie, N. J. and Gallant, J. (2007). Digital soil mapping with improved environmental predictors and models of pedogenesis. In Lagacherie, P., McBratney, A. B., and Voltz, M., editors, *Digital Soil Mapping - An introductory perspective*, pages 327–349. Elsevier, Amsterdam.
- McKenzie, N. J. and Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89(1–2):67–94.
- Meersmans, J., De Ridder, F., Canters, F., De Baets, S., and Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, 143(1–2):1–13.
- Meersmans, J., van Wesemael, B., De Ridder, F., and Van Molle, M. (2009). Modelling the three-dimensional spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, 152:43–52.
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24.
- Mickey, R. M. and Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, 129(1):125–137.
- Minasny, B. and McBratney, A. B. (2006a). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32(9):1378–1388.
- Minasny, B. and McBratney, A. B. (2006b). Mechanistic soil-landscape modelling as an approach to developing pedogenetic classifications. *Geoderma*, 133(1–2):138–149.
- Minasny, B. and McBratney, A. B. (2007a). Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma*, 142(3–4):285–293.
- Minasny, B. and McBratney, A. B. (2007b). Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma*, 140(4):324–336.
- Minasny, B., McBratney, A. B., Mendonça Santos, M. L., Odeh, I. O. A., and Guyon, B. (2006). Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Australian Journal of Soil Research*, 44:233–244.
- Minasny, B., Vrugt, J. A., and McBratney, A. B. (2011). Confronting uncertainty in model-based geostatistics using Markov

- Chain Monte Carlo simulation. *Geoderma*, 163(3-4):150–162.
- Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., and Van Meirvenne, M. (2009). Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. *Soil Sci Soc Am J*, 73(2):614–621.
- Mittlböck, M. and Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15(19):1987–1997.
- Mol, G., de Vries, F., and Brus, D. J. (2005). Bodemdata.nl; webportaal voor de ontsluiting van gegevens uit het Bodemkundig Informatie Systeem (BIS). *Bodem*, 15(5):175–177.
- Moore, A. C., Howell, D. W., Haydu-Houdeshell, C. A., Blinn, C., Hempel, J., and Smith, D. (2010). Building digital soil mapping capacity in the natural resource conservation service: Mojave Desert operational initiative. In Boettinger, J., Howell, D., Moore, A., Hartemink, A., and Kienast-Brown, S., editors, *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*, number 2 in Progress in Soil Science, pages 357–367. Springer, New York.
- Mosegaard, K. and Sambridge, M. (2002). Monte Carlo analysis of inverse problems. *Inverse Problems*, 18:29–54.
- Müller, D. and Zeller, M. (2002). Land use dynamics in the central highlands of Vietnam: A spatial model combining village survey data with satellite imagery interpretation. *Agricultural Economics*, 27(3):333–354.
- Nelson, M. A., Bishop, T. F. A., Triantafyllis, J., and Odeh, I. O. A. (2011). An error budget for different sources of error in digital soil mapping. *European Journal of Soil Science*, 62(3):417–430.
- Nelson, M. A. and Odeh, I. O. A. (2009). Digital soil class mapping using legacy soil profile data: a comparison of a genetic algorithm and classification tree approach. *Australian Journal of Soil Research*, 47:632–649.
- Nol, L., Heuvelink, G. B. M., Veldkamp, A., de Vries, W., and Kros, J. (2010). Uncertainty propagation analysis of an N₂O emission model at the plot and landscape scale. *Geoderma*, 159(1-2):9–23.
- Norberg, T., Rosén, L., Baran, A., and Baran, S. (2002). On modelling discrete geological structures as Markov random fields. *Mathematical Geology*, 34(1):63–77.
- Oberthur, T., Goovaerts, P., and Dobermann, A. (1999). Mapping soil texture classes using field texturing, particle size distribution and local knowledge by both conventional and geostatistical methods. *European Journal of Soil Science*, 50(3):457–479.
- Odeh, I. O. A., McBratney, A. B., and Chittleborough, D. J. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, 63(3-4):197–214.
- Orton, T. G., Goulding, K. W. T., and Lark, R. M. (2011). Geostatistical prediction of nitrous oxide emissions from soil using data, process models and expert opinion. *European Journal of Soil Science*, 62(3):359–370.
- Palmgren, J. (1981). The Fisher information matrix for log linear models arguing conditionally on observed explanatory variable. *Biometrika*, 68(2):563–566.
- Papritz, A. (2009). Why indicator kriging should be abandoned. *Pedometron*, 26:4–7.
- Pardo-Igúzquiza, E. (1998). Maximum likelihood estimation of spatial covariance parameters. *Mathematical Geology*, 30(1):95–108.
- Pardo-Igúzquiza, E., Chica-Olmo, M., García-Soldado, M., and Luque-Espinar, J. (2009). Using semivariogram parameter uncertainty in hydrogeological applications. *Ground Water*, 47(1):25–34.
- Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30:683–691.
- Ponce-Hernandez, R., Marriot, F. H. C., and Beckett, P. H. T. (1986). An improved method for reconstructing a soil profile from analyses of a small number of samples. *Journal of Soil Science*, 37(3):455–467.
- R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Reijneveld, A., van Wensem, J., and Oenema, O. (2009). Soil organic carbon contents of agricultural land in the Netherlands between 1984 and 2004. *Geoderma*, 152(3-4):231–238.
- Rhemtulla, J. M., Mladenoff, D. J., and Clayton, M. K. (2007). Regional land-cover conversion in the U.S. upper Midwest: Magnitude of change and limited recovery (1850-1935-1993). *Landscape Ecology*, 22(SUPPL. 1):57–75.
- Ribeiro Jr, P. and Diggle, P. J. (2001). *geoR: A package for geostatistical data analysis using the R software*. *R-News* 1(2). Available at: <http://cran.r-project.org/doc/rnews>.
- Rivero, R. G., Grunwald, S., Osborne, T. Z., Reddy, K. R., and Newman, S. (2007). Characterization of the spatial distribution of soil properties in Water Conservation Area 2A, Everglades, Florida. *Soil Science*, 172(2):149–166.
- Rosing, H., Thijssen, G. L., and Brouwer, F. (2006). Actualisatie en modernisering van de Bodemkaart van Nederland, schaal 1:50 000; Een test in de omgeving van Helmond. Technical Report 1057, Alterra.
- Rossiter, D. G. (2000). *Methodology for Soil Resource Inventories*. ITC Lecture Notes.
- Rossiter, D. G. (2004). Digital soil resource inventories: status and prospects. *Soil Use and Management*, 20(3):296–301.
- Sanchez, P., Ahamed, S., Carré, F., Hartemink, A. E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A., McKenzie, N. J., Mendonça Santos, M. L., Minasny, B., Montanarella, L., Okoth, P., Palm, C., Sachs, J., Sheperd, K., Vågen, T-G. Vanlauwe, B., Walsh, M., Winowiecki, L., and Zhang, G.-L. (2009). Digital soil map of the world. *Science*, 325:680–681.
- Schelling, J. (1970). Soil genesis, soil classification and soil survey. *Geoderma*, 4(3):165–193.
- Schulp, C. J. E. and Veldkamp, A. (2008). Long-term landscape - land use interactions as explaining factor for soil organic matter variability in dutch agricultural landscapes. *Geoderma*, 146(3-4):457–465.
- Selvaradjou, S.-K., Montanarella, L., Spaargaren, O., and Dent, D. (2005a). European digital archive of soil maps (Eu-DASM) Soil maps of Africa. Technical Report EUR 21657 EN, Office for Official Publications of the European Communities, Luxembourg.
- Selvaradjou, S.-K., Montanarella, L., Spaargaren, O., Dent, D., and Filippi, N. (2005b). European digital archive of soil

- maps (EuDASM) Metadata of the soil maps of Asia. Technical Report EUR 21820 EN, Office for Official Publications of the European Communities, Luxembourg.
- Skidmore, A. K., Watford, F., Luckananurug, P., and Ryan, P. J. (1996). An operational GIS expert system for mapping forest soils. *Photogrammetric Engineering and Remote Sensing*, 62(5):501–511.
- Smith, P. (1994). Autocorrelation in logistic regression modelling of species' distributions. *Global Ecology & Biogeography Letters*, 4:47–61.
- Soil Survey Division Staff (1993). *Soil Survey Manual*. Soil Conservation Service. U.S. Department of Agriculture Handbook 18.
- Sonneveld, M. P. W., Bouma, J., and Veldkamp, A. (2002). Refining soil survey information for a Dutch soil series using land use history. *Soil Use and Management*, 18(3):157–163.
- Sonneveld, M. P. W., Hack-ten Broeke, M. J. D., van Diepen, C. A., and Boogaard, H. L. (2010). Thirty years of systematic land evaluation in the Netherlands. *Geoderma*, 156(3-4):84–92.
- Spek, T. (2004). *Het Drentse esdorpenlandschap : een historisch-geografische studie*. Phd thesis, Wageningen University.
- Stacey, K. F., Lark, R. M., Whitmore, A. P., and Milne, A. E. (2006). Using a process model and regression kriging to improve predictions of nitrous oxide emissions from soil. *Geoderma*, 135:107–117.
- Steur, G. (1961). Methods of soil surveys in use in the Netherlands Soil Survey Institute. *Boor en Spade*, 11:59–77.
- Steur, G. G. L., de Vries, F., and van Wallenburg, C. (1985). *Bodemkaart van Nederland 1:250.000: beknopte beschrijving van de kaartenheden*. Stiboka, Wageningen.
- Steur, G. G. L. and Heijink, W., editors (1991). *Bodemkaart van Nederland, schaal 1:50.000; Algemene begrippen en indelingen (4e editie)*. Staring Centrum, Wageningen.
- Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M., and de Bruin, S. (2009). Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, 149(1-2):161–170.
- Stum, A. K., Boettinger, J. L., White, M. A., and Ramsey, R. D. (2010). Random forests applied as a soil spatial predictive model in arid Utah. In Boettinger, J., Howell, D., Moore, A., Hartemink, A., and Kienast-Brown, S., editors, *Digital Soil Mapping - Bridging Research, Environmental Application, and Operation*, number 2 in Progress in Soil Science, pages 179–189. Springer, New York.
- Suring, L. H., Goldstein, M. I., Howell, S. M., and Nations, C. S. (2008). Response of the cover of berry-producing species to ecological factors on the Kenai Peninsula, Alaska, USA. *Canadian Journal of Forest Research*, 38(5):1244–1259.
- ten Cate, J. A. M., van Holst, A., Kleijer, H., and Stolp, J. (1995). Handleiding bodemgeografisch onderzoek: richtlijnen en voorschriften. Technical Report 19, DLO-Staring Centrum.
- Thompson, J. A. and Kolka, R. K. (2005). Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. *Soil Sci Soc Am J*, 69(4):1086–1093.
- van Beek, C. L., Pleijter, M., and Kuikman, P. J. (2011). Nitrous oxide emissions from fertilized and unfertilized grasslands on peat soil. *Nutrient Cycling in Agroecosystems*, 89(3):453–461.
- van der Klooster, E., van Egmond, F. M., and Sonneveld, M. P. W. (2011). Mapping soil clay contents in Dutch marine districts using gamma-ray spectrometry. *European Journal of Soil Science*, pages no-no.
- van der Pouw, B. J. A. and Finke, P. A. (1999). Development and perspective of soil survey in the Netherlands. In Bullock, P., Jones, R. J. A., and Montanarella, L., editors, *Soil resources of Europe*, volume EUR 18991 EN. Office for Official Publications of the European Communities, Luxembourg.
- van der Salm, C., de Vries, W., and Kros, J. (1996). Modelling trends in soil solution concentrations under five forest-soil combinations in the Netherlands. *Ecological Modelling*, 88(1-3):19 – 37.
- van Kekem, A. J., Hoogland, T., and van der Horst, J. B. F. (2005). Uitspoelingsgevoelige gronden op de kaart; werkwijze en resultaten. Technical Report 1080, Alterra.
- Van Kuilenburg, J., de Gruijter, J. J., Marsman, B. A., and Bouma, J. (1982). Accuracy of spatial interpolation between point data on soil moisture supply capacity, compared with estimates from mapping units. *Geoderma*, 27(4):311–325.
- van Lynden, K. R., van Soesbergen, G. A., van Wallenburg, C., and Westerveld, G. J. W. (1985). De interpretatie van de bodemkaarten in Nederland. *Cultuurtechnisch Tijdschrift*, 25(2):58–68.
- van Meirvenne, M., Maes, K., and Hofman, G. (2003). Three-dimensional variability of soil nitrate-nitrogen in an agricultural field. *Biology and Fertility of Soils*, 37(3):147–153.
- van Meirvenne, M., Scheldeman, K., Baert, G., and Hofman, G. (1994). Quantification of soil textural fractions of Bas-Zaire using soil map polygons and/or point observations. *Geoderma*, 62(1-3):69–82.
- Vanwallegghem, T., Poesen, J., McBratney, A., and Deckers, J. (2010). Spatial variability of soil horizon depth in natural loess-derived soils. *Geoderma*, 157(1-2):37–45.
- Vasques, G. M., Grunwald, S., Comerford, N. B., and Sickman, J. O. (2010). Regional modelling of soil carbon at multiple depths within a subtropical watershed. *Geoderma*, 156(3-4):326–336.
- Viscarra Rossel, R. A., Cattle, S. R., Ortega, A., and Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. *Geoderma*, 150(3-4):253–266.
- Viscarra Rossel, R. A. and Chen, C. (2011). Digitally mapping the information content of visible-near infrared spectra of surficial Australian soils. *Remote Sensing of Environment*, 115(6):1443–1455.
- Viscarra Rossel, R. A. and Walter, C. (2004). Rapid, quantitative and spatial field measurements of soil pH using an Ion Sensitive Field Effect Transistor. *Geoderma*, 119(1-2):9–20.
- Visschers, R., Finke, P. A., and de Gruijter, J. J. (2007). A soil sampling program for the Netherlands. *Geoderma*, 139(1-2):60–72.
- Voltz, M., Lagacherie, P., and Louchart, X. (1997). Predicting soil properties over a region using sample information from

- a mapped reference area. *European Journal of Soil Science*, 48(1):19–30.
- Voltz, M. and Webster, R. (1990). A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *European Journal of Soil Science*, 41(3):473–490.
- Walter, C., Lagacherie, P., and Follain, S. (2007). Integrating pedological knowledge into soil digital mapping. In Lagacherie, P., McBratney, A. B., and Voltz, M., editors, *Digital Soil Mapping - An Introductory Perspective*, volume 31, pages 281–300. Elsevier, Amsterdam.
- Walvoort, D., Brus, D., and Heuvelink, G. (2011). Building a three dimensional soil model by combining data sources of various degrees of uncertainty. *Proceedings of the Pedometrics 2011 Conference*. Trest, Czech Republic, 30 August–3 September 2011.
- Walvoort, D. J. J., Brus, D. J., and de Gruijter, J. J. (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*, 36(10):1261–1267.
- Webster, R. (1994). The development of pedometrics. *Geoderma*, 62(1-3):1–15.
- Webster, R. and Heuvelink, G. B. M. (2006). The Kalman filter for the pedologist’s tool kit. *European Journal of Soil Science*, 57(6):758–773.
- Webster, R. and McBratney, A. B. (1989). On the Akaike Information Criterion for choosing models for variograms of soil properties. *Journal of Soil Science*, 40(3):493–496.
- Webster, R. and Oliver, M. (2007). *Geostatistics for environmental scientists*. Statistics in practice. John Wiley & Sons, Chichester, second edition edition.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Use R. Springer, New York.
- Wösten, J. H. W. and van Genuchten, M. T. (1988). Using texture and other soil properties to predict the unsaturated soil hydraulic functions. *Soil Sci Soc Am J*, 52(3-4):1762–1770.
- Zhu, A. X., Hudson, B., Burt, J., Lubich, K., and Simonson, D. (2001). Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Sci Soc Am J*, 65(5):1463–1472.
- Zhu, Z. and Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):24–44.



Summary

Soils are back on the global political agenda. Renewed interest of the soil resource is fuelled by an increasing awareness about the importance of sustainable soil management to secure production of food and fiber for a quickly growing world population, and about the major role of soils in the global carbon cycle. With this has come great demand for accurate, up-to-date and detailed geographical soil information. The current generation soil information systems typically store data from conventional surveys. Besides soil data at points, these systems contain soil maps that are often restricted to soil type; thematic soil maps are mostly missing. The maps are frequently outdated, lack detail and quantitative information on accuracy, or have no full spatial coverage. Consequently, these data are of limited use in today's soil data applications.

The Dutch soil information system *BIS* is no exception to this situation. The main source of soil information in the Netherlands, the nationwide 1:50 000 soil map, is becoming outdated, particularly for the areas with peat soils, and needs to be updated. Furthermore, maps of basic soil properties with quantified accuracy are lacking. Such maps are essential input for environmental process models that predict the effect of policy measures on for example soil acidification, pesticide leaching and greenhouse gas emission. Now the urgent need is felt to update the national soil map and to extend *BIS* with full-coverage thematic maps of all major soil properties with quantified accuracy. Efficient, quantitative methods for (geo)statistical modelling of soil maps, referred to as digital soil mapping (DSM), might be very useful for this purpose. Yet, despite growing global popularity DSM has not been applied in an operational way in the Netherlands so far. The main objective of this thesis is therefore to investigate and evaluate the merit of DSM for updating soil information in the Netherlands. Research focuses on DSM methods for updating soil type maps as well as maps of continuous soil properties. The province of Drenthe with large areas of peat soils is selected as case study area to illustrate and evaluate the developed methods.

After the general introduction in Chapter 1, Chapter 2 describes a study on the possibility of updating the 1:50 000 soil map using a simple generalized linear regression

model and legacy soil point data from *BIS*. Map unit-specific multinomial logit models (MLM) were used to predict probability distributions of soil types within ten map units of the simplified soil map 1:50 000. For this purpose a framework for selecting an MLM was taken from the literature and adapted for soil mapping. Updating not only focused on peat soils but also on mineral soils to investigate if the purity of these map units could be increased through disaggregation with high-resolution covariates. Validation showed a modest 6% improvement in map purity compared to the existing, outdated soil map. This improvement was mainly attributed to better representation of soil distribution within the peat map units of the simplified map. However, map unit purities and class representations of the four peat soils as depicted on the updated map were still small.

Digital soil type maps offer new possibilities for mapping individual soil properties. Chapter 3 describes the development of a model that exploits the information from such soil type map for spatial prediction of continuous soil properties. This model has important advantages compared to the conventional geostatistical model. First, actual (observed) soil type at sampling locations can be used as covariate instead of the mapped soil type. This has the advantage that the relationship between soil property and soil type is not confounded by impurities in the map units. Second, using actual soil type as covariate in the model makes it possible to quantify the proportion of the prediction variance that arises from uncertainty of the actual soil type at prediction locations. The developed model is applied to map the soil organic matter (SOM) content using the digital soil type map created in Chapter 2. Validation showed that the prediction performance of the proposed model was slightly better than that of the conventional geostatistical model.

In Chapter 4 a method is proposed for three-dimensional mapping of SOM that combines general pedological knowledge with geostatistical modelling. A conceptual SOM depth profile was constructed by stacking building blocks (model horizons) for each soil type depicted on the updated digital map from Chapter 2. The vertical distribution of SOM within each building block was described by a function. The combination of building blocks—stacked in pre-defined order—with their associated parameters (thickness, average SOM content, exponential decay parameters) describes a soil type-specific depth profile. The parameters of each of these depth profiles were spatially predicted by geostatistical interpolation with covariates. A probability distribution of soil type-specific depth functions was then obtained by combining these predictions with the digital soil type map from Chapter 2. The depth functions and their associated probabilities were used to map the SOM stock for four depth intervals using the methodology described in Chapter 3. Validation of the predicted stocks with an independent probability sample showed accurate results for the topsoil. Results for deeper soil layers, however, were modest. Prediction performance of pedometric depth functions was comparable to that of conventional depth functions.

The main drawback of the MLM, which was applied for soil type mapping in Chapter 2, is that spatial dependency in the data is not exploited for spatial prediction. Chapter 5 addresses this issue and investigates if a soil type map predicted by a spatial model is more accurate than one predicted by a non-spatial model. As spatial model the generalized linear geostatistical model (GLGM) was chosen. The GLGM is central to the methodological framework of model-based geostatistics, which is considered state-of-the-art in DSM. A pragmatic approach was adopted in which each of the five soil types in the case study area in the cultivated peatlands was modelled separately with a binomial logit-linear GLGM. Predictions with the soil type-specific GLGMs resulted in five binomial probabilities at each prediction location, which were scaled to multinomial probabilities so that they sum to one. Validation showed that use of a spatial model for digital soil type mapping did not result in more accurate predictions of soil type than those with the non-spatial MLM.

Chapter 6 compares the efficiency of DSM methods with that of conventional soil mapping (CSM) methods for updating soil type and property maps. In addition, the effect of mapping effort (expressed in a monetary unit per ha) on accuracy is assessed for digital soil type and property maps. For digital soil type mapping the GLGM was used. For soil property mapping (SOM content en peat thickness) two methods are considered for both DSM and CSM. For DSM these are the method from Chapter 3 and the conventional geostatistical method (universal kriging). For CSM these are the representative profile descriptions (RPD) and map-unit-means (MUM) methods. For DSM both methods gave similar results in terms of accuracy. The MUM method gave better results than the RPD. For CSM the MUM method gave better results than the RPD. Validation results further showed that DSM produced soil type and property maps that were of similar accuracy as those produced by CSM. Furthermore, DSM maps were produced much more efficiently than the CSM maps: costs per hectare were a factor three to four smaller without compromising accuracy. This shows that for future updating of soil information DSM can be an attractive alternative to CSM.

Finally, Chapter 7 presents a synthesis of the results and the main findings of Chapters 2 to 6. Implications of the results for the soil information system *BIS* and future updating of soil information in the Netherlands are discussed and an outlook on future research is given. It is argued that soil survey is shifting from conventional, qualitative soil survey to quantitative soil survey. This means that a toolbox with quantitative, state-of-the-art methods for soil mapping is not sufficient for effective and successful operational use of DSM. It requires the development of a next-generation soil information system based on new strategies and methods for collecting, storing, processing, visualizing and disseminating soil information. This thesis presents a first step on the road towards such system.



Samenvatting

Bodems zijn terug op de wereldwijde politieke agenda. Hernieuwde interesse in de bodem wordt gevoed door een groeiend bewustzijn voor het belang van duurzaam bodemmanagement voor het veiligstellen van de productie van voedsel en vezels voor een snel groeiende wereldbevolking, en van de belangrijke rol die de bodem speelt in de globale koolstofcyclus. Dit heeft geleid tot een sterk toenemende vraag naar actuele en gedetailleerde geografische bodeminformatie. De huidige generatie bodemkundige informatiesystemen bevatten voornamelijk gegevens afkomstig van conventionele bodemkarteringen. Naast bodemgegevens van puntlocaties, bevatten deze systemen bodemkaarten die meestal alleen het bodemtype aangeven; thematische bodemkaarten ontbreken meestal. De kaarten zijn vaak verouderd, bevatten geen detail en kwantitatieve informatie over de nauwkeurigheid of dekken de ruimte niet volledig. De bodemgegevens zijn daarvoor beperkt bruikbaar voor de huidige toepassingen.

Het Nederlandse bodemkundig informatiesysteem *BIS* is geen uitzondering op deze situatie. De belangrijkste bron van bodeminformatie in Nederland, de landelijke 1:50 000 bodemkaart, raakt verouderd, vooral voor de gebieden met veengronden, en moet daarom worden geactualiseerd. Daarnaast ontbreken kaarten van basale bodemkenmerken met gekwantificeerde nauwkeurigheid, die essentieel zijn als invoer voor procesmodellen die het effect van beleidsmaatregelen op bijvoorbeeld bodemverzuring, uitspoeling van pesticiden en uitstoot van broeikasgassen voorspellen. Daarom is er nu de dringende behoefte om de nationale bodemkaart te actualiseren en om *BIS* uit te breiden met landsdekkende, thematische kaarten van alle belangrijke bodemkenmerken met gekwantificeerde nauwkeurigheid. Efficiënte, kwantitatieve methoden voor het (geo)statistisch modelleren van bodemkaarten, oftewel *digitale bodemkartering* (DBK), kunnen hiervoor uitermate geschikt zijn. Ondanks groeiende wereldwijde populariteit wordt in Nederland DBK tot nu toe niet operationeel toegepast. Het hoofddoel van dit proefschrift is daarom te onderzoeken wat de waarde van DBK is voor het actualiseren van bodeminformatie in Nederland. Het onderzoek richt zich op DBK-methoden voor het actualiseren van zowel bodemtypekaarten als thematische bodemkaarten. De provincie Drenthe, met grote gebieden met veengronden, is gekozen als studiegebied om de ontwikkelde metho-

den te illustreren en beoordelen.

Na een algemene introductie (Hoofdstuk 1) beschrijft Hoofdstuk 2 het onderzoek naar de mogelijkheid om de 1:50 000 bodemkaart te actualiseren gebruikmakend van een eenvoudig gegeneraliseerd lineair regressie model en puntwaarnemingen uit het *BIS*. Multinomiale logit-modellen (MLM) voor specifieke kaarteenheden werden gebruikt om kansverdelingen van bodemtypes te voorspellen binnen tien kaarteenheden van een vereenvoudigde bodemkaart 1:50 000. Hiervoor werd een raamwerk voor selectie van een MLM uit de literatuur genomen en aangepast voor bodemkartering. Actualisatie richtte zich niet alleen op veengronden maar ook op minerale bodems, om te onderzoeken of de zuiverheid van deze kaarteenheden kon worden verhoogd door desaggregatie met hulpvariabelen van hoge resolutie. De validatiere-sultaten toonden een bescheiden verbetering in kaartzuiverheid van 6% ten opzichte van de bestaande, verouderde bodemkaart. Deze verbetering was voornamelijk toe te schrijven aan een betere representatie van de ruimtelijke spreiding van bodemtypen binnen de veenkaarteenheden van de bestaande, vereenvoudigde bodemkaart. De zuiverheden van de kaarteenheden en de klassevertegenwoordiging van de vier veengronden zoals afgebeeld op de geactualiseerde kaart waren echter nog steeds laag.

Digitale bodemtypekaarten bieden nieuwe mogelijkheden voor het in kaart brengen van afzonderlijke bodemkenmerken. Hoofdstuk 3 beschrijft de ontwikkeling van een model dat informatie uit een dergelijke bodemtypekaart benut voor ruimtelijke voorspelling van continue bodemkenmerken. Dit model heeft belangrijke voordelen ten opzichte van het conventionele geostatistische model. Ten eerste kan het eigenlijke (waargenomen) bodemtype op bemonsteringslocaties worden gebruikt als hulpvariabele in plaats van het gekarteerde bodemtype. Dit heeft als voordeel dat de relatie tussen bodemkenmerk en bodemtype niet verstoord wordt door onzuiverheden binnen de kaarteenheden. Ten tweede maakt gebruik van het eigenlijke bodemtype als hulpvariabele in het model het mogelijk om te kwantificeren welk aandeel onzekerheid over het eigenlijke bodemtype op de voorspellingslocaties heeft in de voorspellingsvariantie. Het model is toegepast om het organischestofgehalte (OS) in de bovengrond in kaart te brengen, gebruikmakend van de digitale bodemtypekaart uit Hoofdstuk 2. Validatie liet zien dat voorspellingsprestatie met het voorgestelde model iets beter was dan die van het conventionele geostatistische model.

In Hoofdstuk 4 wordt een methode voorgesteld voor driedimensionale kartering van OS, die algemene bodemkundige kennis combineert met geostatistisch modelleren. Voor elk bodemtype, afgebeeld op de geactualiseerde digitale kaart uit Hoofdstuk 2, werd een conceptueel OS-diepteprofiel geconstrueerd door middel van het stapelen van bouwstenen (modelhorizonten). De verticale verdeling van OS binnen elke bouwsteen wordt beschreven door een functie. De combinatie van bouwstenen—gestapeld in voorgedefinieerde volgorde—met hun bijbehorende parameters (dikte, gemiddeld OS-gehalte, exponentiële afname parameters) beschrijft een bodemtype-

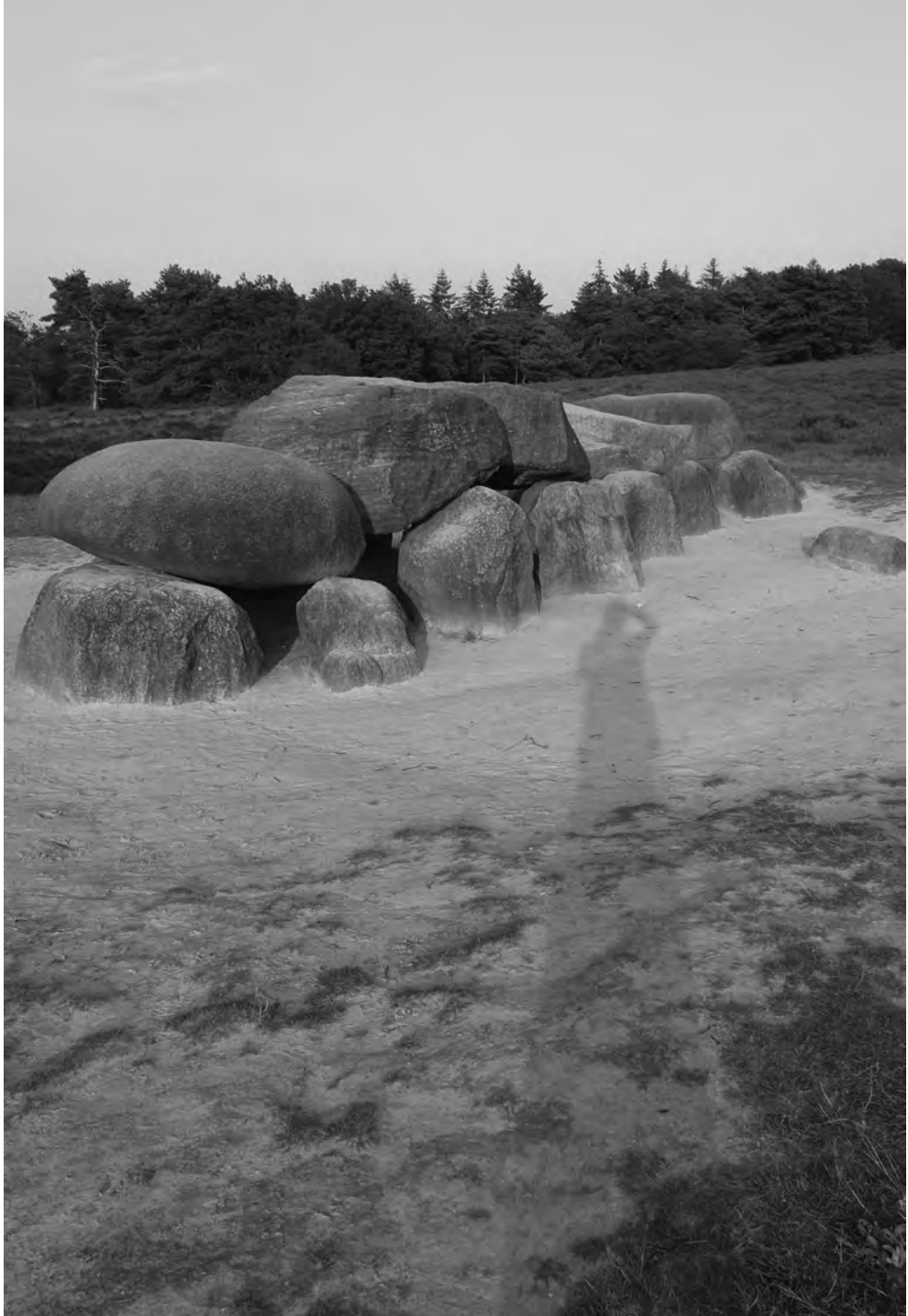
specifiek diepteprofiel. De parameters van elk van deze diepteprofielen werden vervolgens ruimtelijk voorspeld door geostatistische interpolatie met hulpvariabelen. Door combinatie van deze voorspellingen met de digitale bodemkaart uit Hoofdstuk 2 werd een kansverdeling van bodemtype-specifieke dieptefuncties verkregen. Deze dieptefuncties met hun bijbehorende kansen werden vervolgens gebruikt om de OS voorraad voor vier diepte-intervallen in kaart te brengen, gebruikmakend van de methode die in Hoofdstuk 3 is beschreven. Validatie van de voorspelde voorraden met een onafhankelijke kanssteekproef liet nauwkeurige resultaten zien voor de bovengrond. Resultaten voor diepere bodemlagen waren minder nauwkeurig. De voorspellingen met pedometrische dieptefuncties waren even nauwkeurig als die met conventionele dieptefuncties.

De belangrijkste tekortkoming van het MLM, gebruikt voor digitale bodemtypekartering in Hoofdstuk 2, is dat ruimtelijke afhankelijkheid in de gegevens niet wordt benut bij het ruimtelijk voorspellen. Hoofdstuk 5 richt zich op deze kwestie en onderzoekt welke ruimtelijke voorspellingen van bodemtypen nauwkeuriger zijn: voorspellingen met een ruimtelijk model, of voorspellingen met een niet-ruimtelijk model? Als ruimtelijk model werd het gegeneraliseerde lineaire geostatistische model (GLGM) gekozen. Het GLGM staat centraal in het methodologische raamwerk van model-gebaseerde geostatistiek, welke beschouwd wordt als *state-of-the-art* in DBK. Uit pragmatische overwegingen werd elk van de vijf bodemtypen in een studiegebied in de veenkoloniën individueel gemodelleerd met een binomiaal logit-lineair GLGM. Voorspellingen met de bodemtype-specifieke modellen resulteerden in vijf binomiale kansen op iedere voorspellingslocatie, die werden geschaald tot multinomiale kansen zodat deze optellen tot één. Validatie liet zien dat met een ruimtelijk model voor bodemtypekartering de bodemtypen minder nauwkeurig werden voorspeld dan met het niet-ruimtelijke MLM.

Hoofdstuk 6 vergelijkt de efficiëntie van DBK-methoden voor het actualiseren van kaarten van bodemtype en -kenmerken met die van conventionele methoden voor bodemkartering. Daarnaast is het effect van karteringsinspanning (uitgedrukt in een monetaire eenheid per ha) op nauwkeurigheid onderzocht voor de digitale kaarten van bodemtypen en bodemkenmerk. Voor digitale bodemtypekartering werd het GLGM gebruikt. Om de bodemkenmerken OS-gehalte en veendikte in kaart te brengen werden voor zowel DBK als conventionele bodemkartering (CBK) twee methoden onderzocht. Voor DBK zijn dit de methode uit Hoofdstuk 3 en de conventionele geostatistische methode (*universal kriging*). Voor CBK zijn dit de representatieve profielbeschrijving (RPB) en methoden die zijn gebaseerd op kaarten-gemiddelden (KEG). Voor DBK gaven beide resultaten van vergelijkbare nauwkeurigheid. Voor CBK presteerde de KEG-methode beter dan de RPB-methode in termen van voorspelnauwkeurigheid. Validatieresultaten toonden verder aan dat kaarten van bodemtype en bodemkenmerken, geproduceerd met DBK, even nauwkeurig waren als kaarten, geproduceerd met CBK. Bovendien werden de DBK-kaarten een stuk efficiënter geproduceerd dan de CBK-kaarten: de kosten per hectare waren een factor

drie tot vier lager. Deze resultaten laat zien dat voor toekomstige actualisaties van bodeminformatie DBK een aantrekkelijk alternatief voor CBK kan zijn.

Ten slotte presenteert Hoofdstuk 7 een synthese van de resultaten en de belangrijkste bevindingen van Hoofdstukken 2 tot en met 6. Naast een bespreking van de implicaties van de resultaten voor het bodemkundig informatiesysteem *BIS* en voor toekomstige actualisatie van bodeminformatie in Nederland blijkt Hoofdstuk 7 vooruit op toekomstig onderzoek. Er wordt betoogd dat de bodeminventarisatie in Nederland verschuift van conventionele, kwalitatieve bodeminventarisatie naar kwantitatieve bodeminventarisatie. Dit betekent dat een gereedschapskist met kwantitatieve, *state-of-the-art*-methoden voor bodemkartering niet genoeg is voor effectief en succesvol operationeel gebruik van DBK. Het vereist de ontwikkeling van een nieuw bodemkundig informatiesysteem dat is gebaseerd op nieuwe strategieën en methoden voor het verzamelen, opslaan, verwerken, visualiseren en ontsluiten van bodeminformatie. Dit proefschrift presenteert een eerste stap op de weg naar zo'n systeem.



About the author

Curriculum Vitae

Bas Kempen was born on 3 March 1980 in Boxmeer, the Netherlands. He attended secondary school at Scholengemeenschap Stevensbeek where he obtained his VWO diploma in 1998. Not knowing whether he should study chemistry or Earth science, he started studying Molecular Sciences at Wageningen University in that same year. However, after obtaining his propaedeutic diploma (with distinction) in 1999, he changed to Soil, Water & Atmosphere at the same university. In 2004, he obtained an MSc degree in Soil Science, specializing in Soil Inventory and Land Evaluation with a major thesis on landslide modelling in New Zealand, a minor thesis on land evaluation in the Mekong delta in Vietnam and an internship with Alterra. After graduating, Bas continued his studies in Wageningen with a Master's in Geo-Information Science and Remote Sensing. It was during this time that he was introduced to the exciting world of pedometrics and digital soil mapping. In 2004, he attended the first global workshop on digital soil mapping in Montpellier, France. In 2005, he obtained a second MSc degree (with distinction) with a thesis on digital soil mapping in Senegal. This thesis was awarded the Hissink Award of the Dutch Soil Science Society for the best MSc thesis in soil science in 2005 and the KLV Thesis Award for the best MSc thesis in environmental science at Wageningen University in 2005. From January to October 2006, Bas worked as an assistant instructor for the Laboratory of Geo-Information Science and Remote Sensing at Wageningen University. During this time he assisted with the course 'Introduction to Geo-Information Science' and developed new course materials. In October 2006, Bas started his PhD research at Wageningen University with the Soil Inventory and Land Evaluation Group (later the Land Dynamics Group). Since October 2009, Bas works as a researcher at Alterra for the Soil Geography team. In his spare time Bas likes to play squash and he is a fervent player of strategic board and card games, which keeps him busy several evenings a week.

Publications

Refereed articles

1. **Kempen, B.**, Brus, D.J., Heuvelink, G.B.M. Soil type mapping with the generalized linear geostatistical model: a case study in a Dutch cultivated peatland. *Submitted to Geoderma*.
2. **Kempen, B.**, Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B.M., de Vries, F. Efficiency comparison of conventional and digital soil mapping for updating soil maps of a cultivated peatland. *Submitted to Soil Science Society of America Journal*.
3. Brus, D.J., **Kempen, B.**, Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. *European Journal of Soil Science* 62(3), 394-307.
4. **Kempen, B.**, Brus, D.J., Stoorvogel, J.J., 2011. Three-dimensional mapping of soil organic matter using soil type-specific depth functions. *Geoderma* 162(1-2), 107-123.
5. **Kempen, B.**, Heuvelink, G.B.M., Brus, D.J., Stoorvogel, J.J., 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. *European Journal of Soil Science* 61(3), 333-347.
6. **Kempen, B.**, Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. *Geoderma* 151(3-4), 311-326.
7. Stoorvogel, J.J., **Kempen, B.**, Heuvelink, G.B.M., de Bruin, S., 2009. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal, *Geoderma* 149(1-2), 161-170.

Conference contributions

1. **Kempen, B.**, Brus, D.J., Stoorvogel, J.J., 2010. Three-dimensional mapping of soil organic matter using soil type-specific depth functions. 4th Global Workshop on Digital Soil Mapping, May 24-26, Rome, Italy.
2. Heuvelink, G.B.M., Brus, D.J., de Vries, F., Vašát R., Walvoort, D.J.J., **Kempen, B.**, Knotters, M., 2010. Implications of digital soil mapping for soil information systems. 4th Global Workshop on Digital Soil Mapping, May 24-26, Rome, Italy.
3. **Kempen, B.**, Heuvelink, G.B.M., Brus, D.J., Stoorvogel, J.J., 2009. Pedometric mapping of soil organic matter using uncertain soil type information from a soil map. *Pedometrics* 2009, August 26-28, Beijing, China.
4. Brus, D.J., **Kempen, B.**, 2009. Sampling for validation of digital soil maps. *Pedometrics* 2009, August 26-28, Beijing, China.
5. **Kempen, B.**, Brus, D.J., Heuvelink, G.B.M. Stoorvogel, J.J., 2008. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. 3rd Global Workshop on Digital Soil Mapping, September 30-October 3, Logan, UT, USA.
6. **Kempen, B.**, Stoorvogel, J.J., Heuvelink, G.B.M., de Bruin, S., 2007. Implementation and validation of expert knowledge for digital soil mapping in Senegal, *Pedometrics* 2007, August 27-30, Tuebingen, Germany.
7. Trung, N.H., **Kempen, B.**, Tri, L.Q., van Mensvoort, M.E.F., Bregt, A.K., 2006. Application du SIG et evaluation multi-critères pour l'aménagement du territoire dans la zone côtière du delta du Mékong, Vietnam. *Colloque International: Application de la télédétection, des SIG et des GPS pour la réduction des risques naturels et le développement durable*, March 11-14, Hanoi, Vietnam.

Other publications

1. van der Salm, C., Boumans, L.J.M., Brus, D.J., **Kempen, B.**, van Leeuwen, T.C., 2011. Validatie van het nutriënten-emissiemodel STONE met meetgegevens uit het Landelijk Meetnet effecten Mestbeleid (LMM) en de Landelijke Steekproef Kaartenheden (LSK). *Wettelijke Onderzoekstaken Natuur & Milieu, Werkdocument* 228.
2. Knotters, M., Brus, D.J., Heuvelink, G.B.M., **Kempen, B.**, de Vries, F., Walvoort, D.J.J., 2010. Vaste grond onder de voeten? Geactualiseerd bodemkundig informatie systeem informeert over onzekerheid. *Bodem* 20(5), 22-25.

PE&RC Education Certificate

With the educational activities listed below the PhD candidate has complied with the educational requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities).



Writing of project proposal (4.5 ECTS)

- Digital soil mapping in man-dominated landscapes (2007)

Post-graduate courses (5.5 ECTS)

- Land science: bringing concepts and theory into practice; PE&RC, LAD (2007)
- Advanced statistics; PE&RC (2007)

Laboratory training and working visits (4.5 ECTS)

- Working visit; Cornell University (2011)

Invited review of (unpublished) journal manuscript (5 ECTS)

- Geoderma: geopedological soil mapping (2009)
- Environmental Monitoring and Assessment: soil quality assessment (2010)
- Geoderma: uncertainty assessment for mapping depth functions (2010)
- European Journal of Soil Science: effects of management and physical factors on SOC content (2010)
- CATENA: spatial prediction of soil groups (2011)

Deficiency, refresh, brush-up courses (4.2 ECTS)

- Hands-on geostatistics: merging GIS & spatial statistics (2007)
- Sampling for monitoring (2007)
- Basic statistics (2007)

Competence strengthening / skills courses (3.2 ECTS)

- PhD Competence assessment; WGS (2006)
- The art of writing; CENTA (2007)
- Techniques for writing and presenting a scientific paper; WGS (2008)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.2 ECTS)

- PE&RC Weekend (2007)
- Symposium "The role of long-term landscape reconstruction for understanding current-day rates of landscape change" (2008)
- Symposium "Soils and spectroscopy" (2009)

Discussion groups / local seminars / other scientific meetings (4.8 ECTS)

- Spatial methods (2006-2011)
- NBV Themadagen (2006-2010)
- Symposium Bodeminformatie (2008)
- Workshop going "All-terrain" (2009)
- Symposia BIS2014 (2010-2011)

International symposia, workshops and conferences (8.8 ECTS)

- Pedometrics; Tuebingen, Germany (2007)
- 3rd Global Workshop on Digital Soil Mapping; Logan UT, US (2008)
- Pedometrics; Beijing, China (2009)
- 4th Global Workshop on Digital Soil Mapping; Rome, Italy (2010)

Lecturing / supervision of practical's / tutorials (43 days)

- Frontiers of land science (2007)
- Bodem-water II (2007-2009)
- Veldpracticum geologie, bodem en landschap Nederland (2008-2011)
- Bodem-water I (2008-2009)

This research is part of the strategic research programme 'Sustainable spatial development of ecosystems, landscapes, seas and regions' which is funded by the Dutch Ministry of Economic Affairs, Agriculture and Innovation, and carried out by Wageningen University and Research Centre.

This thesis is printed on FSC mixed sources paper.

Cover design by Dirk Oosterink, rrrrronk.blogspot.com.